Johannes Jansson

ANALYS AV BIBLIOTEKSDATA MED HJÄLP AV SJÄLVORGANISERANDE KARTOR

Pro gradu-avhandling i informationssystem

Handledare: ED Anna Sell

Fakulteten för samhällsvetenskaper och ekonomi

Åbo Akademi

Åbo 2018

ABSTRAKT

ÅBO AKADEMI - Fakulteten för samhällsvetenskaper och ekonomi
Ämne: Informationssystem
Författare: Johannes Jansson
Arbetets titel: Analys av biblioteksdata med hjälp av självorganiserande kartor
Handledare: ED Anna Sell

Abstrakt:

Företag och andra institutioner samlar in och besitter en hel mängd data i dagens läge. Data samlas in bland annat med hjälp av sociala media, streckkoder, kassaapparater och RFID. Även de finska biblioteken som organisationer har under årens lopp samlat in data om deras kunder och verksamhet. Data består bland annat av besöksstatistik, utlåning och kostnadskalkyler. På basen av insamlad data har nyckeltal beräknats som illustrerar biblioteksväsendets verksamhet.

Avhandlingen använde nyckeltalen som beräknats av de finska allmänna biblioteken för att skapa en självorganiserande karta för fortsatt analys. Visualiseringsmetoder som U-matris och komponentplan användes för att analysera av hur olika variabler fördelar sig på kartan. Dessutom gjordes klusteranalyser för att se vilka biblioteksväsen som liknar varandra och vilka som sticker ut ur mängden. Komponentplanen och klusteranalyserna jämfördes med varandra för att se vilka variabler som påverkat vilka kluster och till vilken grad.

Resultaten av analysen visar att de stora städernas biblioteksväsen finns i samma kluster och flera av dem i samma neuron. Vad det visar är att de stora städernas biblioteksväsen är till karaktären likartade. De har alla medelhöga till höga besökssiffror och låga kostnader. Utöver de stora städerna fanns det enskilda kluster med få biblioteksväsen i sig som stack ut ur mängden. De karaktäriseras av låga till moderata besökssiffror och höga kostnader.

Nyckelord: biblioteksdata, självorganiserande karta, visualisering, klusteranalys, kluster

Datum: 14.12.2018	**Sidoantal**: 74

INNEHÅLLSFÖRTECKNING

BILDFÖRTECKNING

TABELLFÖRTECKNING

1. INLEDNING

Intresset för ämnet väcktes i samband med studierna på fördjupad nivå inom huvudämnet informationssystem och studierna inom biämnet informationsvetenskap. De fördjupade studierna bestod av kurser som handlade om användning och analys av stora mängder data. Idén med att analysera den data som biblioteken redan besitter bottnar i två specifika intresseområden: bibliotek och dataanalysmetoder.

Problemområdet som avhandlingen fokuserar på är hur biblioteken kan dra nytta av den data som de samlat in i samband med deras verksamhet. Denna data kan beskrivas som en del av vad som kallas Big Data vilket innebär data som företag och institutioner använder som underlag för beslut och som styr deras verksamhet. Bibliotek som vilka andra organisationer som helst i dagens samhälle samlar in och besitter en mängd data. Inom biblioteksväsendena finns det ett intresse att utnyttja insamlad data för verksamheten men det finns inte alltid den nödvändiga kunskapen att göra det.

Genom att tillämpa självorganiserande kartor på biblioteksdata kan man producera en visuell representation av data. Visualiseringen gör att data blir mera överskådligt och lättare för bibliotekspersonal och beslutsfattare att ta till sig. Med hjälp av själv-organiserande kartorna kan man välja att fokusera på individuella nyckeltal för att se hurudan bild de ger av bibliotekens verksamhet. Nyckeltal som t.ex. öppettider i förhållande till mängden besökare vilket sedan kan användas för att avgöra vilken och hur mycket personal som krävs under vilka tider. Det är även möjligt att se om satsningar på biblioteksväsendet har lönat sig i form av ökat besökarantal och utlåningar.

1.1 Forskningsfrågor

Det som är av intresse för avhandlingen är vad karaktäriserar storstädernas biblioteksväsen. Storstäderna har korta avstånd och hög befolkningstäthet jämfört med landsbygden vilket kan leda till att de har karaktärsdrag som skiljer dem från mängden.

Utöver storstäderna är det intressant att ta reda på om det finns avvikande fall bland de övriga biblioteksväsendena.

Konkret vill avhandlingen svara på två forskningsfrågor:

1) Finns det avvikande fall?
2) Går det med hjälp av självorganiserande kartor att hitta biblioteksväsen vars verksamhet på basen av nyckeltalsanalys skiljer sig från övriga biblioteksväsen?

1.2 Metod

Avhandlingen stöder sig på den data som Finlands allmänna bibliotek har matat in i databasen på tilastot.kirjastot.fi. Till sin natur är metoden kvantitativ. Den metod som används för analysen kallas för självorganiserande kartor (*eng. Self Organizing Maps*). Denna metod har valts för att den möjliggör en visuell analys av hur de olika entiteterna, i detta fall kommunernas biblioteksväsen står i förhållande till varandra. Det här innebär att man kan utreda vilka av biblioteken faller inom samma kluster på basen av den data som fås ur databasen. Genom självorganiserande kartor kan man se ifall det finns biblioteksväsen som avviker från mängden.

1.3 Förväntade resultat

Det som kan uppnås med denna avhandling är en ökad förståelse för dataanalys inom biblioteken. Det vill säga hur de kan använda dataanalys för att optimera sin verksamhet och kunna på så sätt hållas relevanta i en digital värld. De som kan ha nytta av denna avhandling är beslutsfattare inom biblioteksväsendet och även personalen på biblioteken. Det är även möjligt att ledningen inom kommunernas biblioteksväsen kan ha nytta av denna avhandling.

1.4 Disposition

Avhandlingen är indelad i sex kapitel plus källförteckningen. Det första kapitlet innefattar inledningen. Kapitlen två och tre innefattar litteraturen som ligger till grund för avhandlingen. Kapitel två handlar om datautvinning och kapitel tre handlar om biblioteksdata. I kapitel fyra diskuteras och presenteras datasetet och analysverktyget. Resultatet av analysen presenteras och diskuteras i kapitel fem och i kapitel sex förs slutdiskussionen och dras slutsatser. Tillsist finns källförteckningen som innefattar all litteratur som använts i avhandlingen.

2. DATAUTVINNING

Kapitlet handlar om de begrepp som är relevanta för avhandlingen. Först definieras och diskuteras begreppet datautvinning, för att begreppet är den kärna runt vilken avhandlingen byggs upp. Alla andra begrepp som diskuteras i denna avhandling bygger till en större eller mindre del direkt på datautvinning. I samband med detta begrepp diskuteras även kort olika analysmetoder som hänger ihop med begreppet.

2.1 Bakgrund

Enligt Olson & Delen (2008, kap. 1) handlar datautvinning (eng. *data mining*) om att utvinna information ur stora mängder data som lagras i datorer. Fayyad m.fl. (1996) i sin tur definierar datautvinning som tillämpningen av dataanalys och matematiska algoritmer för att upptäcka mönster i data. Enligt Fayyad m.fl. (1996) är datautvinning en del av ett bredare begrepp som heter *upptäckande av kunskap i databaser* (eng. *Knowledge Discovery in Databases*) och förkortas KDD. Vad begreppet syftar på den icke-triviala process som går ut på att hitta giltiga, tidigare okända, potentiellt användbara och förståeliga mönster i data. Enligt Mourya och Gupta (2012, s. 17) är KDD och datautvinning synonymer. Mourya och Gupta (2012, s. 17) definierar datautvinning i korthet som icke-trivial utvinning av obetingad, tidigare okänd och potentiellt användbar information ur data.

Datautvinning har även under sin historia kallats för utforskande dataanalys (eng. *exploratory data analysis*) vilket innebär att man utan förbehåll utforskar vilken information data innehåller (Brown 2014, kap. 1). Data som används inom datautvinning genereras ur kassaapparater, streckkoder, sociala media och ur databaser. Dessa data innefattar även kundernas användarprofiler och relaterade data. Till exempel har Facebook en hel mängd data om sina användare (Olson och Delen 2008, kap. 1; Mourya och Gupta 2012, ss. 18-19).

Olika matematiska modeller används sedan för att förutspå vad som kan vara av intresse för företaget t.ex. försäljning, kunders reaktioner på reklam samt för att förutspå företagets vinst (Brown 2014, kap. 1; Mourya och Gupta 2012, s. 25). För att alls klara av datautvinning måste man vara insatt i statistik. Nuförtiden är det även möjligt att använda sig av AI d.v.s. artificiell intelligens för att automatisera datautvinningen. Trots den tekniska utvecklingen är en systematisk utvinning av information ur data med hjälp av statistiska metoder fortfarande det som utgör ryggraden inom datautvinning. Vidare kan man säga att trots att teknikens utveckling gjort det möjligt att använda informationsteknologi för att automatisera utvinningen krävs det mänsklig intelligens. Intelligensen kommer i form av att en person väljer de metoder och modeller som används inom utvinningsprocessen (Olson och Delen 2008, kap. 1).

Företag och andra aktörer som besitter dessa data använder sig av datautvinning för olika syften. Ett av dessa syften är att företagen vill bättre förstå hur olika processer fungerar inom företaget och hur dessa processer påverkar varandra sinsemellan. Företagen vill även veta hur de kan bäst optimera processerna i syfte att göra dem mera effektiva. Till exempel om det lönar sig att satsa pengar på marknadsföring via sociala medier för att öka försäljningen (Brown 2014, kap. 1; Mourya och Gupta 2012, s. 28).

Man kunde alltså säga att datautvinning handlar om hur man använder olika data-analysmetoder för att utvinna användbar information ur data för ett praktiskt syfte. För att konkretisera detta kan man säga att de inom företaget som fattar beslut behöver kunna förstå inte bara det som hänt tidigare utan det som kommer att hända. Det är här som datautvinning kan dra sitt strå till stacken inte bara genom att analysera det som har hänt och utan även berätta varför det hänt. Med hjälp av dataanalys kan man även förutsäga vad som kommer att hända på basen av det som hänt tidigare (Brown 2014, kap. 1; Mourya och Gupta 2012, s. 26). Det är även möjligt för företag att identifiera vilka kunder som är mest lönsamma, vilka av dem som är i risk att gå över till en konkurrent och vilka produkter som de sannolikt köper. Denna information kan sedan användas för marknadsföring (Olson och Delen 2008, kap. 1; Mourya och Gupta 2012, s. 28).

Som ett exempel på en konkret tillämpning av datautvinning kan man titta på ett fall där ett försäkringsbolag lyckades optimera sin verksamhet genom datautvinning. Bolaget hade genom datautvinning konstaterat att ett av dess filialkontor handlägger en viss typs ansökningar snabbare än de andra kontoren. Den informationen ledde till att företaget kunde utveckla processer för att förbättra handläggningen inom hela företaget och samtidigt hitta det bästa tillvägagångssättet för denna typ av arbete (Brown 2014, kap. 1).

Olson och Delen (2008, kap. 1) lyfter fram ett fall där datautvinning kan användas inom dagligvaruhandeln för att hålla reda på de produkter som säljs och de som köps in. Process med produkternas rörelse in och ut ur företagen leder till att det samlas en hel mängd data. Data har samlats in genom att varje produkt har en egen streckkod, koderna läses när produkter anländer till en butik och när det säljs. Genom datautvinning kan butikerna ha översikt över vilka produkter som säljs och vilka som har dålig åtgång. På detta sätt kan de sköta lagerhanteringen (Olson och Delen 2008, kap. 1).

Det finns ytterligare exempel på användning av datautvinning. Till exempel år 2004 användes datautvinning av båda partierna i USA:s presidentval för att få fram information om väljarna. Inom sjukvården har datautvinning används för att identifiera de metoder som gett bästa möjliga resultat vid vård av patienter. Vid Mayo-kliniken i USA har datautvinning används för att ta reda på hur vissa vårdmetoder hade fungerat på ett urval av deras senaste patienter (Olson och Delen 2008, kap. 1).

För att kunna utföra datautvinning krävs alltså att man identifierar ett problem man vill lösa, insamling av relevanta data och datamodeller för statistisk och annan analys. Det är även möjligt att visuellt representera data med hjälp av datorprogram eller grafiskt genom statistisk analys som t.ex. korrelationsanalys (Olson och Delen 2008, kap. 1).

2.2 Processer

I detta avsnitt presenteras och definieras de vanligaste datautvinningsprocesserna. De finns två stycken utvinningsprocesser som är av intresse för avhandlingen. Dessa två processer heter CRISP-DM och SEMMA. Utöver dessa diskuteras nio lagar eller principer som ligger till grund för datautvinning.

Datautvinning är en process med vars hjälp man försöker hitta en lösning på ett problem med ett klart mål i sikte. De som idkar utvinning sitter inte och mållöst siktar igenom data i hopp om att hitta någonting intressant. För att datautvinningen ska vara till nytta måste den som utför arbetet förstå företagets och beslutsfattarnas problem samt målsättning. Det gäller att utvinna information som ligger till grund för bra beslutsfattande. Den som inom företaget idkar och är ansvarig för datautvinning kan bidra med ett nytt perspektiv på företagets verksamhet och problem som ledningen inte har (Brown 2014, kap. 4).

Olson och Delen (2008, kap. 2) lyfter fram två olika sätt att gå till väga när man idkar datautvinning. Den första kallas för CRISP-DM och den andra för SEMMA. CRISP-DM står för Cross-Industry Standard Processing for Data Mining. SEMMA i sin tur står för Sample, Explore, Modify, Model, Assess. Dessa två sätt eller rättare sagt processer innehåller ett antal steg i datautvinningsprocessen (Olson och Delen 2008, kap. 2).

I avhandlingen används CRISP-DM-processen eftersom det är den som enligt Olson och Delen (2008, kap. 2) används brett inom industrin, medan Mariscal m.fl. (2010) karakteriserar CRISP-DM som en *de facto* standard för utveckling av datautvinnings-projekt. SEMMA i sin tur är en specifik process som används av endast en mjukvaruleverantör, SAS Institute (Olson och Delen 2008, kap. 2; Azevedo 2008).

CRISP-DM är utvecklat av datautvinnare för datautvinnare. Deltagare från över 200 organisationer har bidragit till utvecklingen av processen (Brown 2014, kap. 5). Bland dessa organisationer återfinns DaimlerChrysler, SPSS och NCR (Azevedo 2008).

Processen är indelad i sex faser och är avsedd som en cyklisk process. Enligt Olson och Delen (2008, kap. 2) är det inte nödvändigt att använda sig av alla faser i varje datautvinningsprojekt utan man använder sig av de faser man anser vara nödvändiga för projektet (Olson och Delen 2008, kap. 2).

SEMMA i sin tur är en process som är unik för SAS Institute och processen används i deras mjukvaror för datautvinning. Denna process är iterativ i motsats till cyklisk till sin natur vilket innebär att man utvärderar resultatet efter varje steg i processen och vid behov återgår till utforskningsskedet för att ytterligare förfina data (Olson och Delen 2008, kap. 2).

Den första fasen i CRISP-DM kallas *Affärskännedom* (eng. *Business Understanding*). I denna fas ska man förstå det problem som ska lösas, hur det påverkar organisationen och målsättningen för att lösa problemet. För att lösa problemet måste man utvärdera nuvarande situation, ställa upp konkreta mål för datautvinningen och göra upp en preliminär projektplan. Detta steg saknar motsvarighet i SEMMA (Olson och Delen 2008, kap. 2; Mariscal 2010).

I den andra fasen i CRISP-DM som kallas *Datakännedom* (eng. *Data Understanding*), går man igenom befintliga data, dokumenterar dem och identifierar möjliga problem med data. (Brown 2014, kap. 5) I denna fas kan man även utföra datainsamling, utforska data, beskriva data och utföra en kvalitetskontroll av dem. Man väljer alltså data man behöver för att lösa det problem som nämnts i föregående fas. Beroende på problemet behöver man olika sorters data som t.ex. demografiska data eller transaktionsdata (Olson och Delen 2008, kap. 2; Mariscal 2010). Utöver detta nämner Mariscal (2010) att man även kan hitta intressanta delmängder (eng. *subset*) av data och utifrån dem formulera hypoteser för dold information.

Denna fas motsvaras av det första steget i SEMMA som kallas för *Sampla* (eng. *Sample*). I detta steg använder man en del av alla data, ett statistiskt representativt sampel, för att använda samplet i ett senare steg. Orsaken till detta är att man vill

förkorta tiden för processen när man jobbar med väldigt stora dataset (Olson och Delen 2008, kap. 2; Azevedo 2008).

Det andra steget i SEMMA-processen kallas för *Utforska* (eng. *Explore*) och motsvarar även den det andra steget i CRISP-DM. Detta innebär att man söker efter avvikelser och oväntade trender i data för att få en bättre förståelse över datasetet. Samtidigt försöker man även hitta trender med hjälp av visuella och statistiska verktyg. Man kan använda sig av klusteranalys för att, genom att individuellt analysera klusterna, hitta trender som inte hittas genom att analysera hela datasetet (Olson och Delen 2008, kap. 2; Azevedo 2008).

I fasen som kallas *Dataförberedning* (eng. *Data Preparation*) i CRISP-DM förbereder man data inför användning vilket innebär att man, t.ex. fyller i värden som saknas eller tar bort dubbletter. (Brown 2014, kap. 5) I denna fas går man igenom datan för att se om misstag har krupit sig in den. Misstag som värden som är orealistiska, t.ex. att en person med en årsinkomst på 250000 dollar har klassats som fattig. Man bör komma ihåg att den mjukvara man använder sig av för datautvinning ställer vissa krav på hur data ska vara (Olson och Delen 2008, kap. 2; Mariscal 2010). Enligt Mariscal (2010) väljer man även ut de attribut, tabeller och dokument som används av verktyg för datamodellering.

Det motsvaras av det tredje steget i SEMMA, som kallas *Modifiera data* (eng. *Modify data*), där man väljer på vilka variabler man ska basera de modeller man tänkt använda för analysen. Förutom att välja variabler kan det krävas att man skapar nya variabler eller bearbetar existerande sådana. Bearbetningen baserar sig på de upptäckter man gjort i föregående steg som t.ex. avvikelser eller trender. Det kan röra sig om att man t.ex. grupperar kunder i olika segment. Man kan även ta bort variabler som är irrelevanta för analysen. Datautvinningsprocessen är dynamisk och iterativ vilket innebär att man tvingas att uppdatera metoderna och modellerna vartefter man får tillgång till nya data (Olson och Delen 2008, kap. 2; Azevedo 2008).

Den fjärde fasen i CRISP-DM kallas *Användande av modeller* (eng. *Modeling*). I denna fas använder man sig av matematiska metoder för att identifiera eventuella mönster i data. (Brown 2014, kap. 5) För att göra detta använder man sig av datorprogram. Först kan man göra en klusteranalys och en visuell analys av datan. Efter detta kan man använda olika matematiska modeller beroende på typen av data (Olson och Delen 2008, kap. 2; Mariscal 2010).

Denna fas motsvaras av det fjärde steget i SEMMA som kallas för *Skapande av modell* (eng. *Model*). I detta steg söker man efter den kombination av variabler som ligger till grund för den matematiska modell som kan svara på den frågeställning eller problem man vill att datautvinnings-processen ska svara på eller lösa. Det finns olika tekniker man kan använda för att konstruera modellen. Bland dessa tekniker finns artificiella neuronnät, beslutsträd, ungefärliga uppsättningar (eng. *rough sets*), stödvektormaskiner (eng. *support vector machines*) och logiska modeller. Man kan även använda sig av statistiska modeller som tidsserieanalys , minnesbaserat resonemang (eng. *memory-based reasoning)* och principalkomponentanalys (eng. *principal component analysis*). Valet av modell eller modeller avgörs utifrån de data man har tillgång till och vad man vill ta reda på (Olson och Delen 2008, kap. 2; Azevedo 2008).

