TIEDONLOUHINTA VAKUUTUSAINEISTOSTA ASIAKKUUDENHALLINNASSA

Artturi Alaharjula

28.01.2005

Joensuun yliopisto
Tietojenkäsittelytiede
Pro gradu -tutkielma

Tiivistelmä

Ennen oli asiakas, joka osti palvelua tai tuotteita sen enempää vaatimatta palvelua tai erilaisempia tuotteita. Nykyisin yritysten toiminnan siirryttyä yhä enemmän asiakaskeskeiseen toimintamalliin asiakas esittää tarpeen ja yritys yrittää löytää siihen sopivan palvelun tai tuotteen. Tällöin menestyy se yritys, joka pystyy parhaiten vastaamaan tähän tarpeeseen. Tutkielmassa lähestytään aluksi asiakkuudenhallintaa luomalla katsaus kirjallisuudessa esiintyviin tiedonlouhinnan menetelmiin ja tekniikoihin.

Tutkielmassa keskitytään asiakkuudenhallintaan vakuutusyhtiössä. Työn tarkoitus on löytää irtisanotuista vakuutuksista sellaisia asiakkaiden piirteitä, joiden avulla vakuutusyhtiön olisi mahdollista ennustaa jo etukäteen vakuutuksen irtisanomista harkitseva tai muuten yhteydenottoa tarvitseva asiakas. Aineistona käytetään vahinkovakuutusaineistoa, jonka päättyneiden vakuutusten sekä niiden muuttujien lukumäärää pienennetään eri ryhmittelyissä (3 ryhmittelyä). Ryhmittely toteutetaan tiedonlouhinnalla, ja algoritminä käytetään satunnaista paikallishakua.

Tutkimuksestani ilmeni lopuksi, että aineistoa ei pystytty ryhmittelemään riittävän tarkasti, jotta siitä olisi löydetty lopputuloksena jotakin täydellisesti uutta ja mielenkiintoista tietoa asiakaspoistuman tueksi. Ryhmittelyn lopputuloksissa havaittiin olevan suuriakin korrelaatioita joidenkin muuttujien kesken, mutta aineistosta johtuvia riippuvuuksia näiden muuttujien välille ei ollut helppo löytää. Kuitenkin tämä ryhmittely vahvisti aikaisempaa uskomusta siitä, että asiakaspoistuman aiheuttaa usein unohdettu asiakas. Lopputuloksena voidaan todeta, että tiedonlouhinta on varmasti hyvä työkalu asiakkuuksienhallintaan ja tässäkin tapauksessa hieman erilaisemmalla muuttujien valinnalla olisi lopputulos ollut varmasti toisenlainen.

Avainsanat: tietokanta, tietomallit, tiedonlouhinta, asiakkuudenhallinta, ryhmittelyanalyysi, itseoraganisoituva kartta, k-means, random local search.

Esipuhe

Alkusyyn tälle tutkielmalle antoi erikoistyö, jossa tutkin ajoneuvovahinkoasiakkaiden, ajoneuvojen ja vahinkotilanteiden ominaisuuksien samankaltaisuuksia eri vahingoissa. Aineistosta löytyi mielenkiintoisia ryhmiä, joissa oli mukana ehdottomasti niihin kuuluvia ominaisuuksia mutta myös käsittämättömiä toisen ääripään ominaisuuksia. Ryhmittelyanalyysin algoritmina käytin silloin itseorganisoituvaa karttaa, jonka toiminnasta en kuitenkaan vakuuttunut riittävästi. Tämän johdosta kehittyi mielenkiinto tutkia laajemmin asiakkuuksienhallintaa vakuutusyhtiössä, johon käytettäisiin yhtä nykyisistä parhainta ryhmittelyanalyysi-algoritmia: satunnaista paikallishakua (RLS).

Ohjaajani professori Pasi Fräntiä haluan kiittää vahvasta mukanaolosta ja aidosta kiinnostuksesta työtäni kohtaan. Jossakin vaiheessa työn motivaation hiipuessa olisi ilman hänen aktiivista otettaan tämä työ saattanut jäädä loppuusaattamatta.

Lopuksi haluan sanoa suuren kiitoksen perheelleni tuesta ja kannustuksesta työn parissa vietettyjen tuntien vähentäessä yhteistä ajankäyttöä. Lisäksi haluan lausua kiitokset kannustuksesta kavereilleni, jotka jaksoivat aina ja aina vain kysyä työn valmistumisen ajankohtaa.

SISÄLLYSLUETTELO

1 Johdanto

Nopea tekninen kehitys tietotekniikassa ja elektroniikassa on mahdollistanut suurten tietomäärien keräämisen sekä varastoimisen erilaisiin massamuisteihin. Tietojen käsittelystä ja hallinnasta on muodostunut siten yksi yritysten tärkeimmistä tietotekniikan sovellusalueista. Nämä yritysten rekistereissä olevat tiedot varastoidaan suuriin tietovarastoihin, joista sitten aina tarvittaessa louhitaan tallentajan liiketoimintaa hyödyttävää tietoa. Näistä hyvänä esimerkkeinä ovat vakuutusyhtiöiden, pankkien, lentoyhtiöiden, yliopistojen ja kauppojen asiakasrekisterit. Mikäli tämä louhinta ei olisi teknisesti mahdollista, tietojen yksityiskohtaisempi kerääminen olisi turhaa. *Tiedonlouhinta-menetelmissä* etsitään hyödyllistä tietoa analysoimalla tietovarastoa automaattisesti tai puoliautomaattisesti. Perinteisistä analyyseistä poiketen louhinnalla on mahdollista saada esille sellaista tietoa, jota käyttäjä ei olisi tullut edes heti ajatelleeksi. Tiedonlouhinnan arvioidaan lisääntyvän tulevaisuudessa merkittävästi, koska digitaaliset aineistot kasvavat ja laskentamenetelmät sekä niiden kapasiteetit kehittyvät ohjaamaan uusia käyttäjiä tietoanalyysien soveltamiseen. Merkittäviä uusia sovellusalueita nykyisten rinnalle voisin kuvitella löytyvän lääketieteestä kuten esimerkiksi sairauksien ennustamisessa.