Den femte fasen i CRISP-DM kallas för *Utvärdering* (eng. *Evaluation*). Detta innebär att man utvärderar de mönster som hittats och ser ifall de är användbara; man tolkar de mönster man hittat. (Brown 2014, kap. 5) Detta leder till att man får konkret information som man sedan kan tillämpa inom organisationens verksamhet. Man bör komma ihåg att korrekt tolkning leder till bra resultat medan inkorrekt tolkning leder till dåliga resultat (Olson och Delen 2008, kap. 2; Mariscal 2010).

Tillämpning (eng. *Deployment*) kallas den sista fasen i CRISP-DM. I denna fas tillämpas resultaten för att förbättra verksamheten. (Brown 2014, kap. 5) Man måste kunna se om man lyckats lösa det problem som det var tänkt att datautvinnings-processen skulle lösa. En annan sak man bör beakta är att de variabler som ligger till grund för datautvinningen t.ex. kundbeteende ändras med tiden. Vartefter tekniken

utvecklas och ekonomin förändras kommer kunderna att reagera på detta och modifiera sitt beteende (Olson och Delen 2008, kap. 2; Mariscal 2010).

Dessa två faser motsvaras av sista steget i SEMMA som kallas för *Utvärdera* (eng. *Assess*) där man utvärderar resultatets användbarhet och pålitlighet. Ett sätt som vanligtvis används när man utvärderar en datautvinningsmodell är tillämpning av modeller på data som inte används för att konstruera modellen. Alltså den data som man inte använde för samplet man tog i det första steget. Om modellen är lyckad borde den fungera lika bra på dessa data. Man kan även tillämpa modellen på data vars egenskaper man känner till. Exempelvis på data om kunder vars köpbeteende redan har undersökts (Olson och Delen 2008, kap. 2; Azevedo 2008).

Både SEMMA och CRISP-DM är breda ramverk som man tvingas anpassa för det man tänkt använda dem till. Likheter finns mellan dessa två ramverk. Likheterna kan illustreras med följande tabell:

CRISP-DM	SEMMA
Affärskännedom	Antar en välformulerad fråga
Datakännedom	Sampla Utforska
Dataförberedning	Modifiera data
Användande av modeller	Skapande av modell
Utvärdera Tillämpa	Utvärdera

Tabell 1. CRISP-DM och SEMMA (Olson och Delen 2008, kap. 2).

I tabellen ovan kan man se vilka steg inom CRISP-DM som motsvarar steg inom SEMMA. Först kan man se att det som inom CRISP-DM kallas för *Affärskännedom* i SEMMA motsvaras av ett antagande som egentligen inte är en del av processen. I SEMMA utgår man från att man redan har en välformulerad fråga som man vill ha svar på med hjälp av datautvinning innan man börjar tillämpa stegen. I CRISP-DM ingår frågeformuleringen som en del av det första steget. Det steg som benämnas som *Datakännedom* i CRISP-DM motsvaras av *Sampla* och *Utforska* i SEMMA. Data

preparation motsvaras av *Modifiera data* i SEMMA. *Användande av modeller* motsvaras i sin tur av *Skapande av modell* i SEMMA. Stegen *Utvärdera* och *Tillämpning* i CRISP-DM motsvaras av *Utvärdera* i SEMMA (Olson och Delen 2008, kap. 2).

Enligt Mariscal (2010) har både SEMMA och CRISP-DM en gemensam filosofi eftersom de båda strukturerar datautvinningsprocessen i steg som är kopplade till varandra. Trots dessa likheter finns det även vissa skillnader: SEMMA fokuserar mera på processens tekniska karaktär; CRISP-DM tar ett bredare perspektiv vad gäller datautvinningens affärsmål. Skillnaden illustreras redan i början av utvinnings-processen: SEMMA börjar med sampling; CRISP-DM börjar med ett affärsproblem som sedan blir ett tekniskt problem. På basen av detta kan man dra slutsatsen att CRISP-DM bättre representerar ett datautvinningsprojekt. SEMMA har inte uttryckligen ett steg där man tillämpar resultatet av utvinningsprojektet medan CRISP-DM har ett dylikt steg vilket bättre representerar ett verkligt utvinningsprojekt (Mariscal 2010).

Förutom dessa två processer finns även en uppsättning principer för hur man ska arbeta som datautvinnare. Dessa principer, som benämns lagar, är inte bindande för data-utvinnare utan tänkta som vägledning för arbetet med datautvinning. Dessa principer är utarbetade av Thomas Khabaza, en av pionjärerna inom datautvinning (Brown 2014, kap. 4).

Den första lagen av sammanlagt nio kallas för *Affärsmål*. Det är viktigt att man hela tiden har affärsmålen i tankarna när man idkar datautvinning. Det finns annars en risk att man stirrar sig blind på den teknik och de tekniska lösningar som riskerar att stå i strid med affärsmålen. Den andra lagen kallas för *Affärskunskap*. Detta innebär att man behöver affärskunskap för att hitta mening i data som man utvinner. Man måste veta något om affärsverksamheten i den organisation man bedriver datautvinning åt för att utvunna data ska ha någon mening. Den tredje lagen kallas för *Dataförberedning* (eng. *data preparation*). För att kunna använda data ur olika källor inom och utom ett företag

måste man förbereda det för användning. Mer än hälften av tiden för datautvinningens går till att förbereda data (Brown 2014, kap. 4; Khabaza 2010).

Den fjärde lagen kallas för *Den rätta modellen.* Det är viktigt att man hittar rätt matematisk modell för data man jobbar med. Man hittar denna modell genom att pröva olika modeller och sedan väljer den som passar bäst. Ingen modell är helt hundraprocentig utan man väljer den modell som bäst förklarar de mönster man hittat i data. Den femte lagen kallas för *Mönster*. Detta innebär att man söker efter betydelsefulla mönster, det vill säga betydelsefulla förhållanden mellan de variabler som finns i data. Möjligheten finns att man hittar mönster som inte har något av värde för affärsmålen och företagets verksamhet. Det är viktigt att inte ödsla tid på mönster som inte är relevanta för affärsmålen. Den sjätte lagen kallas för *Förstärkning*. Detta innebär att man genom datautvinning kan förstå sin affärsverksamhet bättre än utan den. Man förstärker alltså sin förståelse av affärsverksamheten (Brown 2014, kap. 4; Khabaza 2010).

Den sjunde lagen kallas för *Förutspående*. Vad detta innebär är att man genom att använda det man redan vet förutspår vad som kommer att hända. Alltså använder man de data man har för att förutspå vad som kommer att hända. Den åttonde lagen kallas för *Värde*. Detta innebär att värdet hos resultaten av datautvinning inte bestäms av den matematiska modellens noggrannhet eller stabilitet. Värdet kommer istället från modellens förmåga att leverera den bästa förutsägelsen. Man använder sig av försök för att komma fram till detta snarare än av statistisk teori. Den nionde och sista lagen kallas för *Förändring*. Med förändring avses det faktum att världen är i ständig förändring vilket innebär att den modell som fungerar bra idag kan vara värdelös imorgon. Detta innebär att de mönster man hittat i data kan förändras allt eftersom nya data kommer till (Brown 2014, kap. 4; Khabaza 2010).

2.3 Metoder

I detta avsnitt presenteras och definieras de vanligaste och mest grundläggande datautvinningsmetoderna. De metoder som är relevanta för avhandlingen presenteras mera ingående i nästa avsnitt. Meningen är att ge en kort översikt över de metoder som oftast förknippas med datautvinning.

2.3.1 Regressionsanalys

Regression är en term inom statistiken där man analyserar förhållandet mellan två eller fler variabler. Den mest grundläggande typen av regression är linjär regression. Detta innebär att man analyserar sambandet mellan två variabler, en beroende och en oberoende variabel. Man kan illustrera det med en tvådimensionell graf där den beroende variabeln är y-axeln och den oberoende är x-axeln. Regressionslinjen beräknas enligt formeln y = a +bx. Till exempel kan en bank förutspå hur mycket pengar en kund har på sitt bankkonto = 1000 € + 0.01 * kundens årsinkomst. Det är möjligt att använda fler än en oberoende variabel men då är det frågan om multipel linjär regression. Vidare kan man använda sig av potenser och kvadratrötter på de oberoende variablerna men då är det inte längre frågan om linjär regression utan om icke-linjär regression (Berson m.fl. 2000).

2.3.2 Klassifikation

Klassifikation är en av de vanligaste metoderna för datautvinning. Ett exempel på klassifikation kan vara ett sjukhus som vill klassificera patienter utifrån den risk de har för att insjukna i viss typs sjukdom. Patienterna kan klassificeras att ha en hög, medel eller låg risk för att insjukna i en viss sjukdom. Ett annat exempel kan vara en opinionsundersökning där de svarande klassificeras utifrån sannolikheten att de röstar på ett visst politiskt parti eller stöder en viss sak. I princip går klassificering ut på att dela upp objekt i uttömmande och exklusiva kategorier som benämns klasser. Dessa klasser är unika och de är meningen att ett objekt inte kan vara i flera klasser samtidigt eller helt sakna klass (Bramer 2016).

Naive Bayes är en klassificeringsmetod som använder sig av sannolikhetsteori för att hitta den klassifikation som passar bäst in på data. Man använder sig av denna metod när man är intresserad av sannolikheten att något av en uppsättning alternativ skall inträffa. Till exempel kommer flyget från Stockholm till Helsingfors vara a) inställt; b) försenat med mer än 10 minuter; c) försenat med mindre än 10 minuter; eller d) anlända inom utsatt tid. Naive Bayes metoden använder sig av två typer av sannolikheter för att göra detta, villkorlig sannolikhet (eng. *conditional probability*) och tidigare sannolikhet (eng. *prior probability*).Villkorlig sannolikhet skulle i detta exempel vara väder, vind eller årstid, medan tidigare sannolikhet skulle vara data om flygets ankomsttider under tidigare flygningar. Dessa två sannolikheter kombineras till en matematisk formel som sedan tillämpas på data (Bramer 2016).

Närmaste granne (*eng. Nearest Neighbour*) är även den en klassificeringsmetod. Denna metod används främst när det är frågan om attributvärden som är numeriska. Det går att använda denna metod, med modifikation, för kategoriska värden som t.ex. färger. Metoden går ut på att klassificera data på basen av den klassificerade data som ligger närmast till den nya oklassificerade. För att använda denna metod behöver man data som redan är klassificerat. Man kan bestämma att man baserar klassifikationen på de fem närmaste instanserna av data. Om en majoritet av instanserna har en viss klassifikation, t.ex. positiv, får den nya oklassificerade instansen klassen positiv (Bramer 2016).

2.3.3 Beslutsträd

Beslutsträd är en prediktiv modell som kan illustreras som ett träd. Modellen gör en förutsägelse på basen av ett antal beslut eller frågor som sedan resulterar i att modellen förgrenar sig på basen av svaren. Dessa svar är mycket enkla av typen ja eller nej. I bilden (Bild 1.) nedan illustreras ett typiskt beslutsträd där svaren på frågorna är antingen ja eller nej. Det handlar om hur lojala kunderna är till en mobiloperatör. Illojala är de som inte förnyat sitt mobilabonnemang och lojala är de som förnyat sitt abonnemang (Berson m.fl. 2000).

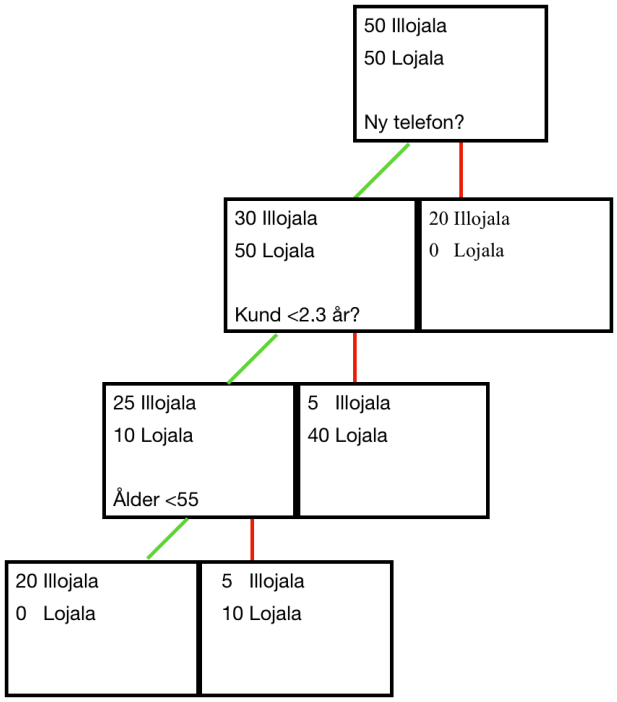

Bild 1. Beslutsträd (Berson m.fl. 2000).

Man kan konstatera på basen av figuren ovan att de som har en modern mobiltelefon och haft sitt abonnemang i mer än 2.3 år är rätt lojala. Mobiloperatören kan använda denna metod för att identifiera de kunder som sannolikt inte kommer att förnya sina abonnemang i framtiden. Operatören kan sedan rikta marknadsföring och erbjudanden till kunderna (Berson m.fl. 2000).

Associationsregler (eng. *Association Rules)* är en metod där man försöker hittar samband och associationer i stora dataset. Evans (2017, s. 358) lyfter fram ett exempel som illustrerar den möjliga nyttan företag kan ha av denna metod. Exemplet handlar om försäljning i dagligvaruhandeln. Ett företag i den branschen kan vara intresserad av vilka produkter som konsekvent köps tillsammans. Data över försäljning samlas in via streckkoder som avläses i samband med att kunden köper varan. Data lagras sedan i en databas och man kan sedan m.h.a. datautvinning få fram den informatorn om försäljningen man behöver. Med denna information i handen kan företaget optimera butikens eller butikernas layout så att kunderna lättare hittar produkterna (Evans 2017, s. 358).

Det finns även avancerade datautvinningsmetoder men de är inte av intresse för avhandlingen. Olson och Delen (2008) diskuterar i sin bok *Advanced Data Mining Techniques* flera olika avancerade metoder som t.ex. *Ungefärliga uppsättningar* (eng. *Rough sets*)*, Oskarpa uppsättningar* (eng. *Fuzzy sets*) och *Stödvektormaskiner* (eng. *Support Vector Machines).*

Metoderna i detta kapitel har även använts inom bibliomining. Azam m.fl. (2013) använde sig av association rules-metoden när de undersökte vilka typer av böcker som lånats hos biblioteket vid North South universitetet i Bangladesh. Azam m.fl. (2013) var intresserade av vilka kategorier av böcker som lånades tillsammans för att sedan kunna använda denna information för att bättre placera böcker på hyllor och för att kunna förbättra bokrekommendationerna. Förutom detta ville Azam m.fl. (2013) få demografisk information om biblioteksanvändarna.

2.4 Kluster, neuronnät och självorganiserande kartor

Detta avsnitt handlar om de metoder som är av intresse för avhandlingen. Dessa metoder är kluster, neurala nätverk och självorganiserande kartor (eng. Self-organizing maps). Den metod som avhandlingen använder sig av är självorganiserande kartor, men eftersom både kluster och neurala nätverk hör till samma kategori som självorganiserande kartor behandlas även de i detta avsnitt.

2.4.1 Klusteranalys

Klusteranalys är ett sätt att analysera innehållet i en databas och ge en överblick över det. Metoden går ut på att gruppera posterna i databasen i kluster. Ett kluster i detta fall är en grupp poster som har gemensamma attribut som t.ex. ålder, inkomst eller utbildning. Detta kan appliceras sedan t.ex. inom marknadsföring där klusterna representerar olika kundsegment. Ett kundsegment kan t.ex. vara högutbildade kvinnor i åldern 40-50 år som bor i förorter. (Berson m.fl. 2000) Ett kluster är alltså en gruppering poster, objekt eller observationer som är nära relaterade till varandra. Objekten i ett

kluster bör vara likartade, medan objekt som inte är inom samma kluster bör vara olikartade (Evans 2017, s. 336).

Själva idéen med klusteranalys är att den kan ta stora mängder observationer och reducera dem till mindre homogena grupper eller kluster som sedan är lättare att tolka. Till skillnad från andra analysmetoder är klusteranalys i huvudsak deskriptiv och man kan inte dra statistiska slutsater av dess resultat. Kluster som identifierats med denna metod är inte unika och beror på hur man utför analysen. Klusteranalys ger ett nytt perspektiv på data, men inte ett definitivt svar (Evans 2017, s. 336).

Enligt Evans (2017, s. 336) finns det två huvudsakliga sätt eller metoder för att tillämpa klusteranalys. Den ena är hierarkisk klusteranalys och den andra K-means klusteranalys. Hierarkisk klusteranalys går ut på att data inte delas in i särskilda kluster i ett enda steg utan i stället till följd av en serie steg. Detta kan göra på två sätt antingen börjar man med att all data är i ett enda kluster och delar in data stegvis i flera kluster eller att alla observationer, poster eller objekt var för sig bildar ett kluster och sedan sammanslår man klusterna (Evans 2017, s. 336). K-means klusteranalys i sin tur innebär att man beräknar antalet kärnor runt vilka klustren grupperas. Kärnorna representeras av bokstaven k, vilket framgår av namnet på metoden. Sedan beräknar man medelvärdet av de euklidiska avstånden mellan observationerna. På basen av dessa beräkningar får man fram klusterna för ett visst dataset (Pierson 2017, kap. 6).

Enligt (Berson m.fl. 2000) kan man även använda klusteranalys för att få fram de objekt eller poster inom data som ligger utanför resten av data. Dessa objekt kallas för uteliggare (eng. *outliers*). Ett exempel på nyttan med att använda klusteranalys i detta syfte är ett företag i Kalifornien som sålde kostymer till rabattpris under en kampanj. Alla butiker som hörde till företaget hade ökat sin försäljning under kampanjen, utom ett. Denna butik hade, till skillnad från de andra butikerna, gjort reklam i radio och inte i teve som de andra butikerna och därmed sorterats i sitt eget kluster (Berson m.fl. 2000).

2.4.2 Artificiella neuronnät

Enligt Salvio (2018) kan neurala nätverk beskrivas som digitaliserade neuroner som fungerar lite i samma stil som de biologiska neuronerna i mänskans hjärna. Berson m.fl.

(2000) använder samma beskrivning när de diskuterar neuronnät. Vidare hävdar Berson m.fl. (2000) att en bättre beskrivning vore artificiella neuronnät eftersom äkta neurala nätverk endast existerar inom biologin i form av hjärnor. Liksom hjärnor upptäcker artificiella neuronnät mönster, gör modeller för prediktion och lär sig. Med att lära sig avses att artificiella neuronnät använder sig av maskinlärande (eng. *machine learning*) algoritmer som appliceras på data i stora databaser för prediktion. (Berson m.fl. 2000) För att ytterligare förenkla detta kan man säga att artificiella neuronnät lär sig av tidigare data för att förutspå något om framtiden.

De algoritmer som artificiella neuronnät använder sig av för att lära sig skiljer sig inte nämnvärt från statistik och de algoritmer som används för beslutsträd. Skillnaden mellan t.ex. beslutsträd och artificiella neuronnät finns i hur de behandlar historisk data som används när de lär sig och upptäcker mönster. Beslutsträd behandlar alla instanser av data på en gång medan artificiella neuronnät behandlar en instans åt gången (Berson m.fl. 2000).

Nätet består av neuroner som motsvarar biologiska neuroner i hjärnan. Flera neuroner är sammanbundna av länkar. På Bild 2 kan man se hur ett enkelt neuronnät ser ut (Berson m.fl. 2000).

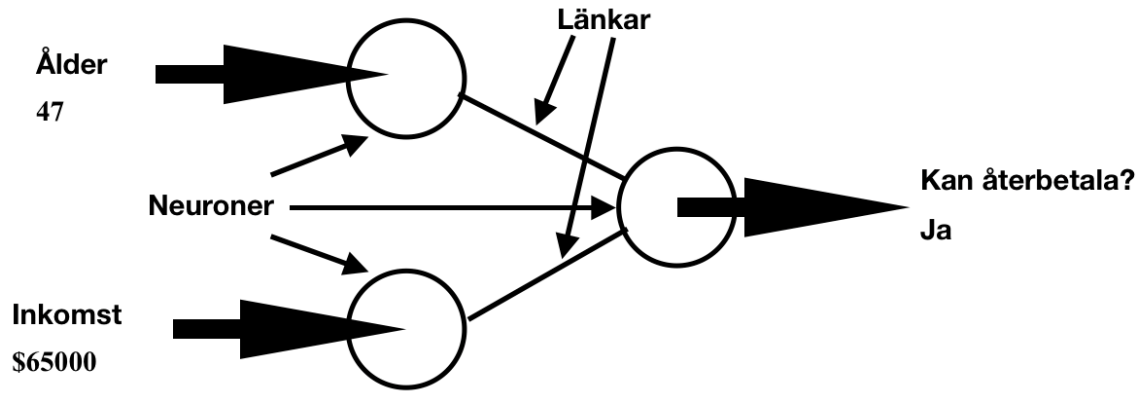

Bild 2. Artificiellt neuronnät (Berson m.fl. 2000).

I neuronerna till vänster finns värdena för ålder och inkomst, i neuronen till höger finns värdet av prediktionen av de två värdena. Prediktionen som nätet gör i detta fall handlar om huruvida en person i åldern 47 år som har en årsinkomst på $65000 kan återbetala

ett lån. Neuronerna till vänster kallas för indata neuroner och neuronen till höger kallas för en output neuron. Värdena för ålder och inkomst är sparade i indata neuronerna, dessa värden multipliceras sedan med värden som är sparade i länkarna. Med länkar avses de sträck som finns mellan neuronerna i figuren ovan. Utdata neuronen spottar ut värden mellan 0.0 och 1.0 där ett värde närmare 0.0 betyder att personen kan betala tillbaka lånet och ett värde närmare 1.0 att personen inte kan betala tillbaka lånet. Eftersom de värden som Utdata neuronerna spottar ut är numeriska måste man kunna översätta dessa värden till kategoriska om man vill veta, som i fallet i figuren ovan, om en person kan betala tillbaka ett lån eller inte (Berson m.fl. 2000).

Förutom indata neuroner och utdata neuroner skapar nätverket gömda neuroner som finns mellan input och utdata neuroner. I de gömda neuronerna sker lärandet inom neuronnät. De gömda neuronerna har, till skillnad från input och output neuronerna, ingen på förhand definierad funktion utan deras funktion bestäms av av neuronnätet vartefter det lär sig. De fungerar som rådgivare till utdata neuronerna i det hänseendet att den information de ger ligger till grund för den output som utdata neuronerna ger ut. Vissa neuroners råd väger tyngre än andra och detta illustreras genom det viktvärde som multipliceras till det värde som den gömda neuronen har. Viktvärdet är den betydelse som en viss gömd neuron har för den slutgiltiga prediktionen som utdata neuronen ger ut. Viktvärdet justeras på basen av de historiska data som används för att träna neuronnätet och det är just det som avses med lärandet i neuronnät. Med hjälp av detta konstrueras sedan modeller som appliceras på ny data (Berson m.fl. 2000).

2.4.3 Självorganiserande kartor

Den sista metoden som detta avsnitt tar upp är självorganiserande kartor eller SOM. Förkortningen SOM härstammar från engelskans Self-Organizing Map. Det är denna metod som är av största intresse för avhandlingen. Metoden används i avhandlingen för att analysera data från olika allmänna bibliotek i Finland.

Enligt Kohonen (2001) är en självorganiserande karta en icke-linjär, ordnad (eng. *ordered*), smidig (eng. s*mooth*) kartläggning (eng. *mapping*) av högdimensionell indata

som sedan representeras visuellt i två dimensioner. Yao (2013) definierar självorganiserande kartor som en icke-övervakad artificiell neuronnätsbaserad metod för visuell klusteranalys. Vesanto (1999) definierar i sin tur självorganiserande kartor som en artificiell neuronnäts algoritm baserat på oövervakat lärande (eng. *unsupervised learning*).

Självorganiserande kartor har visat sig lämpa sig för datautvinning, KDD (eng. *Knowledge Discovery in Databases*) inom såväl fulltext som analys av finansiell data. Man har kunnat konstatera att självorganiserande kartor även lämpar sig för ingenjörer inom upptäckandet av mönster, bildanalys, processövervakning och feldiagnostik. Den största fördelen med självorganiserande kartor är att man kan visualisera komplex data vilket gör data enkelt att förklara och förstå (Vesanto 1999).

Processen för skapandet av självorganiserande kartor

Av den grundläggande algoritmen för självorganiserande kartor finns det två versioner (Vesanto m.fl. 2000, ss.7-9):

1) *Sekventiell träningsalgoritm* (eng. *sequential training algoritm)*, med denna algoritm sker träningen av en självorganiserande karta vartefter ny indata introduceras.
2) *Batch träningsalgoritm* (eng. *batch training algoritm*), med denna algoritm, till skillnad från den förra algoritmen, sker all träning samtidigt eftersom all indata introduceras på en gång.

Enligt Kohonen (2013) är *batch träningsalgoritm* den algoritm som rekommenderas eftersom den *sekventiella träningsalgoritmen* är avsedd för jämförelse med andra självorganiserande modeller. Av denna orsak används *batch träningsalgoritmen* i avhandlingen. Enligt Yao (2013, s.35) finns det två fördelar med denna algoritm: 1) den är mycket mer effektiv ur ett resursperspektiv eftersom kartan uppdateras endast en gång och inte efter varje enskild instans av data; 2) det är lätt att reproducera kartorna, förutsatt att man påbörjar processen på samma sätt.

En karta bestående av en skara (eng. *array*) tvådimensionella neuroner eller noder skapas till först innan man börjar med själva analysprocessen. Skaran kan antingen vara rektangulär eller hexagonal till formen (Holmbom 2015, s. 58). Vesanto (1999) förklarar saken med att säga att neuronerna befinner sig på ett rutsystem, vanligtvis en- eller tvådimensionellt. Högre dimensioner är möjliga men ställer till problem vid visualisering. Själva gallret i rutsystemet kan antingen vara rektangulärt eller hexagonalt. Den hexagonala formen är den bästa formen för visualisering (Vesanto 1999).

Syftet med studien av data, alltså vad man vill får fram ur data, bestämmer hur stor kartan skall vara och hur många neuroner (noder) som behövs. Exempelvis en stor karta med tusentals neuroner är mera lämpad för visualisering medan en liten karta med färre neuroner är mera lämpad för klusteranalys. Neuronerna som finns i indatalagret binds samman med neuronerna i utdatalagret eller kartan (Holmbom 2015, s. 58). Varje utdataneuron på kartan har en associerad referensvektor som benämns m_i (Eklund 2004, s. 60). Vesanto (1999) beskriver det som att varje neuron är representerad av en vektorprototyp.

Efter att indata och utdata neuronerna kopplats samman kan man börja träna den självorganiserande kartan (Kohonen 2001, s. 142). Denna algoritm använder sig av två steg för att åstadkomma en representativ karta på basen av indata (Holmbom 2015, ss. 58-59; Eklund 2004, ss. 60-61):

1) Indata neuronerna x jämförs med referensneuronerna m_i tills den neuron som bäst motsvarar indataneuronen m_c har hittats. Denna neuron m_c benämns också BMU (eng. *Best Matching Unit).*
2) Lärande, alltså justera de neuroner som omringar m_c eller grannskapet h_{ci} mot indataneuronen x.

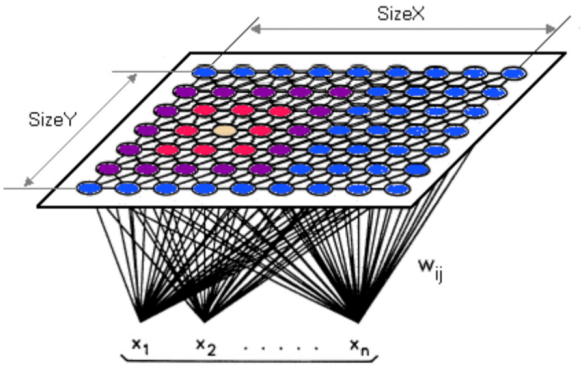

Bild 3. En självorganiserande karta (Demirhan och Güler 2011 i Yao 2013).

I Bild 3 finns indatavektorerna längst nere i figuren och representeras av bokstaven *x* och hör till indatalagret. De hexagonala färgade rutorna representerar utdatalagret. Steg 1 går ut på att hitta den neuron m_c eller BMU (den gula rutan i Bild 3) som bäst passar in på indatavektorn *x*. Detta bestäms på basen av något lämpligt distansmått t.ex. minsta euklidiska avståndet (eng. *Euclidian distance)* enligt $x - m_i$ (Eklund 2004, s. 61). Man kan även uttrycka detta steg som att all indata x_j (där j = 1,2,3…N) körs in i SOM-nätverket, utdataneuronerna m_i (där i = 1,2,3,…M) tävlar sedan om att bli vinnare eller BMU som också benämns m_c. (Holmbom 2015 s. 59) Den vinnande neuronen m_c hittas med formeln (Kohonen 2001, s.110):

Formel 1 $$\|x_j - m_c\| = \min_i\{\|x_j - m_i\|\}$$

Steg 2 kan påbörjas när m_c har hittats. Steget kallas för lärande steget. Inom detta steg justeras de neuroner som omringar m_c mot indatavektorn *x*. Neuroner med ett visst avstånd h_{ci} aktiverar varandra och lär sig något av indatavektorn *x* (Eklund 2004, s. 61). Lärandet kan beskrivas med formeln (Kohonen 2001, s. 111):

Formel 2. $$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)]$$

Formel 2 är den formel som beskriver den *sekventiella träningsalgoritmen*. När man vill använda sig av *batch träningsalgoritm* så använder sig man av Formel 3 där *c(j)* representerar BMU som hör till indatavektor *j* och $m_i(t)$ representerar referensvektorn

för varje indatavektor vid tiden *t* (t = 1,2,3…T) (Holmbom 2015, s.59).

Formel 3 $$m_i(t+1) = \frac{\sum_{j=1}^{N} h_{ic(j)}(t) x_j}{\sum_{j=1}^{N} h_{ic(j)}(t)}$$

Resultatet av denna inlärning innebär att alla BMUs lägen på kartan justeras närmare indatavektorerna enligt Formel 3. Förutom att BMUs justeras så justeras även de neuroner som ligger omkring BMUs eller grannskapet (de röda och violetta rutorna i Figur 4) enligt en jämt minskande faktor (Kohonen 2001, s. 110-111). Enligt Kohonen (2001, s. 111) kan man beskriva denna inlärningsprocess enligt Formel 4 där r_c representerar koordinaten för BMU, r_i representerar koordinaten för referensvektorn och $\sigma(t)$ radien på grannskapet:

Formel 4 $$h_{ic(j)} = \alpha(t) * exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right)$$

För att enkelt illustrera vad processen för skapandet av en självorganiserande karta går ut på kan man säga att varje indata instans skall välja den modell (BMU) som bäst passar indata. Vidare skall denna modell och dess grannar modifieras så att de bättre passar in på indata. (Kohonen 2013) Detta görs sedan kontinuerligt tills man uppnått ett kriterium som gör att processen stoppas. Kriteriet kan antingen vara att ett visst på förhand bestämt antal försök eller att justeringarna i steg två är tillräckligt små (Eklund 2004, s. 61).

Enligt Vesanto (1999) beter sig den självorganiserande kartan, under den iterativa träningsprocessen, som ett flexibelt nät som viker sig runt det moln som data skapar. De vektor prototyper som skapas kan tolkas som villkorliga medelvärden för indata. Eftersom grannskapet h_{ci} används för att bestämma storleken på uppdateringen leder detta till att vektorprototyperna för närliggande neuroner börjar likna varandra. Detta innebär att BMUs för liknande indata sampler ligger nära varandra i kartans rutsystem (Vesanto 1999).

Analys av självorganiserande kartor

För att kunna analysera en självorganiserande karta visualiserar man den. En av de vanligaste sätten är en *U-matris* (eng. *Unified distance matrix*). Denna matris beräknas på basen av medeltalet av avståndet för en neurons referensvektor och jämföra det med närliggande referensvektorer (Eklund 2004, s. 64). Enligt Holmbom (2015, s. 59) representerar en U-matris medelavståndet mellan referensvektorerna för neuroner som är belägna bredvid varandra. Formen på matrisen bestäms av kartans topologi där lanskapet som matrisen lägger på kartan är tredimensionell. Mörkare färger representerar höjder, alltså där avståndet mellan referensvektorerna är längre och ljusare färger där avståndet är kortare. Kluster som har ljusare färger mellan sig är närmare varandra än de som har mörkare färger mellan sig (Holmbom 2015, ss. 59-60).

Förutom *U-matris* kartor finns det även andra sätt att visualisera en självorganiserande karta. Man kan använda sig av komponentkartor (eng. *component planes*). I denna visualiseringsmetod konstrueras en komponentkarta för varje enskild input variabel. Varje neuron i samma position i olika komponentkartor representerar samma data. En uppsättning komponentkartor gör det möjligt att se flera länkade vyer över samma data ur olika variablers perspektiv. Vyerna representeras i färg där varma färger indikerar högre värden och kalla lägre (Yao 2013, s. 37; Vesanto 1999). Man kan även använda dessa kartor för att identifiera olika egenskaper hos kluster genom att se vilka variabler som karaktäriserar vissa kluster. (Eklund 2004, ss. 64-65). Enligt Vesanto (1999) är det möjligt att m.h.a. komponentkartor se möjliga korrelationer mellan olika variabler.

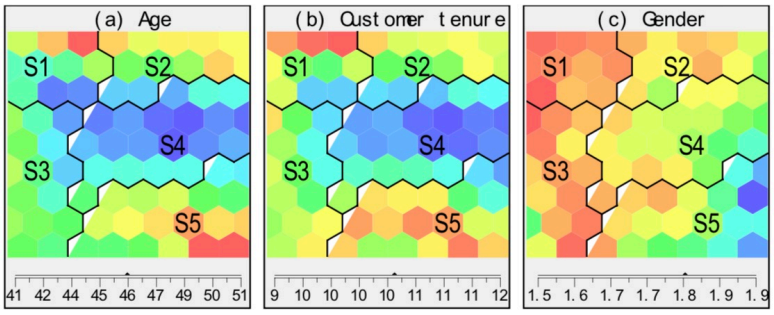

Bild 4. Komponentplan för tre variabler (Yao m.fl. 2012 i Yao 2013).

Bild 4 illustrerar tre komponentplan för tre variabler: a) Ålder; b) Längden på kundförhållande; c) Kön. Man kan jämföra komponentplanen med varandra för att se möjliga likheter eller olikheter. Varma färger anger högre värden medan kalla anger lägre.

Kvaliteten på SOM-kartor

En självorganiserande kartas kvalitet kan mätas på tre sätt genom att beräkna *kvantiseringsfel* (eng. *average quantization error*), *topologiskt fel* (eng. *topological error*) och *förvrängningsfel* (eng. *distortion error*) (Holmbom 2015, s. 62; Desmet 2001, s. 23; Yao 2013, s. 36).

Kvantiseringsfel är det vanligaste sättet och representerar medeldistansen mellan BMUs och indatavektorerna (Eklund 2004, s. 63). Holmbom (2015, s. 62) och Desmet (2001, s. 23) uttrycker det som att *kvantiseringsfel* är ett mått på den kvadrerade avvikelsen mellan indatavektorn x_j och BMUs referensvektor $m_{c(j)}$. Målet är att minska *kvantiseringsfel*. Formeln för beräknandet av *kvantiseringsfel* (Desmet 2001, s. 23; Pölzlbauer 2004):

Formel 5

$$\varepsilon_q = \frac{1}{N}\sum_{j=1}^{N}\left\|x_j - m_{c(j)}\right\|$$

I Formel 5 representerar N antalet indata sampler, x_j indata vektorn och $m_{c(j)}$ BMUs referensvektor (Desmet 2001, s. 23). Holmbom (2015, s. 62) påpekar att många mjukvarupaket beräknar detta kvalitetsmått automatiskt. Enligt Yao (2013, s. 36) minskar *kvantiseringsfel* i jämn takt vartefter kartan görs större.

Det andra kvalitetsmåttet är *topologiskt fel* som är ett mått på hur väl topografin i en SOM-karta är bevarad. Man definierar detta som andelen av alla indatavektorer vars BMU och näst bäst passande enheter inte är angränsande. (Yao 2013, s. 36; Kiviluoto 1996) Formeln för beräknande av *topologiska fel* (Desmet 2001, s. 23):

Formel 6 $$\varepsilon_t = \frac{1}{N}\sum_{j=1}^{N} u(x_j)$$

I Formel 6 anger *N* det totala antalet indata sampler. Om uttrycket $u(x_j) = 1$ innebär det att BMU och näst bäst passande enhet inte är angränsande och om $u(x_j) = 0$ innebär det att de är angränsande (Holmbom 2015, s. 62: Desmet 2001, s. 23).

Det tredje kvalitetsmåttet är *förvrängningsfel.* Enligt Yao (2013, s.36) liknar *förvrängningsfel* väldigt mycket *kvantiseringsfel* med den olikheten att *förvrängningsfel* tar hela grannskapet runt en BMU i beaktande. *Förvrängningsfel* består av tre komponenter: *kvantiseringsfel,* grannskapspartiskhet (eng. *neighborhood bias*) och grannskapets varians. Formel 7 visar hur måttet beräknas (Pölzlbauer 2004):

Formel 7 $$\varepsilon_d = \frac{1}{N}\frac{1}{M}\sum_{j=1}^{N}\sum_{i=1}^{M} h_{ic(j)} \left\| x_j - m_{c(j)} \right\|$$

I Formel 7 anger *N* det totala antalet indata sampel, *M* anger totala antalet output enheter, $h_{ci(j)}$ anger grannskapet, x_j anger indata vektor och $m_{c(j)}$ anger BMU (Pölzlbauer 2004; Holmbom 2015, s. 63).

För att få det mest pålitliga resultatet kan man träna kartan ett flertal gånger. När man gör detta använder man samma data varje iteration av träning för att finjustera parametrarna och bearbetningsätt. På grund av detta sker det ändringar av *kvantiserings-, topologiska* och *förvrängningsfel* mellan de olika självorganiserande kartorna. Det finns inget exakt korrekt värde som skulle påvisa den bästa kartkvalitén. Man kan ändå ha nytta av dessa mått för att utvärdera en karta (Kohonen 2001, s. 161; Holmbom 2015, s. 63).d

Det är möjligt att validera den slutgiltiga självorganiserande kartan genom korsvalidering med tio veck. När man korsvaliderar en karta tränar man kartan tio gånger med samma parametrar med 90% av data eller nio veck och 10% eller ett veck

används för testande. I fall skillnaderna i felgraden är små (ca 1%) kan man anse att kartan är tekniskt sett validerad (Holmbom 2015, s. 63).

Varianter av SOM

Det är inte meningen att redogöra för alla varianter av självorganiserande kartor i avhandlingen utan bara ge ett smakprov på två utvalda exempel. De två exemplen är WSOM och SOTM. Den först nämnda förkortningen står för *viktad självorganiserande karta (eng. Weighted Self-Organizing Map)* och den andra förkortningen står för *självorganiserande tidskarta* (eng. *Self-Organizing Time Map*).

En viktad självorganiserande karta är en anpassning av den vanliga självorganiserande kartan med fokus på de variabler som användaren vill framhäva. Liknande som den vanliga processen för skapande av självorganiserande kartor använder sig viktade självorganiserande kartor av två steg för varje iteration. Skillnaden är att varje referensvektor justeras enligt en viktmotsvarighet för att framhäva vissa vektorer. Formel 8 beskriver hur detta sker (Holmbom 2015, s. 64):

Formel 8

$$\mathbf{m_i(t+1)} = \frac{\sum_{j=1}^{N} \mathbf{w_j h_{ic(j)}(t) x_j}}{\sum_{j=1}^{N} \mathbf{w_j h_{ic(j)}(t)}}$$

I formeln anger vikten w_j hur viktig indatavektor x_j är för inlärningen. Index j anger indatavektorerna som hör till neuron c och N anger antalet indata vektorer. Eftersom viktade självorganiserande kartor är lika självorganiserande kartor kan man använda sig av samma visualiseringsmetoder (Holmbom 2015, s. 64).