Vakuutusyhtiöillä ja pankeilla ei voi olla tuntematonta asiakasta kuten päivittäis-tavarakaupoilla. Näissä yhtiöissä ei voi asioida nimettömänä, mutta jokainen meistä voi kuitenkin ostaa käteisellä kahvipaketin ilman kanta-asiakaskorttia. Vakuutusyhtiöiden toiminnan luonteesta johtuen ne tallentavat ja keräävät satoja tietueita asiakkaastaan vakuutuksen myöntämisen yhteydessä sekä myöhemmin sen vahinko- ja vakuutushistorian aikana.

Asiakkaan saapuessa vakuutusyhtiön palvelupisteeseen myyntitavoitteen omaava vakuutus-virkailija hymyilee kohteliaasti ja tarjoaa vakuutustuotteita uudelle asiakkaalleen. Mikäli kysymyksessä on asiakkaan luokse saapuva asiamies, hän näkee edessään mahdolliset vakuutuksesta saatavat hankintapalkkiot. Tällöin vakuutuksen myymiseen liittyvät motivaatiot ovat molemmilla kohdallaan ja ehkä suurin vaara on, että kauppa tehdään hinnalla millä hyvänsä välittämättä asiakkaan riskiprofiilista. Virkailijan myyntitavoite ja asiamiehen hankintapalkkio kertyvät vasta sen jälkeen, kun asiakas on maksanut vakuutuksensa ensimmäisen erän.

Kuitenkin asiakkaan poistuessa vakuutusvirkailija saattaa miettiä, onko asiakkaan maksuhäiriörekisterissä merkintöjä vai tekikö hän turhaa työtä ja tuntuivatko asiat etenevän liian helposti ja jättäisikö asiakas vakuutusmaksunsa maksamatta. Hän saattaa miettiä myös, miksi asiakkaalla ei ollut irtisanottavaa kotivakuutusta toisessa yhtiössä tai oliko asiakas riskialttiimpi kuin mitä riskivalintaohjeissa on mainittu. Asiakas on taas onnellinen, kun vakuutus tuli nyt vihdoin ja viimeinkin otettua ja riski siirtyi vakuutusyhtiölle. Mahdollista on myös, että asiakkaan entinen vakuutus oli irtisanottu maksamattomana toisesta yhtiöstä ja nythän vakuutus on taas tehty voimaan joksikin aikaa. Asiakkaalle on voinut myös sattua vahinko jo edellisenä päivänä, jonka seurauksesta vakuutus piti tehdä nyt ja mahdollisesti vakuutuspetokseen taipuvaisella asiakkaalla olisi mahdollisuutta hakea korvausta ensi viikolla.

Mikäli tämä asiakas osoittautuu hyväksi asiakkaaksi maksamalla vakuutusmaksunsa ja vahinkoja sattuu vähän, vakuutusyhtiönkin pitäisi alkaa miettiä muuta vakuutustarjontaa ja asiakasetuja tälle hyvälle asiakkaalleen, jotta tämä sitoutuisi jatkossakin olemaan asiakkaana. Juuri tämän uuden asiakkaan hankkiminen maksaa vakuutusyhtiölle yllättävän paljon, koska kustannukset koostuvat markkinoinnista, hankintapalkkioista ja vakuutuksen käsittely-kustannuksista. Laskelmin voidaan osoittaa taulukon 1.1 mukaan, että tariffitasolla myyty vakuutus tulee uuden asiakkaan kohdalta vakuutusyhtiölle kannattavaksi vasta 4 vuoden päästä. Tämä laskelma perustuu kokonaiskulusuhteen (korvausten, hankinta- ja hoitokulujen suhde vakuutusmaksuihin) jäämiseen alle 100 %:n, jonka saavuttaminen on vakuutusyhtiöillä kuitenkin käytännössä harvinaista. Muutoin vaikutus vakuutuksen kannattavuuden saavuttamiseen on vuosina vielä tätäkin pidempi. Vahinkovakuutusyhtiöiden toiminnan tulos muodostuu yleensä pitkälti sijoitustoiminnan tuottojen perusteella, jonka saavuttamiseksi käytetään taseessa olevia pääomia.

Taulukko 1.1: Uuden vakuutuksen kannattavuus tariffitason mukaan.