Den andra av de två exemplen på varianter av självorganiserande kartor är självorganiserande tidskartor. Enligt Sarlin (2013a; 2013b) och Holmbom (2015, ss. 64-65) kan man säga att självorganiserande tidskartor har liknande egenskaper som självorganiserande kartor, men de möjliggör beaktandet av tidsmässiga förändringar i datas struktur. Istället för tid kan man även använda en annan variabel. När en självorganiserande tidskarta skapas följer man samma procedur som för en vanlig självorganiserande karta. Det vill säga först identifieras BMUs och sedan uppdateras

dess referensvektor m.h.a. en tidsbegränsad version av formeln som används i steg två i träningsprocessen för vanliga SOM-kartor (Holmbom 2015, s. 65):

Formel 9

$$\mathbf{m_i(v)} = \frac{\sum_{j=1}^{N(v)} \mathbf{h_{ic(j)}(v) x_j(v)}}{\sum_{j=1}^{N(v)} \mathbf{h_{ic(j)}(v)}}$$

När man visualiserar en självorganiserande tidskarta kan man i viss utsträckning använda sig av samma metoder som för en vanlig självorganiserande karta. När man vill visualisera strukturen på kluster använder man sig av Sammons (1969) kartering (Holmbom 2015, s. 65).

Exempel på användning av självorganiserande kartor

Sarlin och Eklund (2013) har gjort en analys av europeiska bankers finansiella prestationsförmåga i skenet av den senaste bankkrisen. För att göra denna analys har självorganiserande kartor använts. Närmare bestämt en visuell flerdimensionell och temporal finansiell prestationsanalys. Problemet som Sarlin och Eklund (2013) försöker lösa är att bankernas stresstest inte säger något om deras finansiella prestationsförmåga (Sarlin och Eklund 2013).

För att lösa problemet skapades först en självorganiserande karta enligt det sätt som beskrivits i avsnitt 2.4.3 av avhandlingen. Det vill säga indatavektorerna jämförs med referensvektorerna tills en BMU för varje indatavektor har hittats och referens-vektorerna i grannskapet justeras mot BMU. Efter att träningen var genomförd utfördes en klusteranalys. Förutom att göra en vanlig kluster analys där vara indata instans är klart inkluderat i ett visst kluster suddades klustreringen vilket innebär att varje indata instans tillskrevs en grad av tillhörighet till varje kluster. Syftet med att göra klustreringen suddig var för att göra en temporal analys av data. Under en viss tidsperiod kan datapunkter förflytta sig från ett kluster till ett annat (Sarlin och Eklund 2013).

Data som analysen gjordes på hämtades från Banscopes finansiella databas. Datasetet innehöll, innan förbearbetning, 1236 europeiska banker från tidsperioden december 1992 till december 2008. Av dessa valdes ut endast banker från EU. På denna data utfördes analysen som resulterade i komponentplan indelade i kluster. Egenskapen hos klustren varierade från att representera banker som var i gott skick till de som hade det sämre ställt. Det var också möjligt att utläsa individuella bankers tidsvandring från ett kluster till ett annat. Hur hela länders bankväsen vandrat från ett kluster till ett annat under samma tidsperiod var också möjligt. Resultatet av analysen var en fungerande modell över hur självorganiserande kartor kan användas för att utvärdera bankers finansiella prestanda (Sarlin och Eklund 2013).

Li m.fl. (2018) har i sin tur använt självorganiserande kartor för att klassificera vattenkvaliteten vid kustområdet vid Taiwansundet i Kina. Tillsammans togs 282 vattenprov på 94 ställen i området under sommaren och hösten 2012. Proven togs i fyra etapper under den tiden. Bland annat mättes halten tungmetaller, syre, olja, temperatur, salthalt och näringsämnen. (Li m.fl. 2018)

Efter att proven tagits analyserade man resultaten m.h.a. självorganiserande kartor. Komponentplan användes för att visualisera variablerna som representerade vatten-kvalitén där höga värden representerades med varma färger och låga med kalla. Det gjordes även en klusteranalys på basen av kartorna. På basen av analysen kunde det konstateras att föroreningarna hade en unik geografisk spridning. Vidare kunde det konstateras att vissa föroreningar kunde härledas till mänsklig aktivitet som jordbruk och industri. (Li m.fl. 2018)

2.6 Sammanfattning

I kapitlet har de relevanta begrepp och metoder som är relaterade till datautvinning och begreppet datautvinning som ligger till grund för avhandlingen diskuterats. Den metod som har behandlats mest ingående är självorganiserande kartor eftersom det är den metod som används i avhandlingen. Förutom utom att beskriva hur kartorna görs

presenterades även olika sätt att visualisera dem samt alternativa sätt att tillämpa självorganiserande kartor.

3. BIBLIOTEKSDATA

Begreppen som diskuteras i kapitlet är *bibliometri, biblioteksdata* och *bibliomining*. Till först kommer begreppet *bibliometri* att definieras och diskuteras. Begreppet handlar om statistisk analys av olika delar av bibliotekens verksamhet. Avsnittet *biblioteksdata* i sin tur handlar om vilken typ av data biblioteken besitter och vad de består av.

Bibliomining är nära förknippat med datautvinning, men ur ett bibliotekscentriskt perspektiv. *Bibliomining* används för att utvinna data ur det som faller inom begreppet biblioteksdata. Det är även ett begrepp som kommer att definieras och presenteras i detta kapitel. I samband med *bibliomining* kommer de olika metoderna som används för analys av biblioteksdata att definieras.

3.1 Bibliometri

Begreppet bibliometri definierades på 1960-talet av Alan Prichard (1969). Enligt Prichard (1969) definieras bibliometri som tillämpningen av matematik och statistik på böcker och annan kommunikationsmedia. Enligt Ball (2018) är bibliometri ett sätt att kvantifiera mänskors eller institutioners vetenskapliga produktion. Ordet produktion avser här vetenskapliga artiklar, bokkapitel, böcker och konferenspapper. Det som faller utanför bibliometri, men som också används för att mäta akademisk produktion, är akademiska avhandlingar, hur man har lyckats få extern finansiering, antalet patenter, utställningar, medlemskap i olika kommittéer, studerande per professor och antalet utvärderingar (Ball 2018, kap 3; De Bellis 2009, kap 1).

Genom att mäta antalet hänvisningar (eng. *citations*) som en forskare, vetenskaps man eller publikation har är det möjligt att avgöra hur produktiv en forskare eller institution är. Bibliometri kan användas som ett verktyg för att utarbeta den policy som styr forskning och allokering av resurser för forskare och institutioner. Bibliometri har utarbetas p.g.a. behovet att motivera varför man finansierar och bedriver en viss sorts forskning (Ball 2018, kap 3; De Bellis 2009, kap 1).

I ett nötskal kan man definiera bibliometri som kvantifiering av skriftlig akademisk produktion. Man kan klassificera personer och institutioner på basen av bibliometriska indikatorer som man sedan kan presentera som rankningar. Den statistik som ligger till grund för bibliometrin ger indirekta indikationer om individers, institutioners och länders vetenskapliga kvalitet (Ball 2018, kap 6; De Bellis 2009, kap 2).

När man mäter vetenskaplig produktion, som kallas outputanalys, adderar man ihop alla akademiska publikationer som producerats av individer, institutioner eller nationer. Det ger en uppfattning om produktivitet utan att beröra kvaliteten på de akademiska artiklarna. En individ som publicerar ofta och mycket är inte nödvändigtvis en som producerar artiklar av god kvalitet. För att kunna säga något om kvaliteten måste man beakta andra faktorer också (Ball 2018, kap 3).

För att skapa sig en uppfattning om akademiska artiklarnas kvalitet använder man sig av resonansanalys. Detta innebär att man undersöker hur många gånger en artikel blivit hänvisad till under en viss tidsperiod. Ju fler gånger en artikel har blivit hänvisad till desto viktigare och bättre är den. Här antas det att det finns ett samband mellan kvalitet och antal gånger en artikel blivit hänvisad till. Man använder sedan statistiska beräkningar för olika index, rankingar och riktmärken. Till exempel vem är den mest produktiva vetenskapsmannen inom ett visst område eller vilka artiklar har de flesta hänvisningarna (Ball 2018, kap 3).

Ett exempel på en bibliometrisk indikator är *Hirsch Index* även känt som *H-index*. Indexet utarbetades av fysiker Jorge E. Hirsch och lanserades år 2005. Idéen med indexet är att det är ett enkelt sätt att visa en enskild vetenskapsmans prestationer inom vetenskaplig litteratur. Indexet räknas ut först genom att kombinera antalet publikation med antalet hänvisningar därefter ordnas publikationerna fallande ordning enligt antalet hänvisningar (Ball 2018, kap 3; De Bellis 2009, kap 6). Tabell 2 visar exempel på hur man räknar ut H-index.

Antalet Publikationer	Antalet hänvisningar
1	32
2	25
3	21
4	13
5	7
6	5
7	4
8	1

Tabell 2. H-index (Ball 2018).

I figuren ovan kan man se att sex stycken publikationer har åtminstone fem hänvisningar vilket innebär ett H-index på fem. Indexet kombinerar kvantitet, antalet publikationer, med kvalitet, antalet hänvisningar. En svaghet med H-indexet är att endast generella paralleller kan dras mellan två vetenskapsmän med motsvarande H-index men olika antal publikationer och hänvisningar (Ball 2018, kap 3

Ett annat exempel är *Impact Factor*. Denna indikator mäter hur stort inflytande en viss vetenskaplig journal har inom vetenskapen. Genom att räkna ut hur många hänvisningar en journal fått delat med antalet artiklar under de två föregående åren. Till exempel en journals *Impact Factor* för år 2014 beräknas genom att ta antalet hänvisningar från år 2014 och dela det med antalet artiklar för åren 2012 och 2013. En svaghet med denna indikator är att en journal med ett få antal artiklar med högt antal hänvisningar kan ge en hög *Impact Factor* vilket kan leda till att man för uppfattningen att den är inflytelserik (Ball 2018, kap 3; De Bellis 2009, kap 6.1).

Generellt sett måste en bibliometrisk indikator uppfylla fyra kriterier. Först måste indikatorn kunna besvara en konkret fråga om en akademisk publikation. För det andra måste indikatorn kunna mätas, annars är den inte av värde. För det tredje måste man kunna bestämma indikatorn med hög precision och kvalitet. Till sist måste man kunna

bestämma indikatorn genom en korrelation mellan möda och nytta som bestämts på förhand (Ball 2018, kap 3).

Bibliometri är inte ett ämne som är av direkt intresse för avhandlingen utan begreppet tas upp här som ett exempel på analys av journaler och vetenskapsproduktion. Själva principen om att analysera bibliometrisk data tangerar det ämne som avhandlingen tar upp även om icke är helt jämförbar med den typ av analys som avhandlingen beskriver.

3.2 Biblioteksdata

Biblioteksdata är något som är av största vikt för avhandlingen eftersom det är denna typs data som avhandlingen bygger på. Den typ av data som är av intresse för avhandlingen är sådan om gör det möjligt att jämföra akademiska bibliotek som institutioner med varandra. Vad som inte är av intresse är metadata som används för att beskriva de böcker och annat material som biblioteken har. De data som är av intresse är av typen statistik över bibliotekens användning och användardata. I nästa avsnitt presenteras olika sätt man har analyserat biblioteksdata och ges exempel på vilka olika sätt man har haft nytta av datan.

Enligt Hiller (2002) och Wang m.fl. (2016) sitter bibliotek på en hel del data som är till sin natur både numerisk och statistisk samt finns både i tryck och elektronisk form. Data har använts inom biblioteksväsendet för budgetering, användarstatistik över samlingarna och effektivisering av olika processer inom biblioteken. Utöver detta har man även använt biblioteksdata för att mäta kvaliteten på servicen och bibliotekens prestanda. Tidigare har detta mätts genom att titta på bibliotekens storlek och budgeten som indikatorer på kvalitet. Under senare år har förändringar inom biblioteken skett. Man har kunnat observera förändringar i hur kunderna använder bibliotek, e-resurser har blivit vanligare, nya organisationsstrukturer har tillkommit och man måste kunna påvisa nyttan för varje euro som man spenderar på biblioteket (Hiller 2002).

Allt detta har lett att man blivit tvungen att utveckla nya sätt att mäta bibliotekens prestanda och servicekvalitet. I sin artikel nämner Hiller (2002) LibQUAL+ som är ett verktyg för utvärdering av bibliotekens servicekvalitet, E-Metrics projektet som mäter användningen av e-resurser. Utöver detta har National Information Standards Organization (NISO) utarbetat en standard för kategorisering och insamling av data, *Metrics and Statistics for Libraries and Information Providers* (Z39.7) (Hiller 2002). Enligt Wang m.fl. (2017) är det möjligt att tillämpa s.k. "Big Data"-teknologi på biblioteksdata. Det skulle innebär att man samlar in, väljer ut, organiserar, beskriver, skapar modeller, lagrar och presenterar eller visualiserar biblioteksdata. Det som också är viktigt är dataanalys, datamodellering och datastandardisering som kräver en hel del arbete och tid. Samtidigt så kan man säga att nyttan med att utföra detta skulle vara enorm. Biblioteken kunde som följd av dataanalys av biblioteksdata fatta kostnadseffektiva, innovativa beslut eller ge rekommendationer som leder till nöjda användare (Wang m.fl. 2017).

Biblioteksdata ger en inblick i hur användarna utnyttjar bibliotekens samlingar. Till exempel om användarna favoriserar tryckta eller elektroniska resurser. Det är även möjligt att analysera hur de olika användargrupperna inom bibliotek använder resurserna. Förutom statistik över samlingarnas användning finns det siffror över användningen av bibliotekens utrymmen. Biblioteken kan ha olika utrymmen till användarnas förfogande som t.ex. rum man kan boka för grupparbeten eller datorutrymmen (Hiller 2002).

En del av biblioteksdata är metadata som har som uppgift att beskriva innehållet i bibliotekens samlingar. Metadata skapas av biblioteken i syfte att beskriva innehållet och formatet på böcker, artiklar CD-skivor och andra resurser som finns. Förutom data som beskriver innehållet i resurser finns det data som beskriver var man kan hitta dessa resurser och hur dessa resurser är klassificerade. Nyttan med denna data är att det gör det lättare för användare att hitta just den boken eller artikeln de letar efter. Skapande av metadata och klassificering av resurser är en del av bibliotekens kärnverksamhet (Frederick 2017).

3.3 Bibliomining och analys av biblioteksdata

I avsnittet behandlas begreppet *bibliomining* och hur det används inom biblioteksvärlden idag. Utöver detta begrepp behandlas även analys av biblioteksdata eftersom de två tingen är nära besläktade med varandra. Syftet är att ge en inblick i vad *bibliomining* är och hur det används och att ge insikt i analys av biblioteksdata samt klargöra nyttan med dessa två.

3.3.1 Bibliomining

Begreppet *bibliomining* myntades av Nicholson och Stanton (2003). Begreppet innebär att man använder sig av datautvinningsmetoder för att hitta mönster i biblioteksdata. Nyttan med att hitt olika röster i biblioteksdata är att man får inblick i både bibliotekspersonalens och kundernas beteende. Genom att titta på de data som genererats av personalen kunde man se ifall bibliotekets egna interna processer fungerar effektivt. En studie av användarnas genererade data kunde man se ifall de hittat den information och resurser de sökt efter (Nicholson och Stanton 2003).

Enligt Nicholson (2003) går bibliominingprocessen ut på att man först bestämmer det område man vill fokusera på. Det kan hända att man vill veta om det finns intressanta mönster i data eller så har man ett visst mål i sikte, ett problem man vill lösa. Efter man gjort detta identifierar man de lämpligaste källorna för processen. Enligt Nicholson (2003) bör följande två typer av datakällor övervägas: 1) interna datakällor som t.ex. kunddatabas, transaktionsdata och loggar från webbserver; 2) externa datakällor som t.ex. demografisk data. När man har valt källorna man vill använda skapar man ett datamagasin (eng. *data warehouse*) där man in och rensar data. Efter att detta är gjort krävs det endast mindre underhåll framöver (Nicholson 2003).

För att kunna analysera data behöver man välja ett lämpligt analysverktyg. Enligt Nicholson (2003) kan det vara antingen ett MIS (eng. *Management Information Systems*) eller OLAP (eng.*On-Line Analytical Processing*). Med MIS kan beslutsfattarna med hjälp av rapporter och medeltal för olika variabler kan beslutsfattarna inom

biblioteken hålla ett öga på vad som sker i biblioteken. (Nicholson 2003) Förutom traditionella rapporter och medeltal kan detta även öppna dörren för möjligheten att ställa nya frågor. Inte bara frågor om dagsläget utan även om det som skett tidigare vilket är möjligt tack vare att datamagasinet innehåller all nuvarande och tidigare data (Nicholson 2006).

OLAP is sin tur skapar en interaktiv vy för beslutsfattarna och personalen på deras skärmar. Detta är möjligt eftersom OLAP gör tusentals förfrågningar av datamagasinet för att kombinera ett flertal variabler tillsammans med utvalda mått. (Nicholson 2003) OLAP är ett beslutsstödssystem (eng. *decision support system*) som gör det enkelt att skapa sig en översikt över data i datamagasinet (Nicholson 2006).

Förutom dessa två nämna metoderna kan man ytterligare använda sig av visualisering av data vilket gör det möjligt att snabbt upptäcka mönster. Bara genom att titts på siffror är det inte nödvändigtvis möjligt att upptäcka mönster. Visualisering gör det möjligt för personal och beslutsfattare att utforska biblioteksdata (Nicholson 2006).

3.3.2 Analys av biblioteksdata

I detta avsnitt presenteras ett urval artiklar som belyser analys av biblioteksdata. Meningen är inte att avsnittet skall vara uttömmande utan snarare fungera som exempel på vad för slags analyser det görs på biblioteksdata.

Sung och Tolppanen (2013) utförde en studie där de tog reda på huruvida förseningsavgifter vid universitetsbibliotek hade någon inverkan på kundernas beteende. Studien utfördes vid två mellanstora universitetsbibliotek i USA, Eastern Illinois University (EIU) och University of Hawaii i Manoa (UHM). Biblioteket vid EIU tillhandahåller bibliotekstjänster åt 10000 grundexamensstuderande, 1500 magistersstuderande samt 900 lärare och forskare. Det andra biblioteket vid UHM har lite fler studerande och personal, 14000 grundexamens-, 6000 magistersstuderande samt 1700 lärare och forskare (Sung och Tolppanen 2013).

Vid EIU har hade man en förseningsavgift på $0.25 per dag för varje försenad bok med en max avgift på $10 per bok. Deras praxis var att inte debitera avgifter under $2.50 vilket innebar en nådperiod på tio dagar. Konsekvenser av obetalda avgifter blev att studerande inte kunde närvaroanmäla sig följande termin. UHM hade en liknande praxis och avgifter, med det undantaget att deras praxis inte hade någon nådperiod. EIU debiterar ingen förseningsavgift av forskare och lärare medan UHM gör ingen skillnad mellan användargrupperna (Sung och Tolppanen 2013).

Proceduren för att frå fram data vid båda biblioteken var identiska eftersom båda använde sig av Voyagers integrerade bibliotekssystem som är en mjukvara som bibliotek kan använda. För att analysera data som tagits fram använde sig Sung och Tolppanen (2013) av SPSS (Statistical Package for Social Sciences) som är ett mjukvarupaket för statistisk analys. Den metod som valdes var ANOVA eller variansanalys (eng. *analysis of variances*) som lämpar sig för att analyser olikheter mellan olika grupper (Sung och Tolppanen 2013).