(vakuutusmaksujen kasvu 2 % , korvaukset 60 %, hankintakulut 70 % ja hoitokulut 23 % vakuutusmaksuista sekä sijoitustoiminnan tuotto 6 % vakuutuskatteesta)

	Vakuutusmaksut	Korvaukset	Hankintakulut	Hoitokulut	Vakuutuskate	Sijoitustoiminta	Tulos	Kertymä
1. Vuosi	1000	-600	-700	-230	-530	-32	-562	-562
2. Vuosi	1020	-612	0	-235	173	10	183	-379
3. Vuosi	1040	-624	0	-239	177	11	188	-191
4. Vuosi	1061	-637	0	-244	180	11	191	0
5. Vuosi	**1082**	**-649**	**0**	**-249**	**184**	**11**	**195**	**195**

Tariffitason alentaminen ylimääräisellä 10 %:n alennuksella vaikuttaa siten, että kannattavuuden saavuttaminen kestää peräti 7 vuotta. Laskelmassa ei ole otettu kuitenkaan huomioon taseessa olevia vakavaraisuuspääomia sijoitustuottojen osalta. Nämä varaukset ja omapääoma parantavat sijoitustoiminnan tuottoja merkittävästi, jolloin myös uuden vakuutuksen kannattavaksi saamiseen kuluva aika lyhenee. Myöskään vuodet eivät ole aina samanlaisia vahinkosuhteiden osalta.

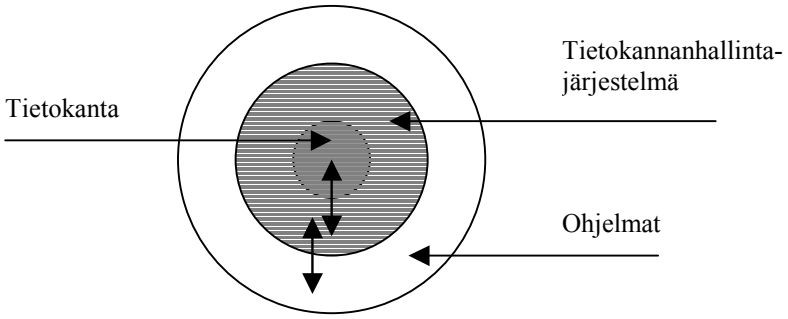
Tämän työn tavoitteena on tarkastella asiakkaan toimesta irtisanottujen vakuutusten poistuman ennustamista etukäteen vakuutuksesta tallennettujen tietojen perusteella. Nykyisissä markkinastrategioissa vanhan asiakkaan pitäisi antaa yritykselle jotakin todellista lisäarvoa. Siksi vakuutusyhtiöiden pitäisi pystyä tunnistamaan ne asiakkaat, jotka ovat kiinnostuneita hankkimaan uusia lisävakuutuksia ja keskittämään vakuutukset tai jotka harkitsevat vakuutusyhtiön vaihtamista. Näiden asiakkaiden tunnistaminen voidaan toteuttaa eri menetelmin. Tiedonlouhintaa voidaan mahdollisesti käyttää asiakkaiden tunnistamisessa, mutta tällöin on välttämätöntä ymmärtää, kuinka menetelmät toimivat, jolloin voidaan valita parhaiten sopiva menetelmä ongelman ratkaisemiseksi. Tämän työn tarkoitus on tehdä tutuksi keskeiset tiedonlouhintamenetelmät ja tarkastella tiedonlouhinnan toimivuutta vahinko-vakuutusyhtiön päättyneiden vakuutusten analyysissä.

Luvussa kaksi selvitetään lyhyesti tietokantojen perusteet ja tutustutaan tarkemmin yleisimmin käytettyihin tietomalleihin. Tämän lisäksi tarkastellaan eri tietojärjestelmien kuten operatiivisen järjestelmän, tietovaraston ja tietoanalyysin välisiä eroavaisuuksia. Kolmannessa luvussa tarkastelun kohteeksi on otettu asiakkuuksienhallinta koko elinkaaren aikana. Myös tässä luvussa tarkastellaan ryhmittelyanalyysin toimintaa esimerkkinä ostoskorianalyysin toteutus. Neljännessä luvussa käydään läpi tiedonlouhinnan periaatteet. Tämän jälkeen tutustutaan tarkemmin tietojen esikäsittelyyn ennen louhintaa ja lopuksi tarkastellaan tiedonlouhinnan eri malleja. Viidennessä luvussa tutustutaan ryhmittelyanalyysin toimintaan, jolloin tutustutaan tarkemmin SOM-, K-means- ja RLS-ryhmittelyalgoritmien toimintaan. Tutkimuksessa käytetään myöhemmin RLS-algoritmiä. Itse tutkimuksesta kertoo kuudes luku, jossa käydään aluksi läpi aineiston muodostamisesta ja käsittelystä aiheutuneet ongelmat. Lopuksi tarkastellaan eri aineistojen louhintojen lopputuloksena saatuja ryhmiä. Näiden tulosten perusteella pitäisi löytyä asiakkaita, joiden piirteiden perusteella olisi mahdollista tunnistaa jo tietovarastosta vakuutusten irtisanomista harkitsevat tai muutosta

tarvitsevat asiakkaat. Lopuksi tehdään yhteenveto saavutetuista tuloksista ja havaituista virheistä sekä tarkastellaan parannusehdotuksia tulevia louhintoja varten.

2 Tiedon hierarkia

Tietokannaksi (*database*) kutsutaan tietokoneessa tallennettuna olevaa tietovarastoa (Hyvönen & al. 1993), jota käytetään eri ohjelmien kautta. Tämän johdosta voidaan ajatella minkä tahansa tietokokoelman olevan tietokanta. *Tietokantajärjestelmä* (*database system, DBS*) sisältää tietokannan (kuva 2.1), *tietokannanhallintajärjestelmän* (*database management system, DBMS*) itse *tietokantaa hyödyntävät ohjelmistot* (*application program*) sekä laitteiston (Laine 2000). Tietokantajärjestelmä pitää sisällään loogisen ja fyysisen tietokannan (Laine 2000). Fyysinen tietokanta sisältää tiedostot, kun taas looginen tietokanta käsittää itse tietokannan tietomallin. Sovellusohjelmat käyttävät I/O-operaatioita käskyjen ja vastauksien välittämiseen loogisen tietokannan kautta fyysiseen tietokantaan. Tietokannanhallinta-järjestelmän avulla luodaan, ylläpidetään ja käytetään tietokantoja.