Resultatet av analysen visade att förseningsavgifter har en inverkan på kundernas beteende. Kunder som lånade böcker återlämnade böckerna före förfallodagen i den grad att det kan anses vara statistiskt signifikant. Slutsatsen som Sung och Tolppanen (2013) drog var att förseningsavgifter är ett effektivt verktyg för att försäkra sig om att böcker återlämnas i tid (Sung och Tolppanen 2013).

Kovacevic, Devedzic och Pocajt (2009) lade fram en sätt där datautvinning kunde användas för att rekommendera digitala bibliotekstjänster åt kunder i Serbien. Kovacevic m.fl. (2009) använde sig av REKOB som är ett datautvinningssystem för att hjälpa nya användare av KOBSON-biblioteket hitta de tjänster de letar efter på basen av vilka tjänster liknade användare ansett varit användbara. KOBSON gör det möjligt för serbiska studerande, forskare och lärare att komma åt utländska journaler och resurser som t.ex. Springer, Science Direct, ProQuest osv. (Kovacevic m.fl. 2009).

Datautvinningen gick ut på att analysera användarnas sökhistorik och användarprofil. För att göra denna analys använde sig Kovacevic m.fl. (2009) av K-means klusteranalys. Datautvinningen resulterade i användarmönster som baserats på användarnas sökhistorik och användarprofil. Efter att klusteranalysen gjorts och klusterna skapats kunde nya användare sätta i klusterna. Till slut användes Naive-Bayes klassifikationsalgoritm för att göra lämpliga rekommendationer (Kovacevic m.fl. 2009).

Data som Kovacevic m.fl. (2009) använde sig av KOBSON-bibliotekets verkliga användardata. För att användarna skall få tillgång till biblioteket måste de först registrera sig. Användarprofildata samlades in i samband med registreringen och modifierades en aning för att kunna användas vid datautvinningen. För att skydda användarnas identitet ersattes viss data med koder. Endast behöriga var medvetna om användarnas verkliga identitet (Kovacevic m.fl. 2009).

Slutsatsen av denna studie var att det går att använda datautvinningsmetoder som k-means klusteranalys och Naive-Bayes klassifikation tillsammans för att förbättra tjänster vid digitala bibliotek (Kovacevic m.fl. 2009).

Zaugg, McKeen, Hill och Black (2017) utförde en studie som hade som syfte att skapa ett digitalt inventarie (DI) för ett stort privat akademiskt bibliotek. Zaugg m.fl. (2017) ville visa processen och nyttan med att utföra ett dylikt inventarie. För att göra inventariet i studien valdes Harold B. Lee-biblioteket vid Brigham Young-universitetet (BYU) i USA.Vid universitetet finns ca. 33000 studerande och biblioteket har 156 heltidsanställda (Zaugg m.fl. 2017).

Metoden som användes var intervjuer med biblioteksledningen och bakgrundssökning. På basen av dessa kunde man komma fram till möjliga kategorier för inventariet. Tre stora övergripande kategorier identifierades och samtidigt skapades underkategorier för dem. Eftersom data som biblioteket besatt samlats in av mänskor fokuserades inventariet på vem som samlat in data och deras befattning inom biblioteket. Man kom

fram till att vartefter personers befattning ändrades inom organisationen ändrades även insamling av data eller upphörde helt (Zaugg m.fl. 2017).

Syftet med att skapa ett DI var inte för att göra en djup analys utan snarare för att identifiera alla datakällor som biblioteket skapat och indikera var data kunde återfinnas. Idéen var att bibliotekspersonalen kunde med hjälp DI se ifall den data de behövde redan samlas in och så fall var data återfinns. Nytta med detta skulle vara att personalen kunde använda data för att fatta bättre beslut, se vad som behöver förbättras och utöva biblioteks centrerad forskning. Datainventariet bli således inte forskningsdata utan ett verktyg för forskning och planering (Zaugg m.fl. 2017).

Sammanlagt deltog 156 heltids- och deltidsanställda i inventarier. Man lyckades identifiera 612 dataset varav 52% var kvantitativa och 18 kvalitativa. Resten av data var demografiskt, positionsdata och loggar. Sjuttiosju anställda ville att mer data smalas in. Förfrågningar kategoriserades i sju breda kategorier: samling, teknologi, instruktion, professional development (PD), information, marknadsföring och webbsajt (Zaugg m.fl. 2017).

Ett datainventarie (DI) fungerar som ett uttömmande översikt över vilken data som är tillgänglig inom ett bibliotek. Nyttan med detta är att inventariet blir ett starkt verktyg för identifiering av vilken typs data samlas in och var data återfinns. Data indikerar också var det finns möjlighet till samarbete, vilket kan leda till effektivare praxis inom organisationen. Till exempel vilken data som samlas in men inte används eller där insamlad data inte ger något av värde. Ytterligare är det möjligt att med hjälp av ett DI visa för personalen nyttan med datainsamling och vilken typ av data som samlas in. Med ett DI möjliggör man det att personalen hamnar att stiga ur sina roller och bekanta sig med vad kollegerna gör (Zaugg m.fl. 2017).

In sin artikel använder sig Jiang och Carter (2018) av mjukvaran R för att visualisera biblioteksdata. Den första visualiseringen gjordes på University of Alabamas institutional repository (IR) som fungerar som samlingsplats för vetenskapligt material

och forskning. Visualiseringen gjordes som en världskarta där man kunde se var användarna av IR befann sig och hur mycket det laddat ned av materialet. Processen med att samla in data gjordes via Google Analytics och sedan konvertera plats information till geografiska koordinater så att man kunde visa på världskartan var användarna befann sig. Google Analytics används även för att få användardata ur IR, d.v.s. hur många nedladdningar användarna hade gjort. Detta visualiserades på världskartan så att en större punkt betydde fler nedladdningar och en minde färre nedladdningar. Genom att klicka på punkterna kan man få fram information om den orten och hur många nedladdningar som gjorts (Jiang och Carter 2018).

Den andra visualiseringen gjordes i form av en instrumentbräda (eng. *dashboard*) i syfte att följa med alla besök som gjorts till bibliotekets alla filialer och servicepunkter. Nyttan med att hålla reda på besöksantalet är att biblioteken kan då bestämma hur mycket personal och när de behöver vara på plats. Ytterligare kan detta ligga till grund för beslut om öppethållningstider. Data tog ur bibliotekets LibPAS-programvara som sedan visualiserades m.h.a. R. I visualiseringen kan man jämföra besökarantalet i t.ex. oktober 2016 med oktober 2017. Förutom detta kan man se vilka filialer och servicepunkter som besökts mest och under vilka tider (Jiang och Carter 2018).

3.4 Sammanfattning

I kapitlet har termerna bibliometri, biblioteksdata och bibliomining definierats och diskuterats. Förutom termerna har olika exempel på analys av biblioteksdata presenterats och nyttan med dessa analyser diskuterats. Speciellt av värde är Jiang och Carters (2018) användning av mjukvaran R för visualisering av biblioteksdata eftersom det är visualisering av biblioteksdata med hjälp av mjukvaran R som avhandlingen går ut på.

4. METOD

I kapitlet diskuteras val av metod för analysen av biblioteksdata från Finlands allmänna bibliotek. Dessutom diskuteras från vilken databas data är hämtat och vilken typ av data det är frågan om. Till slut diskuteras processen för bearbetningen av data så att de sedan kan användas av mjukvaruprogrammet R.

4.1 Metodval

Valet av metod för avhandlingen föll på självorganiserande kartor eftersom metoden är till sin natur visuell. Avsikten med avhandlingen är att visuellt representera och utforska en uppsättning data. I det här fallet är det frågan om att se vad data som helhet via visuell representation kan säga om de allmänna biblioteken i Finland. Genom att visuellt representera olika variabler och jämföra visualiseringarna med varandra är det möjligt att säga något om biblioteken. Till exempel vilka som satsat mest på verksamheten.

Självorganiserande kartor har använts av andra forskare och därmed tillämpats på andra datasets. Alla dessa tillämpningar har motiverats på olika sätt. Till exempel Holmbom (2015, s. 55) motiverar sitt val av självorganiserande kartor med att säga att självorganiserande kartor är datadrivna och klarar av att hantera stora mängder olika typer av data. Utöver detta kan man visualisera resultatet (Holmbom 2015, s.55). Eklund (2003, s. 2) i sin tur motiverar sitt val med att säga att självorganiserande kartor är lämpade för att hitta mönster i data. Mönster som inte nödvändigtvis hittas bara genom att titta på Excel-tabeller..

Verktyget för skapande av den självorganiserande kartan och utförande av visualiseringen av data är mjukvaruprogrammet R. Valet av R är enkelt eftersom det redan används inom flera kurser vid Åbo Akademi och för att det är gratis. Utöver detta är det möjligt att utöka programmets funktioner med att ladda ned olika kodpaket som ger utökade funktioner. En av dessa utökade funktioner är skapandet av självorganiserande kartor. I denna avhandling används Kohonen-paketet i detta syfte.

Paketet möjliggör skapande av en självorganiserande karta och sedan kan man använda olika visualiseringsmetoder som t.ex. U-matris eller värmekartor för att skapa en bild över data. I avhandlingen används version 3.5.1 av R och version 3.0.7 av Kohonen-paketet, utöver detta används version 1.1.463 av RStudio som är ett grafiskt användargränssnitt för R.

4.2 Beskrivning av data

Data som används i avhandlingen har hämtats från tilastot.kirjastot.fi som upprätthålls av undervisnings- och kulturministeriet. Biblioteken matar själva in statistiska uppgifter i databasen och regionförvaltningsverken granska uppgifternas riktighet innan de publiceras. Uppgifterna matas in årligen för varje statistikår som sträcker sig från 1.1 till 31.12. Den statistiska grundenheten som används är kommunernas biblioteksväsen som helhet, vilket innebär att det inte finns uppgifter för enskilda serviceenheter. (Kirjastot.fi 2018)

I databasen är det möjligt att välja vilket verksamhetsår man vill ha statistik för. Uppgifterna i databasen sträcker sig från 1999 ända till 2017. Den statistik som är av intresse för avhandlingen är från verksamhetsåret 2017, alltså statistik från bibliotek som var verksamma det året . För att få med alla kommuners biblioteksväsen valdes alla kommuner för det valda verksamhetsåret, 2017. Det är även möjligt att begränsa valet av kommuner enligt regionförvaltningsverk, språklig indelning, landskap eller statistisk kommungruppering (urbana kommuner, landsbygdskommuner eller tätortskommuner).

De statistiska uppgifterna är indelade i skilda kategorier som har olika teman. Den kategori som är av intresse för avhandlingen är nyckeltal. Kategorin ifråga innehåller alla nyckeltal som beräknats på basen av de data som biblioteken matat in och är indelad i ett antal underkategorier. Efter att man valt kategori och de underkategorier man vill exporterar man de utvalda data som Excel-fil som laddas automatiskt ned på den dator man använder. För behandlingen av den nedladdade Excel-filen som används

i avhandlingen utnyttjades datorprogrammet Numbers som kommer förinstallerad med Apple-datorer.

Efter att filen importerats in i Numbers ser innehållet ut som vilken tabell som helst med rader, kolumner och celler som är fyllda med värden. I den fil som är till grund för avhandlingens analys återfinns de 277 kommunernas namn i kolumn A och i B kolumnen representeras kommunnamnen av koder som sträcker sig från B001 till B227. Kommunerna gavs en kod p.g.a. att vissa kommuner har väldigt långa namn och det kan vara svåra att utläsa från visualiseringarna i SOM. Genom detta representeras varje kommun av en fyra tecken lång kod oavsett längden på kommunnamnet. Koderna fanns inte i originalfilen utan lades till enkom för avhandlingens analys. Förutom att koder lades till förkortades även vissa kommunnamn eftersom de innehöll de kommuner som ingått i tidigare års kommunsammanslagningar. Till exempel i originalet benämndes Väståboland som Väståboland/Pargas vilket förkortades för avhandlingens analys till endast Väståboland.

Kolumnerna C till O innehåller de nyckeltal som används i analysen. Alla nyckeltal som finns i databasen på tilastot.kirjastot.fi har inte tagits med i analysen utan en hel del har valts bort, främst p.g.a. att de innehöll NULL-värden. Med NULL-värden avses här att cellen i Excel-tabellen är tom. Utöver nyckeltal med NULL-värden valdes nyckeltal för böcker, e-resurser och tidskrifter bort för att i stället fokusera avhandlingens analys på bibliotekets samlingar som helhet och int blanda in individuella kategorier. Följande tabell illustrerar vilka nyckeltal som togs med och de förkortningar som används för dem:

Nyckeltal	Förkortning
Mediebestånd / Invånare	KAL
Anskaffningar / Invånare	HAL
Anskaffningar / Mediebestånd	HK
Fysiska besök / Invånare	FKAL
Fysiska besök / Öppettimmarna	FKAT

Nyckeltal	Förkortning
Omkostnader under statistikåret / Fysiska besök	TKFK
Totalutlåning / Invånare	KLAL
Utlåning / Mediebestånd	LK
Omkostnader under statistikåret / Totalutlåning	TKKL
Omkostnader under statistikåret / Invånare	TKAL
Personalkostnader / Invånare	HENKKAL
Kostnader för medieanskaffningar / Invånare	KAKAL
Kostnader för bokanskaffningar / Invånare	KHKAL

Tabell 3. Nyckeltalen och förkortningarna.

Orsaken att förkortningar används för nyckeltalen är att de har långa namn och för att det ser bättre ut vid visualisering. Efter att nyckeltalen valts ut och förkortats samt kommunnamnen förkortats och kodats exporterades tabellen som CSV-fil.

4.3 Sammanfattning

I kapitlet har val av metod diskuterats och motiverats. Förutom detta har valet av verktyg, mjukvaruprogrammet R och vilka tilläggspaket som valts presenterats. De data data som avhandlingen använder sig av har diskuterats och förbearbetningsprocessen har beskrivits. Med förbearbetningsprocess avses val av variabler och kodningen av kommunnamnen samt exporten av data till en CSV-fil.

5. RESULTAT

Kapitlet handlar om resultaten av analysen och skapande av tillhörande visualiseringar. Till först relateras hur den självorganiserande kartan skapades i R sedan diskuteras olika val som gjorts under skapandet av modellen. Efter att modellen har diskuterats presenteras komponentplanen och sedan diskuteras klusteranalysen.

5.1 Skapande av den självorganiserande kartan i R

För att skapa koden i RStudio användes ett skript vilket gör det lättare att modifiera koden under processens gång. I annat fall hade man varit tvungen att mata in koden ett segment i taget varje gång man velat köra den. Förutom ett skript skapades även ett R-Projekt inom vilket koden, skriptet och sessionen sparades. Användandet av projektfunktionen samlar allt som har med analysen på ett ställe.

Efter att skript- och projektfilerna skapats i den katalog på datorns hårdskiva man tänkt använda för analysen importerades och laddades ned de paket som analysen kräver. Det paket som är av största vikt för analysen förutom RStudios inbyggda funktioner är Kohonen-paketet. När paketen laddats och installerats importerades den fil som innehåller biblioteksdatan. RStudio tolkade filen som en dataram (eng. *data frame)* som innehåller 277 observationer och 15 variabler. Observationerna är de 277 kommuner som diskuterades i kapitel 4 och variablerna är kommunnamnen, koderna och de andra variablerna som presenterades i kapitel 4. Dataramen gavs namnet data.

Till näst plockades de variabler den självorganiserande kartan skulle tränas med och sparades i dataram med namnet data_train. Sammanlagt valdes 13 variabler av 15, två variabler lämnades bort i detta skede. Variablerna för kommunnamn och kod var inte relevanta för träningen av den självorganiserande kartan utan endast de variabler innehåller observationer med numeriskt värde. Eftersom variablerna har olika skalor på de numeriska värden, vissa värden är av typen 1.0 och andra 100.0, är det nödvändigt att tillämpa samma skala på alla värden. Annars för värden av typen 100.0 för stort

inflytande på kartan och genom att justera dem blir kartan rättvis. De justerade värdena sparades som en matris med namnet data_train_matrix.

I det här skedet är data redo att träna den självorganiserande kartan. Innan själva träningen av kartan kan börja skapades ett sex gånger fem rutor stort rutsystem. Gallret i rutsystemet bestämdes som hexagonalt. När kartan tränades användes batch träningsalgoritmen. I Bild 2 kan utläsas träningens framgång.

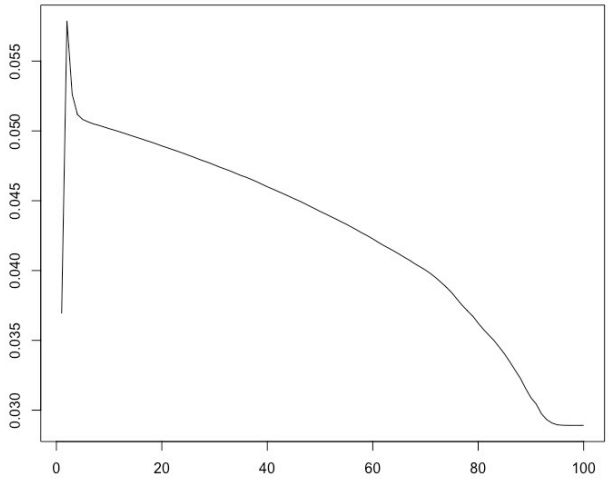

Bild 5. Träningens framgång.

På Bild 5 ser man på y-axel medeldistansen till närmaste enhet och på x-axeln antalet iterationer. Träningsdata data_train_matrix har tränat den självorganiserande kartan 100 gånger eller i 100 iterationer. Resultat av träningen är att medelavståndet till närmaste enhet har minskat ända tills att det nått en botten och planat ut..

Bild 6 visar antalet indatavektorer per neuron eller i detta fall antalet kommuner per neuron. De ljusa färgerna indikerar höga värden och mörka färgerna låga värden. På basen av bilden kan man uppskatta att de flesta neuronerna har kring tio kommuner i sig vilket verkar rimligt eftersom det finns 277 kommuner med i indata och 30 neuroner i SOM-modellen. Man kan konstatera att fjärde neuronen från vänster i mittersta raden har många kommuner i sig, så till den grad att den avviker från massan. Utöver den neuronen finns det även ett antal rödfärgade neuroner som indikerar att det finns ett färre antal kommuner i dem.

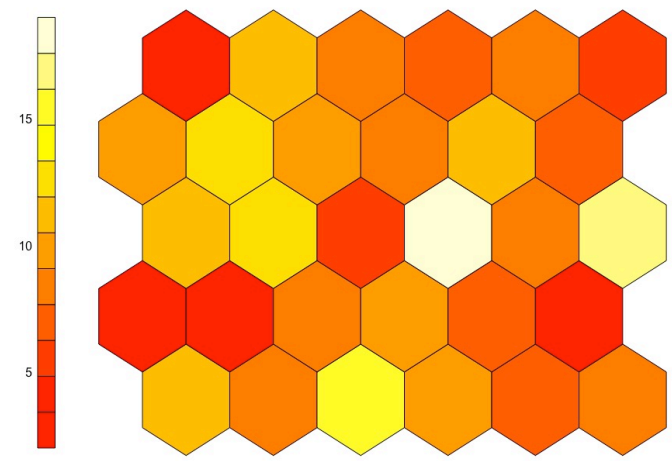

Bild 6. Indatavektorer per neuron.

Bild 7 visar den självorganiserande kartans *U-matris*. De mörka färgerna indikerar kortare avstånd och de ljusa längre. Man kan konstatera att de flesta neuroner ligger nära varandra, men det finns vissa som sticker ut.

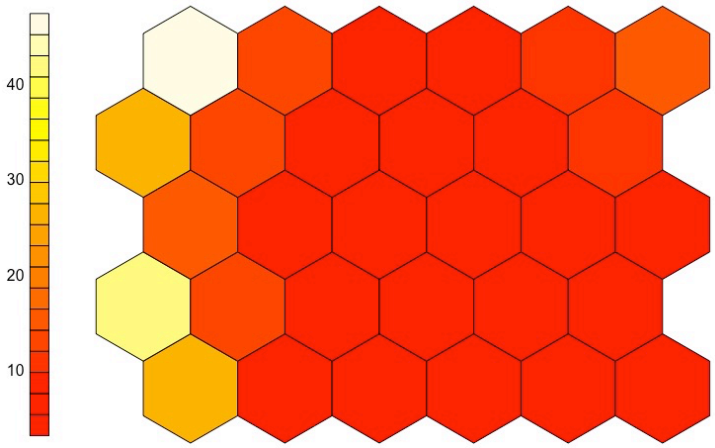

Bild 7. Den självorganiserande kartans U-matris.