Kuva 2.1: Tietokantajärjestelmän eri kerrokset.

2.1 Tietokannat

Tietokannanhallintajärjestelmä on tietokannan hallintaan tarkoitettu ohjelmisto, joka mahdollistaa tietokannalle tietoriippumattomuuden muista ohjelmista, tiedon samanaikaisen käsittelyn, monipuolisen tiedonhaun, tiedon suojauksen, tiedon eheyden varmistamisen, virhetilanteista toipumisen sekä tiedon turvaamisen. Näitä tietokantaratkaisuissa usein tarvittavia ominaisuuksia ei ole perinteisillä ohjelmointikielillä toteutetuissa tiedosto-ratkaisuissa, vaan niissä tiedot on tallennettu peräkkäistiedostoihin, joita ohjelmat lukevat ja käsittelevät muistissaan. Tietokannassa olevia tietoja voidaan käyttää myös moniin eri tarkoituksiin, ja silloin on järkevää tehdä jokaista käyttötarkoitusta varten oma ohjelma, joka toteutetaan siihen parhaiten sopivimmalla ohjelmointikielellä. Kuitenkin eri ohjelmointikielet

määrittelevät käyttämänsä tiedostot hieman eri tavoin ja tämän seurauksena tiedot pitää määritellä ohjelmista ja ohjelmointikielistä riippumattomiksi. Tietokantaratkaisuissa tästä riippumattomuudesta päästään eroon käyttämällä *tietokantakaaviota* (*database schema*).

Tietokannat mahdollistavat tiedon hakemisen ja käsittelyn niiden sisällön perusteella. Esimerkiksi asiakkaan vakuutukset on voitava hakea suoraan tietokannasta ilman, että vakuutukset poimittaisiin yksitellen kaikkien vakuutuksien joukosta käyttäen apuna jotakin silmukkarakennetta. Tietokannat tarjoavat lisäksi paremman samanaikaisuuden hallinnan useammalle käyttäjälle, kun taas tiedostojen useamman käyttäjän yhtäaikainen käyttö ei ole mahdollista, koska tiedoston sisältö ladataan tietokoneen keskusmuistiin ohjelman käytettäväksi. Tällöin tiedostoa ei pystytä ottamaan käyttöön muutoksia varten, kuitenkin jotkut ohjelmistot sallivat lukuoikeuden muille käyttäjille. Tietokannoissa näitä ongelmia ei ole, koska jokaisen muutoksen yhteydessä päivitetään tietokanta.

2.1.1 Operatiiviset tietokannat

Yrityksen tiedosta suurin osa sijaitsee operatiivisten järjestelmien tietokannoissa, jotka palvelevat yrityksen jokapäiväistä toimintaa, kuten laskutusta ja kirjanpitoa. Näillä tapahtumapohjaisilla järjestelmillä on mahdollista olla olemassa monia yhtäaikaisia käyttäjiä. Tallennuskapasiteetin suurentumisen ja halpenemisen seurauksesta tietovarastojen sisällön koko on kasvanut huomattavan suureksi, jolloin niistä suoritettavien hakujen on oltava tehokkaita. Näitä järjestelmiä kutsutaan yleisesti *OLTP-järjestelmiksi (On-line Transaction Processing).*

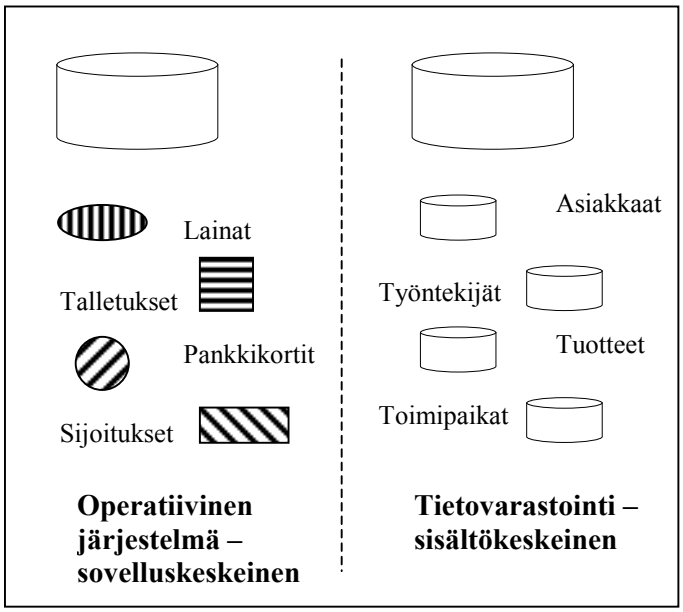
Operatiivisten järjestelmien tietokannat on suunniteltu tehokasta tapahtumakäsittelyä varten, jolloin niissä ei säilytetä historiatietoja kovinkaan kauan (Hovi 1997). Operatiivisissa tietokannoissa tiedot hajautetaan eri tietokantatauluihin, jolloin niihin kohdistuvat operaatiot, kuten rivin lisäys tai poisto, ovat nopeita suorittaa. Näihin tietokantoihin tallennetaan jatkuvasti uusia tapahtumia ja päivitetään vanhoja tietoja. Tämän seurauksesta tietojen analysoinnit, raportoinnit ja kyselyt ovat usein hitaita sekä vaikeita suorittaa. Operatiiviset kannat ovat usein raskaassa tapahtumakäytössä (vakuutusyhtiöt, pankit ja kirjastot) ja tällöin vastausaikojen tulee olla korkeintaan muutamia sekunteja.