På Bild 8 kan man se en kvalitetsplott som illustrerar kvaliteten på den självorganiserande kartan. De mörkare färgerna indikerar kortare avstånd medan de ljusare längre. På basen av kvalitetsplottet kan man säga att kvaliteten på den självorganiserande kartan är relativt hög eftersom de flesta neuroner har mörk färg. Kort sagt representerar de flesta neuroner sina? indatavektorer väl.

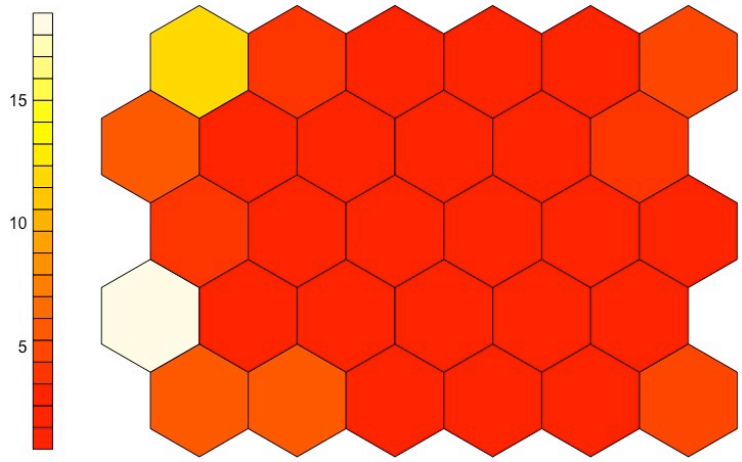

Bild 8. Den självorganiserande kartans kvalitetsplott

Bild 9 visar vilka variabler har påverkat vilka neuroner och till vilken grad. Det är även möjligt att se variablernas spridning på kartan. Vissa neuroner har just inga inflytelserika variabler, vissa några och vissa klart fler. Neuronerna nere till vänster har flera inflytelserika variabler, det gäller även neuroner upp till höger.

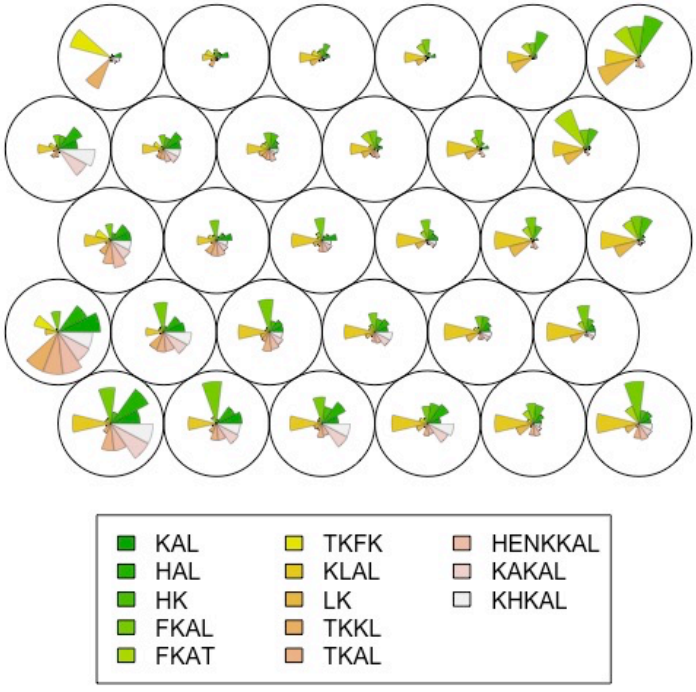

Bild 9. Variablers inverkan på neuroner.

5.2 Komponentplan

Varje enskilt komponentplan representerar en variabel, utom variablerna för kommunnamn och kod. Genom att jämföra komponentplanen sinsemellan kan man upptäcka samband eller andra intressanta mönster. Komponentplanen visualiserades med samma färgpalett vilket gör det enklare att jämföra dem sinsemellan.

5.2.1 Komponentplanen

I avsnittet presenteras och diskuteras komponentplanen i den ordning som variablerna de representerar förekommer i indata.

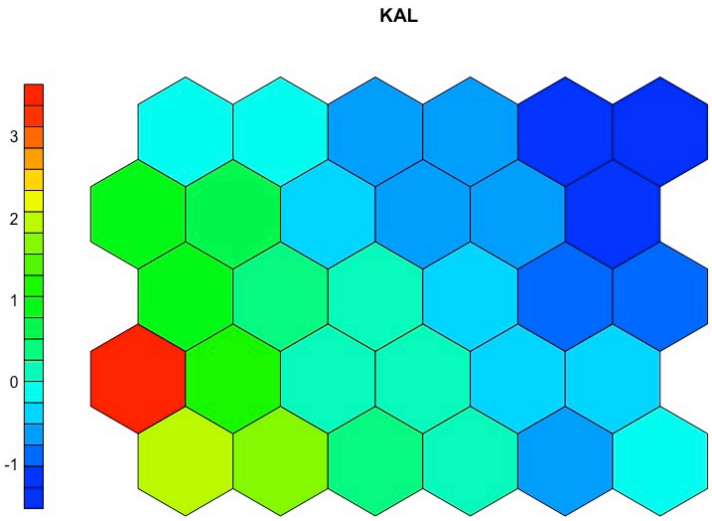

Bild 10. Variabeln KAL.

Bild 10 visar hur mediebeståndet per invånare (KAL) är distribuerat i de olika neuronerna. Det som man kan konstatera är att en neuron har röd färg som tyder på ett högt värde. Kommunerna i den neuronen har högt mediebestånd per invånare. Neuronerna uppe till höger i Bild 7 är färgade mörkblåa som indikation på lågt mediebestånd per invånare. Resten av neuronerna har medelhöga till låga värden.

På Bild 11 kan man se hur anskaffningarna per invånare (HAL) är distribuerade i neuronerna. Kommunerna i de två neuronerna som är färgade röda och gula nere till vänster har mycket anskaffningar i förhållande till befolkningen. Kommuner med lite anskaffningar i förhållande till befolkningen finns i neuronerna i översta raden och upp

till höger. Resten av kommunerna har medelhöga till låga anskaffningar i förhållande till befolkningen.

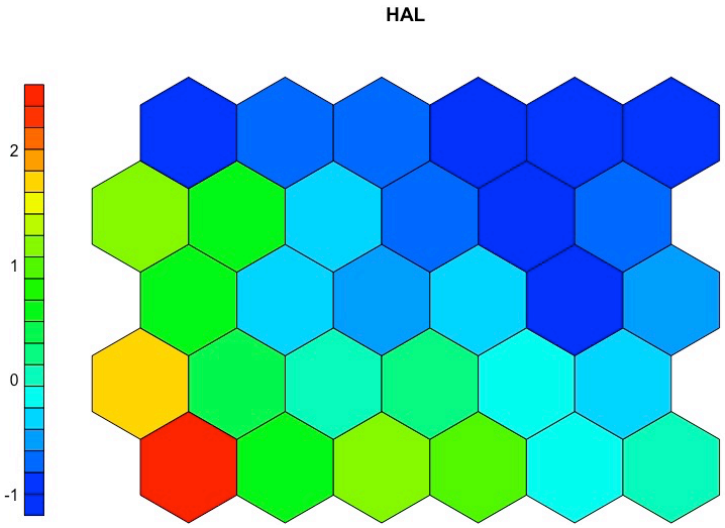

Bild 11. Variabeln HAL.

Bild 12 illustrerar anskaffningarna i förhållande till mediebestånd (HK). Av komponentplanet kan man avläsa att de flesta neuroner har medelhöga till låga värden. Det finns en neuron som sticker ut ur mängden och det är den uppe i högra hörnet som är färgad röd som indikation på att kommunerna i den neuronen anskaffar mycket i förhållande till deras mediebestånd.

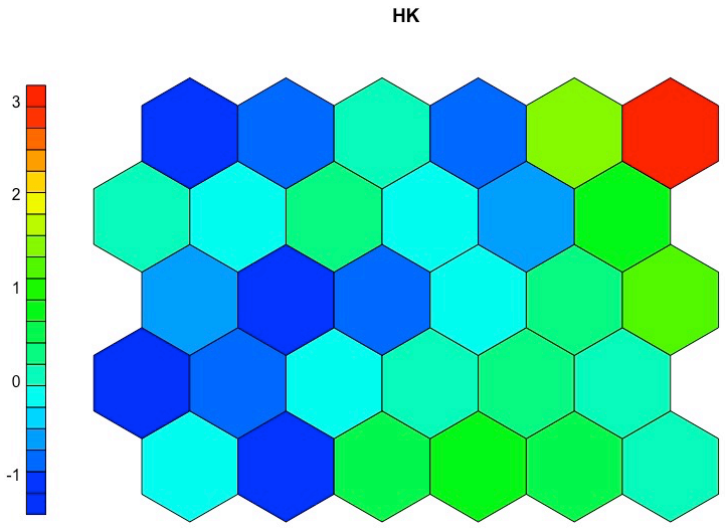

Bild 12. Variabeln HK.

Variabeln FKAL som betyder fysiska besök per invånare illustreras i Bild 13. Man kan se att där finns två röda neuroner som indikerar mycket höga värden och 4 andra

neuroner som indikerar höga värden som innebär många besök per invånare. I övre raden till vänster finns tre neuroner i blått som tyder på låga värden, alltså få fysiska besök per invånare. Resten faller någonstans mittemellan dessa.

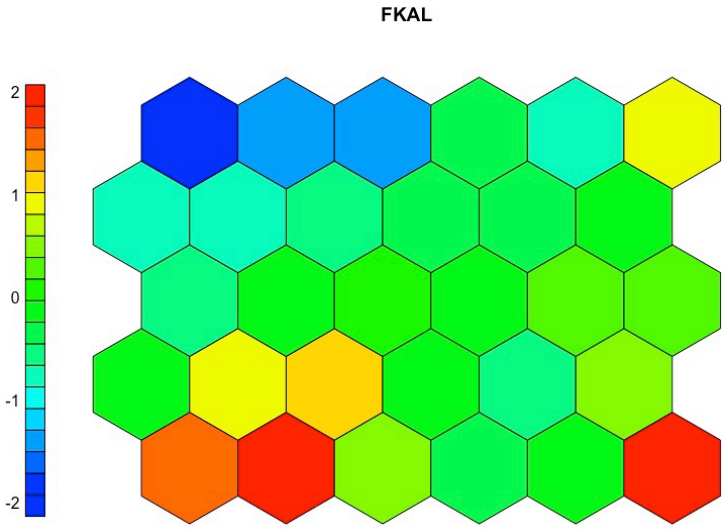

Bild 13. Variabeln FKAL.

Bild 14 visar fysiska besök under öppettimmarna (FKAT). I näst översta raden längst till höger finns en rödfärgad neuron som indikerar många besök under öppettimmarna. Utanför den röda noden till höger finns en gulaktig neuron som även den tider på relativt många besök under öppettimmarna. De blåa neuronerna tyder på få till mycket få besök medan de gröna faller mittemellan.

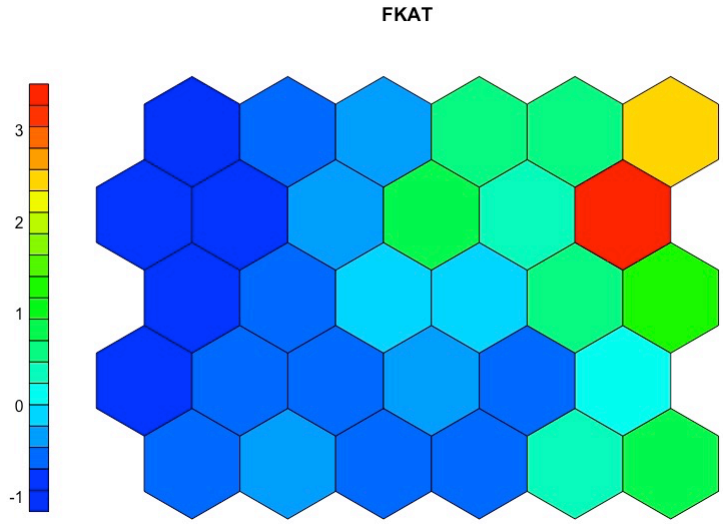

Bild 14. Variabeln FKAT.

Bild 15 visar kommunernas biblioteksväsens omkostnader under statistikperioden per varje fysiskt besök. Av bilden framgår att de flesta neuroner är blåaktiga som tecken på låga omkostnader per besök (TKFK). Uppe till vänster finns en avvikande neuron som är röd vilket innebär att kommunerna där har höga omkostnader per besök.

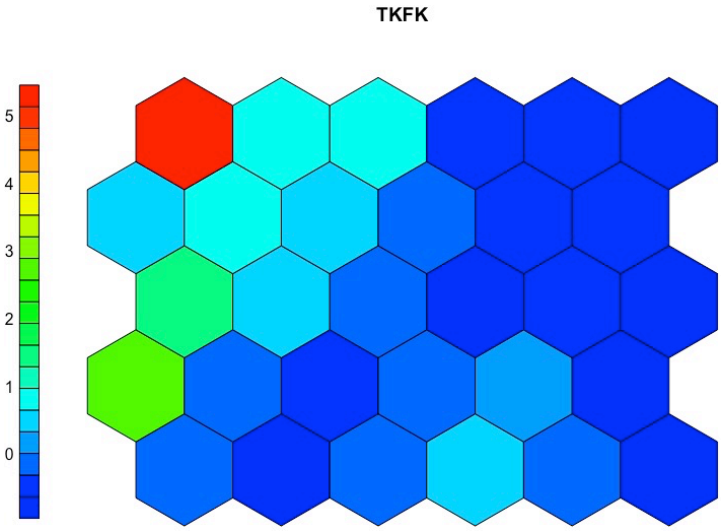

Bild 15. Variabeln TKFK.

Totalutlåningen per invånare (KLAL) illustreras i Bild 16. På basen av bilden kan man konstatera att utlåningen i de flesta kommuner är medelhög till mycket hög. Det finns en blå avvikande neuron uppe till vänster som indikerar mycket lågt lånande av medieresurser vid biblioteken i de kommuner som finns i neuronen.

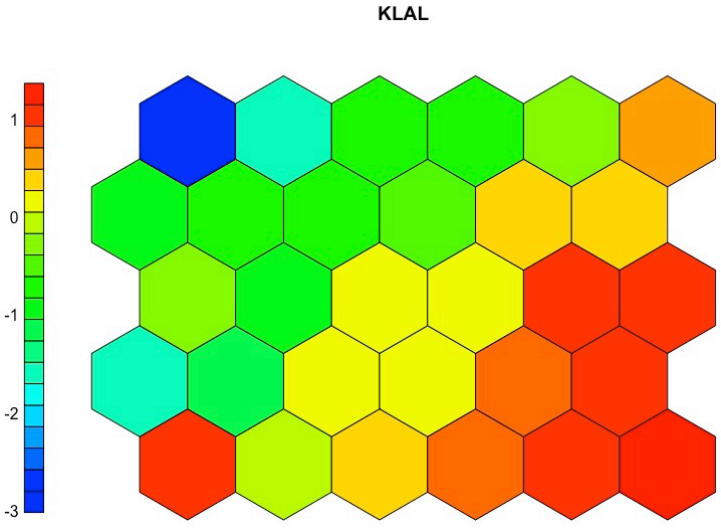

Bild 16. Variabeln KLAL.

På Bild 17 kan man se utlåning i förhållande till mediebestånd. Det finns en röd neuron i hörnet upp till höger på bilden vilket indikerar hög utlåning i förhållande till mediebestånd. Resten av neuronerna indikerar medel högt till låg utlåning i förhållande till mediebeståndet.

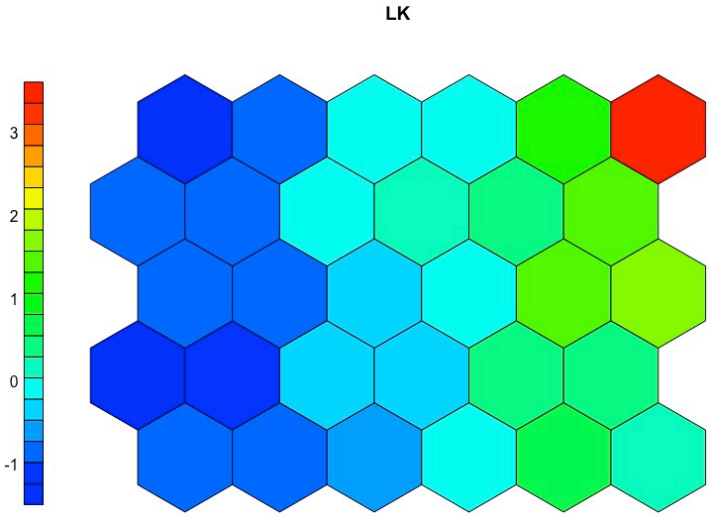

Bild 17. Variabeln LK.

Variabeln TKKL på Bild 18.visar omkostnaderna under statistikåret per totalutlåning. En rödfärgad neuron finns på näst lägsta raden längst till vänster. Den indikerar mycket höga omkostnader i förhållande till totalutlåningen. Uppe i det vänstra hörnet finns en gul neuron som indikerar höga omkostnader i förhållande till totalutlåningen. De blåaktiga neuronerna indikerar låga omkostnader i förhållande till totalutlåningen.

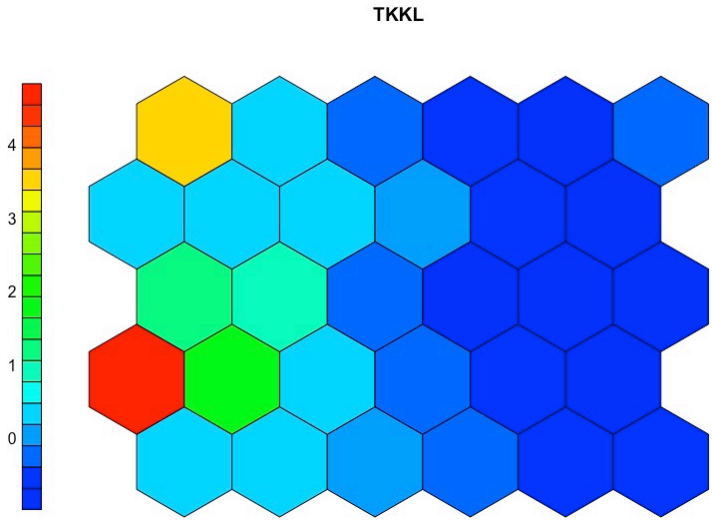

Bild 18. Variabeln TKKL.

På Bild 16 kan man se variabeln TKAL som innebär omkostnaderna under statistikåret per invånare. På den näst lägsta raden finns det en röd neuron längst till vänster som indikerar höga omkostnader per invånare. Resten av neuronerna indikerar medelhöga till låga och mycket låga omkostnader per invånare.

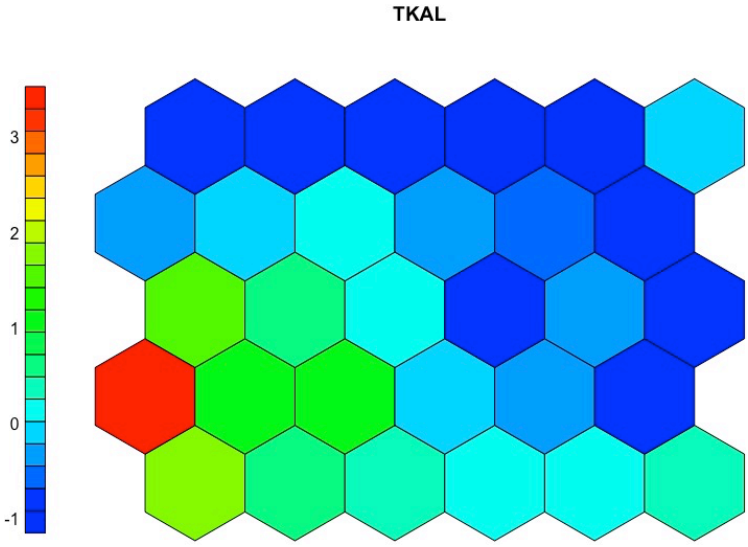

Bild 19. Variabeln TKAL.

Variabeln HENKKAL (se Bild 20) står för kommunernas biblioteksväsens personalkostnader per invånare. På bilden finns en röd neuron som indikerar höga personalkostnader per invånare i kommunen. Den gröna neuronerna indikerar moderata personalkostnader per invånare och de blåa indikerar låga personalkostnader per invånare.

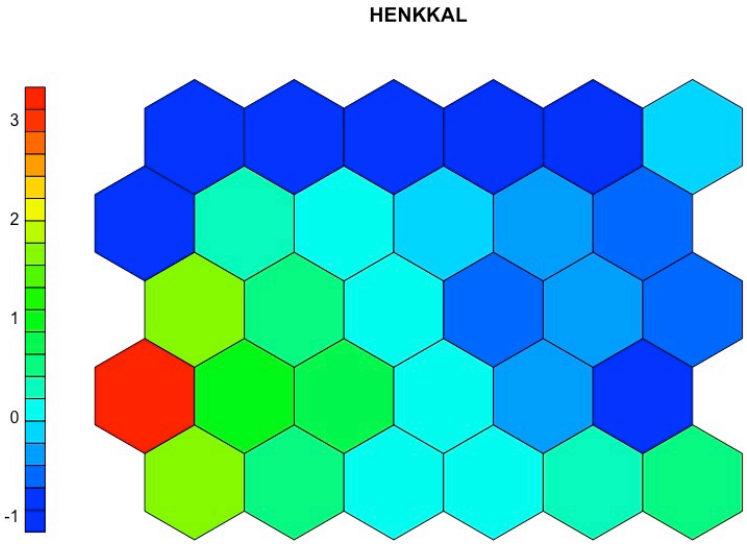

Bild 20. Variabeln HENKKAL.

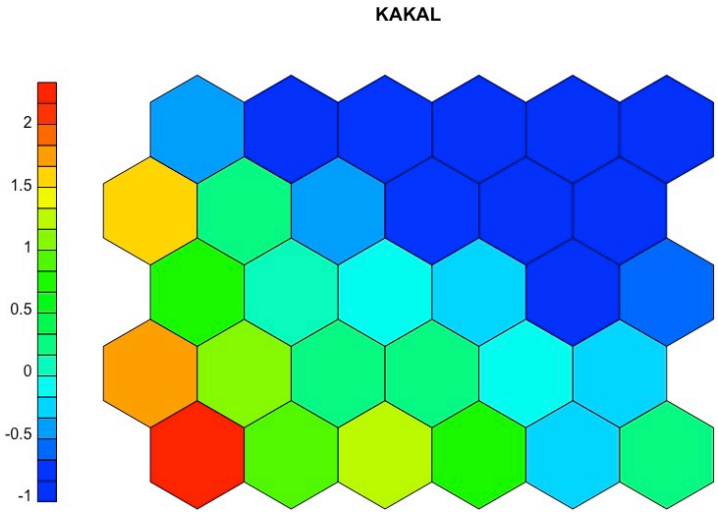

Bild 21. Variabeln KAKAL.

Bild 21 illustrerar kostnaderna för medieanskaffningar per invånare (KAKAL). Det finns en röd neuron på nedersta raden längst till vänster som indikerar mycket höga anskaffningskostnader per invånare. Utöver den neuronen finns det på vänstra sidan i bilden två orangea neuroner som indikerar höga anskaffningskostnader per invånare. De blåa neuronerna indikerar låga anskaffningskostnader per invånare och de gröna moderata anskaffningskostnader per invånare.

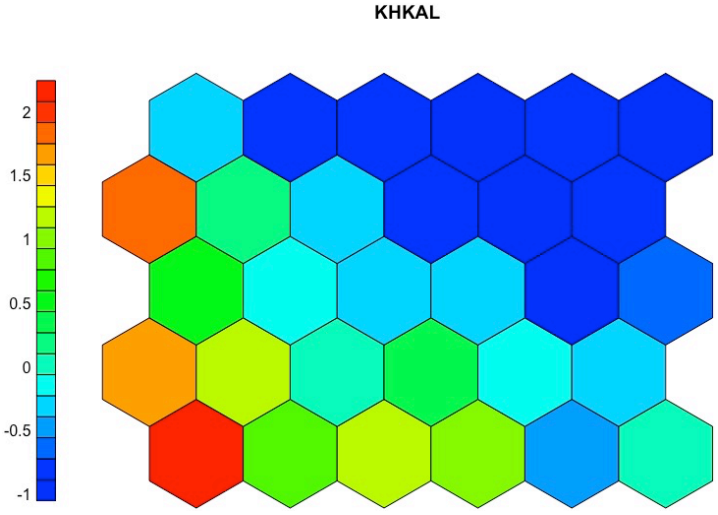

Bild 22. Variabeln KHKAL.

Bild 22 illustrerar kostnaderna för bokanskaffningar per invånare (KHKAL). Det finns en röd neuron på nedersta raden längst till vänster som indikerar mycket höga anskaffningskostnader per invånare. Utöver den neuronen finns det på vänstra sidan i

bilden två oranga neuroner som indikerar höga anskaffningskostnader per invånare. De blåa neuronerna indikerar låga anskaffningskostnader per invånare och de gröna moderata anskaffningskostnader per invånare. Det är naturligt att komponentplanen för KAKAL och KHKAL liknar varandra eftersom båda handlar om medieanskaffningar.

5.2.2 Mönster och likheter mellan komponentplanen

I föregående avsnitt presenterades komponentkartorna som representerar de variabler som ingått i det dataset som använts i avhandlingen. Genom att analysera de presenterade komponentplanen kan man upptäcka mönster och likheter mellan komponentplanen. Det är minst lika intressant att se om det finns inversa samband mellan variablerna i komponentplanen.

Mellan variablerna TKKL (se Bild 18) och KLAL (se Bild 16) verkar det existera ett omvänt samband. Det här tyder på att biblioteksväsen som har låga utgifter per lån har många lån per invånare. Sedan verkar det finnas ett samband mellan variablerna TKAL (se Bild 16) och HENKKAL (se Bild 20). Man kan sluta sig till att de biblioteksväsen som har höga omkostnader under statistikåret per invånare har även höga personal kostnader under statistikåret per invånare. Man kan även säga om att de biblioteksväsen som har låga omkostnader även har låga personalkostnader.

Det verkar också existera ett samband mellan variablerna LK (se Bild 17) och FKAT (se Bild 14). Sambandet är inte fullständigt men likheter finns t.ex. mellan de blåa neuronerna på vänstra sidan av bilderna. Man kan även se ett visst samband mellan de grönaktiga neuronerna på högra sidan av bilderna. De här likheterna tyder på att de biblioteksväsen som har låg utlåning i förhållande till mediebeståndet även har få fysiska besök under öppettimmarna. Man kan även sluta sig till att de biblioteksväsen som har mycket hög utlåning i förhållande till mediebeståndet även har många fysiska besök per invånare på basen av den röda noden upp till höger på Bild 17 och den gulaktiga noden på samma ställe på Bild 14.

Man kan se likheter mellan variablerna KHKAL (se Bild 22) och KAKAL (se Bild 21). Distributionen av de blåa neuronerna på båda bilderna är mycket lika varandra och på båda komponentplanen finns den röda neuronen som indikerar högt värde på samma ställe. Distributionen av de gröna neuronerna liknar mycket varandra på båda bilderna. Man kan sluta sig till att de biblioteksväsen som har höga kostnader för medieanskaffningar per invånare även har höga kostnader för bokanskaffningar per invånare. Det går även att säga att de bibliotek som har låga kostnader för medieanskaffningar per invånare även har låga kostnader för bokanskaffningar.

Mellan variablerna HAL (se Bild 11) och KAL (se Bild 10) kan även utläsas likheter. Man kan konstatera på basen av bilderna att de biblioteksväsen som har högt mediebestånd per invånare även har höga anskaffningar per invånare.De biblioteksväsen som har lågt mediebestånd per invånare har också låga anskaffningar per invånare.

Man kan konstatera att extremvärdena för variablerna HK (se Bild 12) och LK (se Bild 17) finns i samma neuron som är röd och belägen uppe till höger på båda bilderna. Vilket tyder på att de biblioteksväsen som har höga anskaffningar i förhållande till mediebeståndet också har hög utlåning i förhållande till mediebeståndet. Resten av neuronerna som är grön- och blåaktiga är inte helt lika distribuerade på båda bilder. Trots det kan man säga att de grönaktiga neuronerna är belägna mer till höger på båda bilderna.

Variablerna KAKALs (se Bild 21) och KALs (se Bild 10) distribution liknar varandra om man ser på de grön- och blåaktiga neuronerna. Man kan sluta sig till att de biblioteksväsen som har moderata eller låga kostnader för medieanskaffningar per invånare har också moderat eller litet mediebestånd per invånare.

Likheter finns i distributionen av låga värden mellan variablerna TKKL (se Bild 18) och TKFK (se Bild 15). Man kan tolka det som att de kommuners biblioteksväsen som har låga omkostnader under statistikåret i förhållande till totalutlåningen också har låga omkostnader under statistikåret i förhållandet till fysiska besök.

Man kan konstatera ett omvänt samband mellan variablerna TKFK (se Bild 15) och KLAL (se Bild 16). Det är en indikation på att de biblioteksväsen som har låga omkostnader i förhållandet till fysiska besök har hög totalutlåning per invånare. Ett liknande omvänt samband finns mellan variablerna TKKL (se Bild 18) och KLAL (se Bild 16). Man kan tolka det omvända sambandet som att de biblioteksväsen som har hög totalutlåning per invånare har låga omkostnader under statistikåret i förhållande till totalutlåningen.

5.3 Klusteranalysen

Först diskuteras hur klusteranalysen gjordes innefattande metodval och valet av antalet kluster. Sedan diskuteras resultatet av analysen och vad man kan säga om de olika klustren.

5.3.1 Skapandet av klustren

Valet av antalet kluster är en aning godtyckligt eftersom den baserar sig en tolkning av visualiseringar. I avhandlingen baserar sig valet på hierarkisk klusteranalys och på K-means klusteranalys. Den självorganiserande karta som producerats i avhandlingen har sammanlagt 30 neuroner vilket innebär att det maximalt kan finnas 30 kluster. Minimalt kan det finnas ett kluster som i detta fall skulle innehålla kartans samtliga neuroner.

Först gjordes den hierarkiska klusteranalysen med Ward-metoden som resulterade i ett dendrogram. På basen av dendrogrammet fattades beslutet att använda fem kluster för den hierarkiska klusteranalysen. Dendrogrammet illustreras på Bild 23, man kan se på bilden att andra klusterantal även var möjliga. Man kunde ha indelat kartan i två stora kluster (röda strecket på Bild 23), fem kluster (gröna strecket på Bild 23) eller fler. Valet av fem kluster ger två små, ett medelstort och två större kluster.

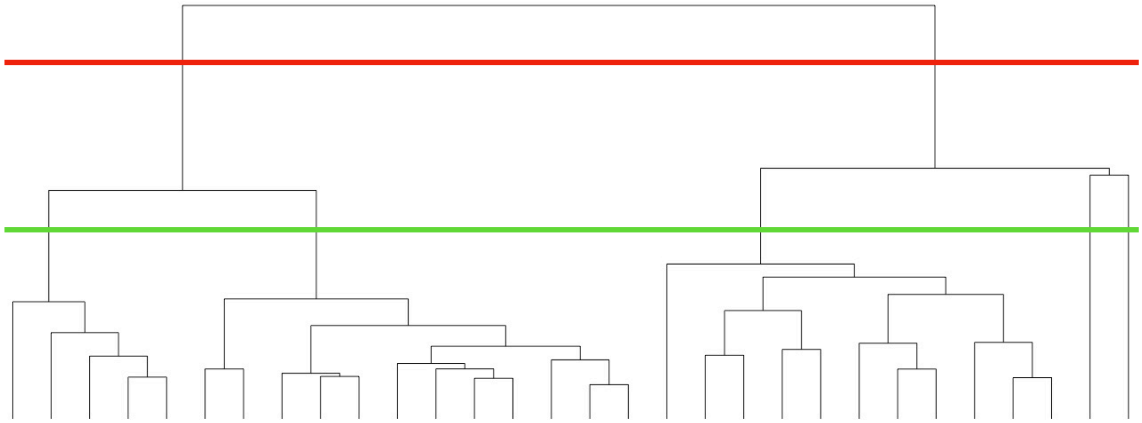

Bild 23. Den hierarkiska klusteranalysens dendrogram.

Efter den hierarkiska klusteranalysen gjordes en K-means klusteranalys. På Bild 24 kan man se resultatet av analysen. Det gäller att hitta ett armbågsveck i grafen och därmed få veta antalet kluster. Grafen behöver inte endast ha ett veck utan kan ha flera. Av Bild 24 kan man utläsa att två kluster var ett alternativ (röda strecket på Bild 24) och fem kluster var ett alternativ (gröna strecket på Bild 24). Det är även möjligt att argumentera för flera kluster än två eller fem, men avhandlingen använder sig av fem kluster för K-means klusteranalysen.

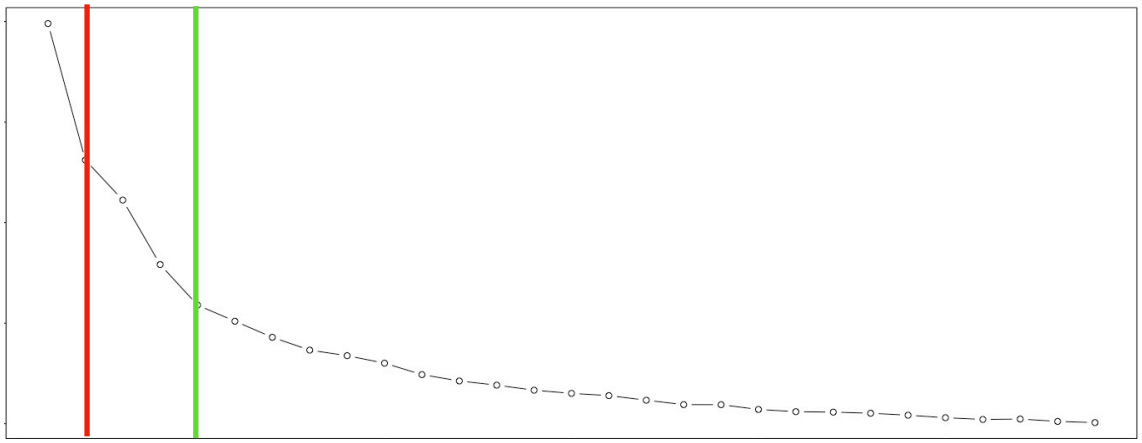

Bild 24. K-means klusteranalysen.

Utöver hierarkisk och K-means klusteranalys gjordes även en Silhouette-analys på indata för att få reda på det optimala antalet kluster. Silhouette-analysens optimala antal kluster var två stycken.

5.3.2. Tolkning av klustren

I föregående avsnitt diskuterades skapandet av klusteranalysen, de metoder som använts och de val som gjorts. Efter att antalet kluster bestämts delades och färgsattes avhandlingens självorganiserande karta med både den hierarkiska och K-means klusteranalysernas klusterindelning i två skilda visualiseringar. Utöver kartans indelning enligt klusteranalyserna lades även till en gräns mellan klustren.

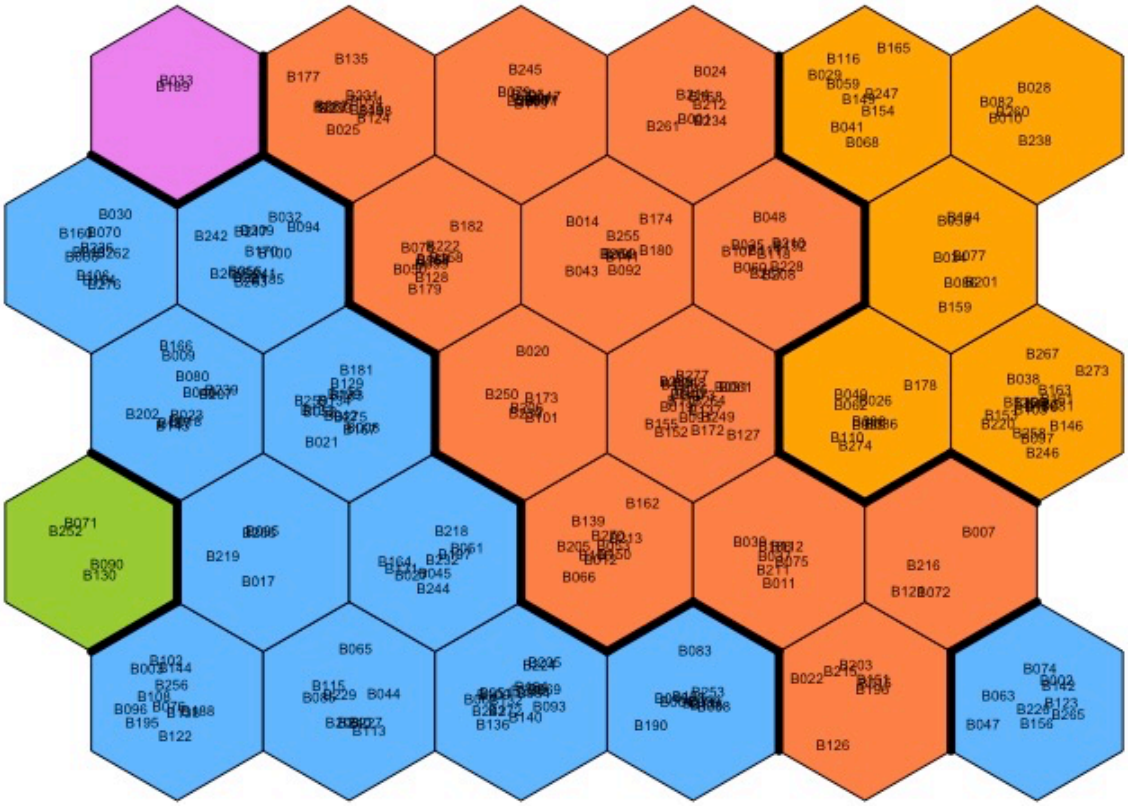

Bild 25. Visualiseringen av hierarkiska klusteranalysen.

På bilderna (se Bild 25 och Bild 26) kan man se klusterindelningen och kommunernas kod. Man kan se att två neuroner till vänster bildar två skilda en neurons kluster. I mitten av bilderna finns två större kluster och uppe till höger finns ett kluster. Man kan konstatera att klusterindelningen är identisk för både hierarkisk och K-means klusteranalys.