Yleensä näissä operatiivisissa järjestelmissä on valmiina erilaisia raportteja, mutta nämä raportit eivät yleensä riitä niiden käyttäjille. Järjestelmien käyttäjien on itsenäisesti vaikea tehdä helppoja kyselyjä ja raportteja, kun operatiiviset järjestelmät eivät tue niitä rakenteeltaan. Nämä operatiiviset tietokannat eivät ole suunniteltu kyselyjä silmällä pitäen. Niinpä tietojen toistoa ei juuri vältetäkään vaan päinvastoin, tietoja toistetaan eli *denormalisoidaan*. Tällöin taulujen määrä ja samalla tarvittavien liitosten määrä pienenee, jolloin kyselyjen suorituskyky on huomattavasti nopeampi kuin täysin normalisoidun kannan. Operatiivisten järjestelmien tietovarastokantaan voidaan myös summata valmiiksi usein kysyttäviä tietoja kuten asiakkaan vakuutusmäärä ja vakuutusmaksut. Nämä summataulut ovat kooltaan murto-osia niiden tapahtumatauluista, jolloin kyselyjen nopeutuessa vastaukset saadaan heti.

2.1.2 Tietovarastot

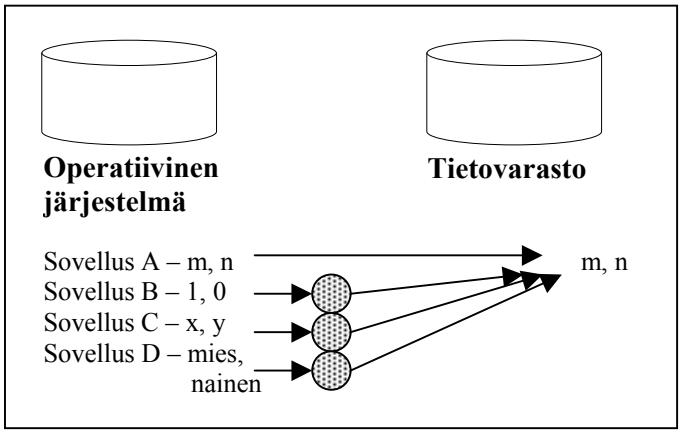
Tietovarastoksi (*data warehouse*) kutsutaan yleisesti yrityksen operatiivisista tietokannoista tai muista ulkopuolisista tietolähteistä kerättyä yhtenäistä, keskitettyä ja jalostettua tietokokoelmaa (Inmon 1996). Tietovarasto antaa yrityksen päätöksentekijöille arkkitehtuurin sekä työkalut tuottaa järjestelmällisesti moninaisia ja oikeisiin tietoihin pohjautuvia tilastoja, raportteja ja analyysejä (Han & Kamber 2001).

Tietovarastotekniikkaa (*data warehousing*) on vasta äskettäin opittu hyödyntämään yritysten päätöksenteon tukena ja perusteluna. Tietovarastojen käyttö on kasvanut yrityksissä räjähdysmäisesti, koska operatiiviset perusjärjestelmät eivät pysty riittävästi tyydyttämään analysointi- ja raportointitarpeita. Tietovarasto sijaitsee erillään operatiivisesta tietokannasta, jolloin se on yksittäinen, täydellinen ja yhdenmukainen tietovarasto. Tietovarastoa päivitetään jatkuvasti operatiivista kannoista ja sen vastausajat kyselyihin voivat olla minuuteista tunteihin (Hovi 1997). Tietovarastoinnin idean isänä pidetään USA:ssa yleisesti W. H. Inmonia, ja hänen mukaansa tietovarasto on sisältökeskeinen, integroitu, aikasidonnainen ja vakaa tietokokoelma (Inmon 2002). Nämä neljä avainsanaa erottavat tietovaraston muista tietokannoista. Sisältökeskeinen tietovarastointi on toimintatapana järjestynyt pääasioiden ympärille, mutta operatiivisten järjestelmien ympäristö on kehittynyt ohjelmien ja toimintojen ympärille (kuva 2.2).



Kuva 2.2: Sovelluskeskeisen ja sisältökeskeisen järjestelmän eroavaisuudet (Inmon 2002).

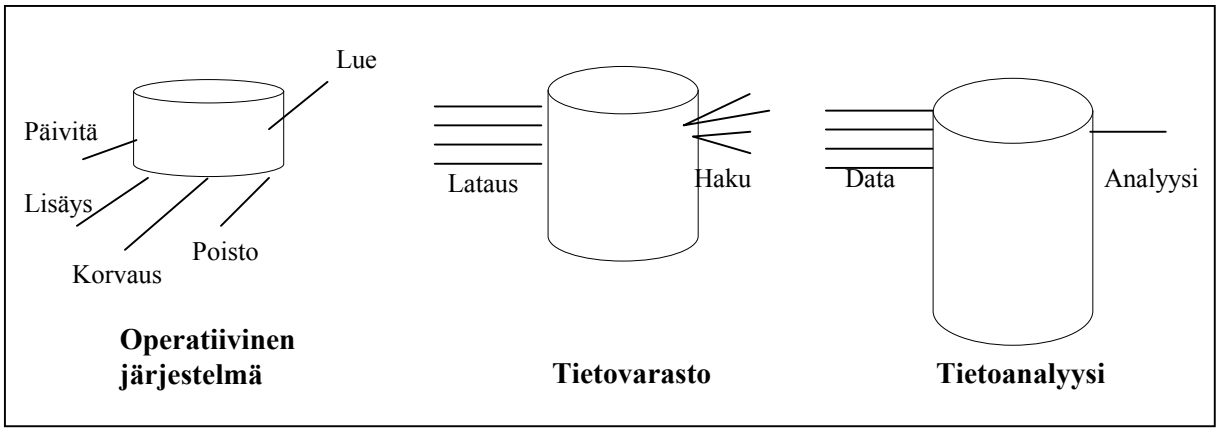
Sovelluskeskeiseen järjestelmään liittyy aina sekä tietokannan että ohjelman suunnittelu, kun taas tietovarastoinnin keskipisteenä on tiedon mallintaminen ja tietokannan hyvä suunnittelu (Inmon 2002). Yhdenmukaistaminen eli *integrointi* on aina poikkeuksetta tarpeen silloin, kun useat sovellukset käyttävät tietoa eri tavalla (Inmon 2002). Vuosien aikana suunnittelijat ovat tehneet erilaisia ohjelmistoratkaisuja, jolloin erot ovat syntyneet ohjelmointikielelle muuntamisesta, avainrakenteiden eroista, fyysisistä tunnusmerkeistä ja nimeämissäännöistä (Inmon 2002). Tämän seurauksena yksinkertainen asia voidaan esittää operatiivisessa järjestelmässä monin eri tavoin (kuva 2.3).



Kuva 2.3: Esitysmuodosta johtuvia tiedon siirron ongelmia sovelluskeskeisestä järjestelmästä (sukupuolten moninainen esitysmuoto eri sovelluksissa) tietovarastoon (Inmon 2002).

Päätöksenteon tukijärjestelmää eli *DSS* (*Decision Support Systems*) käytettäessä päähuomion pitää olla tiedon käytössä, eikä epäilyksenä tiedon uskottavuudesta ja johdonmukaisuudesta (Inmon 2002). Tästä johtuen tieto varastoidaan tietovarastoon yksilöllisesti ja yleisesti hyväksyttävällä tavalla, vaikkakin operatiiviset järjestelmät varastoivat tiedon eri tavalla. Tiedot siirretään ja tallennetaan ajoittain operatiivisista järjestelmistä erilliseen tietokantaan. Tämä tietovarasto voi sisältää tietoa pitkältä ajalta, kuten 5–10 vuotta, kun taas operatiivinen järjestelmä sisältää tietoja 2–3 kuukauden ajalta (Inmon 2002). Tällöin tietovaraston tiedot eivät ole yleensä ajan tasalla, koska siirtotiheys voi olla esim. kuukausi, viikko tai päivä. Tietojen analysointia, raportointia ja kyselyjä varten tarvitaan harvoin tiheämpää ajantasaisuutta kuin korkeintaan päivä.

Operatiivisen tietokannan päivitystoiminnot ovat säännöllisesti lisäys, poisto, korvaus ja päivitys sekä luku. Tietovarastossa perustiedon käsittely on helpompaa, koska siellä on olemassa vain kaksi toimintamenetelmää: tiedon lataus ja sen haku (kuva 2.4). Tietoanalyysissä päivitettyyn tietoaineistoon suoritetaan analyysia eli *tiedonlouhintaa*.



Kuva 2.4: Operatiivisen järjestelmän, tietovaraston ja tietoanalyysin toiminnan erot.

2.2 Tietomallit

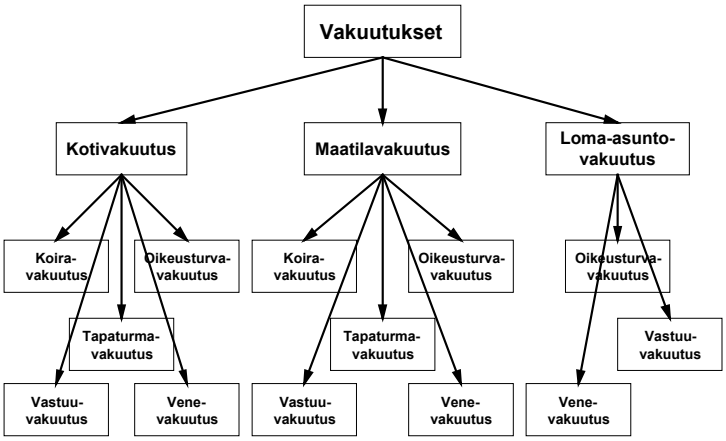
Jokainen tiedonhallintajärjestelmä perustuu jonkin tietomallin toteutukseen. *Tietomallin* (*data model)* avulla koko tietokannan sisällön tiedot ryhmitellään ja esitetään loogisesti (Hyvönen & al. 1993). Tietokantarakenteet voidaan jaotella esimerkiksi tietomallin, arkkitehtuurin, käyttötarkoituksen, käyttäjien lukumäärän tai tietomäärien mukaan. Neljä yleisintä

perustietomallin mukaista tietorakennetta ovat *hierarkkinen malli, verkkomalli, relaatiomalli* ja *moniulotteinen malli*.