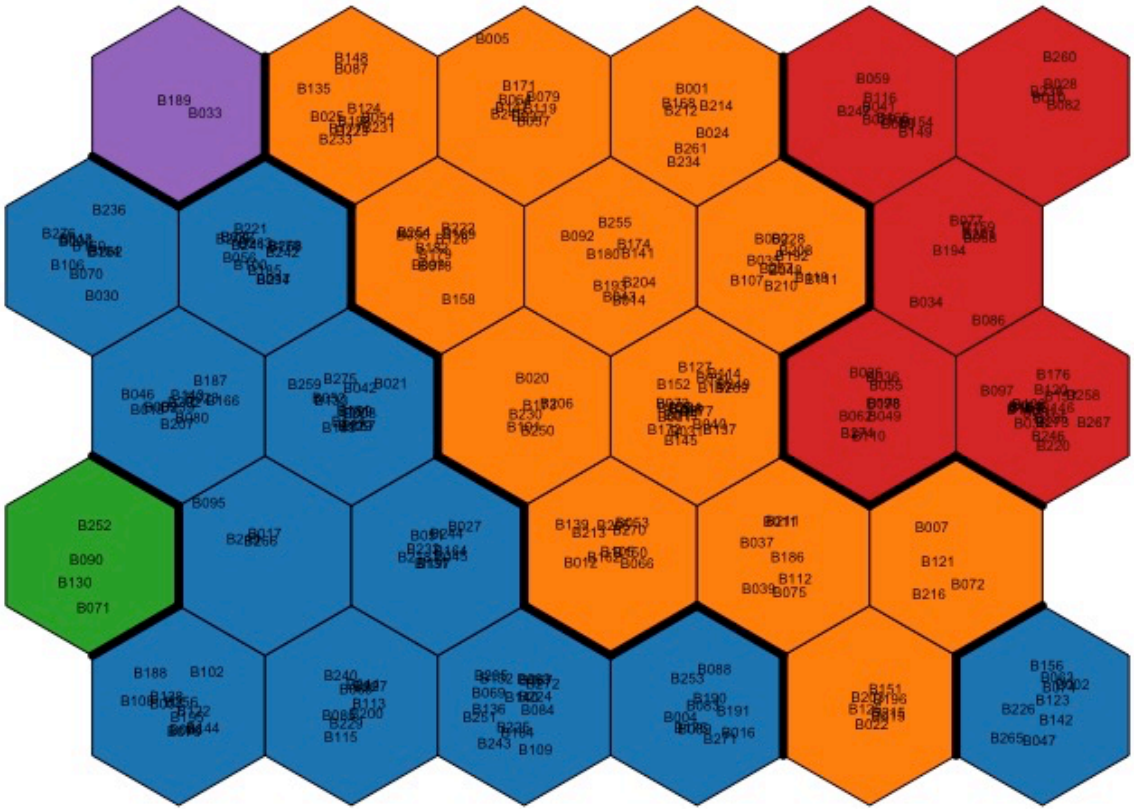

Bild 26. Visualisering av K-means klusteranalysen.

I Tabell 4 kan man se vilka kluster inom den hierarkiska klusteranalysen motsvarar vilka kluster inom K-means klusteranalysen. Man kan även se numret på klustret, vilket används i avhandlingen för att hänvisa till individuella kluster. Antalet kommuner per kluster är även det beräknat i tabellen och är identisk för både den hierarkiska och K-means klusteranalyserna.

Hierarkisk	K-means	Kluster Nr	Antal kommuner
Blå, ljus	Blå, mörk	1	114
Orange	Orange	2	110
Grön, ljus	Grön, mörk	3	4
Gulaktig	Röd, mörk	4	47
Violett, ljus	Violett, mörk	5	2

Tabell 4. Klusterindelningen för hierarkisk och K-means klusteranalys.

Om man jämför klusterindelningen enligt Bild 25 och 26 med U-matrisen (se Bild 7) kan man se att kluster 1 är inbördes relativt heterogent. Klustren 4 och 2 en betydligt mindre grad av heterogenitet medan klustren 3 och 5 är homogena. Kluster 1 har avvikande neuroner på andra och sista raden längst till vänster. Inom de två neuroner finns kommuner som t.ex. Taivassalo (B236) kluster 1och Pyhtää (B186) i kluster 2 På basen komponentplanen i det föregående avsnittet kan man säga att kommunerna i kluster 1 spenderar aningen mera än kommunerna i kluster 2. I sin tur har kommunerna i kluster 2 aningen högre utlåning än kommunerna i kluster 1. Tillsammans finns 224 kommuners biblioteksväsen i klustren 1 och 2.

Klustren 3 och 5 är säregna i det hänseendet att båda innehåller få kommuner och båda omfattar endast en neuron var. Kluster 3 omfattar kommunerna Utsjoki (B252), Kivijärvi (B090), Luhanka (B130) och Kaskö (B071). Alla tre kommuner är små Utsjoki har 1241 invånare, Luhanka har 756 invånare och Kaskö har 1296 invånare (siffrorna tagna från tilastot.kirjastot.fi och gäller år 2017). Det som karaktäriserar kluster 3 är att kommunerna där satsar mycket på sina biblioteksväsen i förhållande till sitt befolkningsunderlag. Besökssiffrorna är moderata till låga och låg utlåning i förhållande till befolkningen.

Kluster 5 omfattar endast två kommuner: Pyhäranta (B189); Hyrynsalmi (B033). Kommunerna har 2075 respektive 2406 invånare. Klustret karaktäriseras av att ha lägsta utlåningen i förhållande till befolkningen och högsta omkostnaderna per fysiskt besök. Klustret har öven höga omkostnader i förhållande till befolkningen och lägsta antalet besök i förhållande till befolkningen.

Det som är intressant med kluster 4 är att alla landets storstäder finns här. Med storstäder avses här Tammerfors (B238), Helsingfors (B028), Vanda (B260), Åbo (B246) och Esbo (B010). Helsingfors, Tammerfors, Vanda och Esbo finns i samma neuron. Neuronen finns på översta raden längst till höger på bilderna 22 och 23. I neuronen finns även Kervo (B082). Samtliga kommuner i neuronen, utom Tammerfors, finns i Nyland. Det som karaktäriserar neuronen är höga anskaffningar och hög utlåning

i förhållandet till befolkningen. Dessutom har neuronen även högst anskaffningar i förhållande till mediebeståndet. I kluster 4 finns även andra större städer som t.ex. Uleåborg (B163), Villmanstrand (B116), Lahtis (B110), Nokia (B154), Joensuu (B049), Kuopio (B103) och Jyväskylä (B055).

Åbo (B246) finns inte i samma neuron, men nog i kluster 4. Det som karaktäriserar Åbos biblioteksväsen i förhållande till de andra nämnda storstäderna är moderata anskaffningar och utlåning i förhållande till mediebestånd. Dessutom har de moderata fysiska besök under öppettimmarna, men mycket hög totalutlåning i förhållande till befolkningen. I samma neuron tillsammans med Åbo finns även bl.a. Ii (B038), Vichtis (B267), Ylivieska (B273) och Muurame (B146). I helhet karaktäriseras kluster 4 av låga kostnader: HENKKAL (se Bild 17); TKKL; (se Bild 15) KAKAL (se Bild 18); KHKAL (se Bild 19); TKFK (se Bild 12). Dessutom har de litet mediebestånd i förhållande till befolkning och medelhög besöksfrekvens.

5.4 Sammanfattning

I kapitlet har resultaten av analyserna presenterats och diskuterats. Dessutom har processen för skapandet av den självorganiserande kartan beskrivits och diskuterats. Resultaten delades in i två grupper varav den första kruppen handlade om komponentplanen och de variabler de illustrerar. Den andra gruppen beskriver klustren och vad man ska säga om dem och de kommuner de innehåller.

6. DISKUSSION OCH SLUTSATSER

I kapitlet besvaras forskningsfrågorna med hjälp av resultaten i kapitel 5 och värdet av självorganiserande kartor för denna typ av analys diskuteras. Dessutom diskuteras avhandlingens begränsningar och möjlig framtida forskning. Till slut diskuteras hur CRISP-DM använts i avhandlingen.

6.1 Forskningsfrågorna

Avhandlingen har två forskningsfrågor som med hjälp av analysen i kapitel 5 besvaras: 1) Går det med hjälp av självorganiserande kartor att hitta biblioteksväsen vars verksamhet på basen av nyckeltalsanalys skiljer sig från övriga biblioteksväsen?; 2) Vad karaktäriserar de stora städernas biblioteksväsen i förhållande till de andra kommunernas biblioteksväsen? Båda frågorna har lika stor vikt för avhandlingen och besvaras i den ordning resultaten presenterats i kapitel 5.

På den första frågan om det går att med hjälp av självorganiserande kartor att hitta biblioteksväsen vars verksamhet på basen av nyckeltalsanalys skiljer sig från övriga biblioteksväsen kan man svara att det går. Kommunerna i klustren 3 och 5 är få till antalet och små till invånarantalet. Dessutom fanns båda kluster med samma kommuner enligt både hierarkisk och K-means klusteranalys. Kluster 5 karaktäriseras av att ha lägst utlåning i förhållande till befolkning och högsta omkostnaderna per fysiskt besök. Klustret har även höga omkostnader i förhållande till befolkningen och lägsta antalet besök i förhållande till befolkningen. Kluster 3 i sin tur karaktäriseras av att kommunerna där satsar mycket på sina biblioteksväsen i förhållande till sitt befolkningsunderlag. Besökssiffrorna är moderata till låga och låg utlåning i förhållande till befolkningen. Båda klustren satsar mycket på sina biblioteksväsen men det har inte resulterat i höga besökssiffror eller hög utlåning.

På den andra frågan kan man svara att de stora städerna finns alla i samma kluster och Helsingfors, Esbo, Vanda och Tammerfors finns t.o.m. i samma neuron. Åbo finns inte i

samma neuron utan i en neuron med bl.a. Ii (B038), Vichtis (B267), Ylivieska (B273) och Muurame (B146). Man kunde ha förväntat sig att Åbo och de andra storstäder funnits i samma neuron eftersom Tammerfors, Vanda, Helsingfors och Esbo gör det. Det som karaktäriserar kluster 4 är låga kostnader och medelhöga till höga besökssiffror. Det stora befolkningsunderlaget och korta avstånd jämfört med landsbygden kan förklara besökssiffrorna.

Numerärt flest kommuner har kluster 1 och 2 vilket man kan tolka som att de flesta kommuner följer samma mönster och gör likartade satsningar. Klustren 3:s och 5:s kommuner satsar mycket på sina biblioteksväsen i förhållande till befolkningen men besökssiffrorna är moderata till mycket låga. Det kan möjligen förklaras med låg befolkningstäthet och långa avstånd.

6.2 Värdet av självorganiserande kartor

Självorganiserande kartor som analysmetod för biblioteksdata fungerar bra. Med hjälp av kartornas komponentplan kan man se vilka vilka variabler som har påverkat vilka neuroner. Dessutom är det möjligt att dela in kartan i olika kluster och genom att jämföra klusterindelningen med komponentplanen se vilka variabler som påverkat vilka kluster. På de sättet kan man få fram olika egenskaper och karaktäristika hos klustren. Av värde har även varit att man kan se i vilka neuroner och kluster kommunernas biblioteksväsen hamnat i. Nyttan med det är att man kan säga något om kommunernas biblioteksväsen på en individuell plan och inte bara de allmänna biblioteken som helhet.

På ett mera allmänt plan kan man säga att visualisering av stora mängder data eller en mindre mängd data ger ett mervärde. Data sett i t.ex. en uppsättning Excel-tabeller ger inte samma värde som en visualisering. Det heter att “en bild säger mer än tusen ord” vilket stämmer in på visualisering av data.

6.3 Begränsningar

Det som begränsar denna avhandling är att analysen av allmänna bibliotekens data endast innefattar år 2017. Konsekvensen av denna begränsning är att analysen endast fungerar som en ögonblicksbild. Om man tagit flera år i beaktande kunde man ha sett hur en kommuns biblioteksväsen förändrats och utvecklats under åren. Dessutom kunde man ha sett i fall klustren ändrats under åren. För att göra det kunde man ha använt sig av flera ögonblicksbilder, t.ex. gjort en en separat analys för olika år och jämfört dem med varandra eller använts sig av en självorganiserande tidskarta.

En annan begränsning är att analysen endast baserar sig på nyckeltal och inte all tillgänglig data. Hade man använt all data kunde komponentplanen och klusteranalyserna se annorlunda ut.

6.4 Framtida forskning

I framtiden kunde en mera ambitiös analys göras som innefattar all data och sträcker sig bakåt i tiden. Alternativt kunde man välja en uppsättning variabler som sträcker sig bakåt i tiden för att sedan användas i prediktiv mening för att förstå biblioteks-väsendenas framtid. Nyttan med en dylik analys kunde vara att den skulle ligga till underlag för beslutsfattande och budgetering.

En annan möjlighet kunde vara att titta på de små kommunernas biblioteksväsen och förutspå något om deras framtid. Man kunde även titta på hur bibliotekens mediebestånd utvecklas och hur utlåningen av olika media utvecklas. Nyttan med det kunde vara att se ifall det lönar sig att satsa på tryck eller elektroniskt material.

6.5 CRISP-DM

I avhandlingen användes CRISP-DM modellen för datautvinning. Modellen innefattar sex faser för datautvinningsprocessen. Första fasen som kallas *Affärsförståelse* utfördes

genom att forskningsfrågorna formulerades och med att självorganiserande kartor valdes som metod. Andra fasen som kallas *Datakännedom* utfördes i samband med att tilastot.kirjastot.fi databasens data utforskades. Syftet med den utforskningen var att se vilka variabler som finns, vilken tidsperiod data omfattar och vilka kommuner som fanns med. Tredje fasen som kallas *Dataförberedning* innefattar valet av variabler, kodning av kommunnamnen och exporten av data till R.

Fjärde fasen som kallas *Användande av modeller* innefattar själva skapandet av en självorganiserande karta, komponentplanen och klusteranalyserna. Femte fasen som kallas *Utvärdera* innefattar tolkningen av resultatet av analysen. Sjätte och sista fasen kallas *Tillämpning* innebär svarandet på de forskningsfrågor som avhandlingen vill svara på.

7. KÄLLFÖRTECKNING

Azam, I., Sohrawardi, J., Das, H.S., Alam, M.S., Alvy, M.S. och Rahman, R.M. (2013). Bibliomining on North South University Data. I *Eighth International Conference on Digital Information Management (ICDIM 2013)*, Islamabad, Pakistan, 10-12 September.

Azevedo, A. and Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. I *Proceedings of the IADIS European Conference on Data Mining 2008*, Amsterdam, Nederländerna, 24-26 Juli, ss. 182–185.

Ball, R. (2018). *An Introduction to Bibliometrics (1st ed.)*. Kidlington, UK: Elsevier Limited.

Berson, A., Smith, S. och Thearling, K. (2000). *Building Datamining Applications for CRM.* New York: McGraw-Hill.

Bramer, M. (2016). *Principles of Data Mining (3rd ed.)*. London: Springer-Verlag London Limited

Brown, M.S. (2014). *Data Mining for Dummies*. Hoboken, NJ: John Wiley & Sons, Inc.

De Bellis, N. (2009). *Bibliometrics and Citation Analysis*. Lanham, MD: Scarecrow Press, Inc.

Demirhan, A., & Güler, İ. (2011). Combining stationary wavelet transform and self-organizing maps for brain MR image segmentation. *Engineering Applications of Artificial Intelligence*, **24**(2), ss. 358-367.

Desmet, P. (2001). Buying behavior study with basket analysis: pre-clustering with a Kohonen map. *European Journal of Economic and Social Systems,* **15**(2), ss. 17-30.

Eklund, T. (2004). *The Self-Organizing Map in Financial Benchmarking.* TUCS Dissertations No 56. Turku, Finland: Turku Centre for Computer Science.

Evans, J. (2017). *Business Analytics*. Harlow, England: Pearson Education Limited.

Fayyad, U., Piatetsky-Shapiro, G. och Smyth, P. (1996). Knowledge discovery and data mining: towards a unifying framework. I E. Simoudis, J. Han and U. Fayyad (Red.), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96),* Portland, OR, 2-4 Augusti: AAAI Press, ss. 82-88.

Holmbom, A.H. (2015). *Visual Analytics for Behavioral and Niche Market Segmentation*. TUCS Dissertations No 195. Turku, Finland: Turku Centre for Computer Science.

Jiang, Z. och Carter, R. (2018). Visualizing library data interactively: two demonstrations using R language. *Library Hi Tech News,* **35**(5), ss. 14-17.

Khabaza, T. (2010). *9 Laws of Data Mining*. Hämtat 6 December, 2018, från http://www.khabaza.com.

Kiviluoto, K. (1996). Topology Preservation in Self-Organizing Maps.
Kohonen, T. (2001). *Self-Organizing Maps (Third Edition).* Berlin: Springer-Verlag.

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, **37**, ss. 52-65.

Kovacevic, A., Devedzic, V. och Pocajt, V. (2009). Using data mining to improve digital library services. *The Electronic Library*, **28**(6), ss. 829-843

Li, T., Sun, G., Yang, C., Liang, K., Ma, S. och Huang, L. (2018). Using self-organizing map for coastal water quality classification: Towards a better understanding of patterns and processes. *Science of The Total Environment,* **628–629**, ss. 1446-1459.

Mariscal, G., Marbán, O. och Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review,* **25**(2), ss. 137-166.

Mourya, S.K. och Gupta, S. (2012). *Data Mining and Data Warehousing*. Oxford, UK: Alpha Science International

Nicholson, S. och Stanton J. (2003). Gaining Strategic Advantage through Bibliomining: Data Mining for Management Decisions in Corporate, Special, Digital, and Traditional Libraries. I H. Nemati och C. Barko (Red.), *Organizational data mining: Leveraging enterprise data resources for optimal performance* (ss. 247-262). Hershey, PA: Idea Group Publishing.

Nicholson, S. (2003). The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. *Information Technology and Libraries,* **22**(4).

Nicholson, S. (2006). The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. *Information Processing and Management,* **42**, ss. 785 - 804.

Olson, D.L. och Delen, D. (2008). *Advanced Data Mining Techniques (2nd ed.)*. Berlin, Heidelberg: Springer-Verlag.

Pierson, L. (2017). *Data Science for Dummies (2nd ed.).* Hoboken, NJ: John Wiley & Sons, Inc.

Prichard, A. (1969). Statistical Bibliography or Bibliometrics?. *Journal of Documentation,* **25**(4), Dec 1969, ss. 348–349.

Pölzlbauer, G. (2004). Survey and Comparison of Quality Measures for Self-Organizing Maps. I *Proceedings from the Fifth Workshop on Data Analysis,* Tatranska Polianka, Slovakien, 24-27 Juni.

Sarlin, P. (2013a). Self-Organizing Time Map: An Abstraction of Temporal Multivariate Patterns, *Neurocomputing*, **99**(1), ss. 496-508.

Sarlin, P. (2013b). Replacing the time dimension: A Self-Organizing Time Map over any variable. I *Proceedings of the Workshop on New Challenges in Neural Computation (NC^2),* Saarbrucken, Germany, ss. 17-24.

Sarlin, P. och Eklund, T. (2013). Financial performance analysis of European banks using a fuzzified Self-Organizing Map. *International Journal of Knowledge-based and Intelligent Engineering Systems,* **17**, ss. 223–234

Sung, J. och Tolppanen, B. (2013). Do Library Fines Work?: Analysis of the Effectiveness of Fines on Patron's Return Behavior at Two Mid-sized Academic Libraries. *The Journal of Academic Librarianship,* **39**(6), ss. 506-511.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent data analysis,* **3**(2), ss. 111-126.

Vesanto, J., Sulkava, M. och Hollmén, J. (2003). On the Decomposition of the Self-Organizing Map Distortion Measure. I *Proceedings of the workshop on self-organizing maps,* Kitakyushu, Japan, 11-14 September.

Vesanto, J., Himberg, J., Alhoniemi, E. och Parhankangas, J. (2000). *SOM Toolbox for MATLAB 5; Report A57*. Espoo, Finland: Libella Oy.

Wang, C., Chen, L., Xu, S. och Chen, X. (2017). Exposing Library Data with Big Data Technology: A Review. I *IEEE/ACIS 15th International Conference on Computer and Information Science*, Juni 2016, ss. 1-6.

Yao, Z., Sarlin, P., Eklund, T. och Back, B. (2012). Combining visual customer segmentation and response modeling. I *Proceedings of the 20th European Conference on Information Systems*, Barcelona, Spain, 10-13 Juni.

Yao, Z. (2013). *Visual Customer Segmentation and Behavior Analysis: A SOM-Based Approach.* TUCS Dissertations No 163. Turku, Finland: Turku Centre for Computer Science.

Zaugg, H., McKeen, Q., Hill, B. och Black, B. (2017). Conducting and Using an Academic Library Data Inventory. *Technical Services Quarterly,* **34**(1), ss. 1-12.