2.2.1 Hierarkkinen malli

Hierarkkinen malli (*hierarchical model*) on vanhin tietomalli, joka on kehitetty jo 1960-luvulla. Tätä mallia käytettiin ensimmäisissä kaupallisissa tietokannoissa, joiden edelläkävijä oli IBM:n kehittämä *IMS/I* (*Information Management System-I*) (Ullman 1982). Tärkein toimintatapa hierarkkisilla malleilla on tietojen kuvaaminen luonnollisella tavalla, joka mukaillee yrityksen loogista toimintaa. Hierarkkiseen malliin perustuvia sovelluksia käytettiin aikaisemmin pääsääntöisesti suurkoneympäristössä aina tuotannonohjauksesta materiaali-varastonhallintaan, mutta niiden joustamattoman tiedonkäsittelyn johdosta niitä käytetään nykyisin harvoin.

Hierarkkisessa mallissa tiedot on järjestetty monitasoiseksi hierarkiaksi, jossa tietojen sidokset muodostavat puumaisen tietorakenteen (Hyvönen & al. 1993). Tietokanta muodostuu tietueista, joiden voidaan kuvata olevan erillisten puiden muodostamassa metsässä. Tietueiden väliset suhteet kuvataan linkkien avulla (1:N). Tällöin jokaisella tietueella on vain yksi ”äititietue”, mutta kuitenkin yhdellä ”äititietueella” voi olla useita ”lapsitietueita”. Tämän avulla mallissa päästään porautumaan jokaisella tasolla aina syvemmälle rakenteeseen esittäen lopulta tuote yksilöitynä (kuva 2.5). Käyttäjän on tiedettävä tarkoin hierarkkisen mallin rakenne, koska haussa käydään pahimmassa tapauksessa läpi kaikki tietokannan tietueet ennen kuin löydetään vastaus.

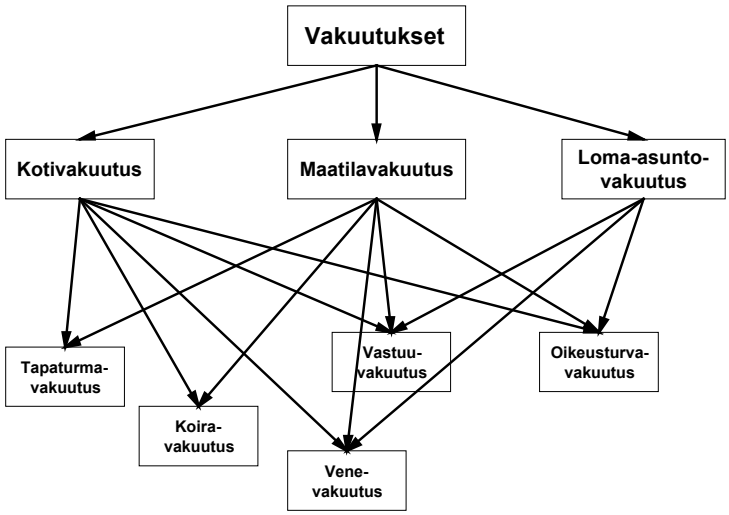


Kuva 2.5: Pelkistetty hierarkkinen malli vakuutusten tietokannasta.

2.2.2 Verkkomalli

Hierarkkisessa mallissa esiintyvät muutamat puutteet aiheuttivat *verkkomallin* (*network model*) kehittymisen seuraavana. Hierarkkisen mallin pahin puute oli, ettei se tukenut eri tietueiden välisiä monesta moneen -suhteita (N:N). Tällöin ”lapsitietueella” ei voinut olla useita ”äititietueita”, eikä myöskään ”lapsitietue” voinut olla orpo.

Verkkomallin kehittämisen aloitti CODASYL 1960-luvun lopulla (Ullman 1982). Verkkomallin menestyksen salaisuus on sen joustava ja tehokas toiminta, sekä sen toimintatapa vähentää toistojen tarvetta verrattuna hierarkkiseen malliin (kuva 2.6). Kuitenkin verkkomallin haittapuolina ja ongelmina ovat sen vaikea ymmärrettävyys sekä käytettävyys. Ylläpito aiheuttaa yleensä ongelmia, josta voidaan todeta esimerkkinä Internetin www-sivut. Mikäli joku www-sivu poistetaan käytöstä, sille sivulle jää kuitenkin osoittamaan epäkelvot linkit toisilta sivuilta. Näiden linkkien poistaminen on työlästä, ja ne aiheuttavat käyttäjille harmaita hiuksia.



Kuva 2.6: Pelkistetty verkkomalli vakuutusten tietokannasta.

Verkkomallissa tieto kootaan erillisiksi palasiksi ja palaset yhdistetään kokonaisuudeksi tätä tehtävää varten määritettyjen osoittimien avulla. Verkkomalli on tavallaan hierarkkisen mallin yleistys. Verkkomalli tukee suoraan tiedon voimakasta hajauttamista ja useat käytännön laitejärjestelmät toimivat suoraan jo fyysisen sijoittelunsa takia verkkomallin mukaisesti.

2.2.3 Relaatiomalli

Tunnetuin tietomalleista on 1970-luvulla kehitetty *relaatiomalli* (*relational model*) (Codd & al. 1993), johon perustuvat nykyisin eniten käytetyt tietokantajärjestelmät. Relaatiomalli sisältää kolme osa-aluetta: yksinkertaisen tietomallin, tehokkaan *tietokantakielen* (*SQL*) ja tietokantojen hallintajärjestelmän (Kuvaja 1995). SQL *(Structured Query Language)* on standardoitu relaatiotietokantojen rakenteinen kyselykieli, josta on käytössä kaksi erilaista standardia. Suositumpi standardi on aito ja alkuperäinen sekä huolellisesti tehty ANSI/ISO-standardin mukainen SQL (Hursch 1991). Toinen unohtuneempi standardi perustuu IBM:n 1970-luvulla kehittämälle relaatiotietokantojen tiedonmäärittely- ja käsittelykielelle, joita käytettiin IBM:n DB2-tietokannoissa (Hursch 1991).

Relaatiomallisessa tietokannassa tieto on järjestetty nimetyiksi taulukoiksi (taulukko 2.1) eli relaatioiksi (Rantanen & al. 1989). Relaatiokannat sopivat myös tietovarastoiksi, ja siksi niitä käytetään yhä useamman yrityksen operatiivisessa tiedonhallinnassa. Relaatiotietokannat ovat omimmillaan suurten luettelomaisten tietomäärien varastoinnissa ja ylläpitämisessä. Mutta relaatiokantoihin sisältyy tietovarastoinnin kannalta tarpeettomia osia, kuten lukitus, virheistä toipuminen, tapahtumien hallinta ja tapahtumaloki (Hovi 1997). Tietovarastokantahan on vain lukua varten eikä lukitus ole silloin tarpeen, koska käyttäjän ei tarvitse käsittelyn aikana päivittää tietoja. Kannan mahdollisesti vioittuessa se palautetaan kokonaan varmistus-nauhalta, jolloin tietovaraston toipuminen virheongelmista on nopeaa.

Taulukko 2.1: Pelkistetty relaatiomalli vakuutusten tietokannasta.

Vakuutukset	Lajit	Lisäturvat
	Kotivakuutus	Koiravakuutus
	Kotivakuutus	Oikeusturvavakuutus
	Kotivakuutus	Vastuuvakuutus
	Kotivakuutus	Venevakuutus
	Kotivakuutus	Tapaturmavakuutus
	Maatilavakuutus	Koiravakuutus
	Maatilavakuutus	Oikeusturvavakuutus
	Maatilavakuutus	Vastuuvakuutus
	Maatilavakuutus	Venevakuutus
	Maatilavakuutus	Tapaturmavakuutus
	Loma-asuntovakuutus	Oikeusturvavakuutus
	Loma-asuntovakuutus	Vastuuvakuutus
	Loma-asuntovakuutus	Venevakuutus

Relaatiomalli perustuu joukko-oppiin ja matemaattiseen relaatiokäsitteeseen. Relaatiomallille on ominaista rivien ja kenttien väliset operaatiot sekä relaatiot toisiin relaatiomallisiin tietokantoihin. Taulukosta 2.2 voidaan huomata tiedoston, taulukon ja relaation samaa tarkoittavien käsitteiden vertailu.

Taulukko 2.2: Tiedoston, taulukon ja relaation käsitteiden vertailu.

Tiedosto	**Taulukko**	**Relaatio**
Tietue	Rivi	Monikko
Kenttä	Sarake	Attribuutti

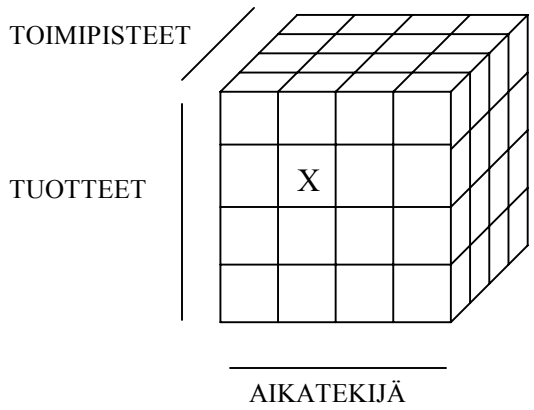
Relaatiotietokantojen suunnittelu perustuu *normalisointiin*, jolloin tieto hajotetaan pienempiin osiin ja niiden toistoja pyritään välttämään. Näissä kannoissa tiedot on jaettu aiheen mukaan eri tauluihin, joita sitten tietoja haettaessa yhdistellään muodostamalla riippuvaisuuksia taulujen välillä. Relaatiotietokannan ideana on jakaa tieto pienempiin, toisistaan riippumattomiin osiin ja luoda tämän jälkeen uutta tietoa näitä osia yhdistelemällä. Näiden kantojen etuna tietovarastoinnissa voidaan pitää avointa SQL-rajapintaa, toimivia välineitä tietojen yhdistelyyn ja jalostukseen sekä summaamiseen. Ongelmina voidaan pitää, että tietoja halutaan monesti tarkastella monista eri näkökulmista, yhteenlaskien ja ryhmiteltynä. Tällaisen tekeminen liitoksineen ja ryhmittelyehtoineen on usein vaikea ja aikaa vievä tehtävä jopa asiantuntijallekin.

2.2.4 Moniulotteinen malli

E.F. Codd kehitti vuonna 1993 yhdessä S.B. Coddin ja C.T. Salleyn kanssa *OLAP-menetelmän (On-Line Analytical Processing)* työkaluksi moniulotteisen tietorakenteiden analysointiin (Codd & al. 1993). Hän kritisoi SQL:ää ja nykyisiä relaatiokantatuotteita niiden huonosta tuesta joustavalle analyysille ja esitti tilalle erikoisia analyyttisiä tietokantatuotteita. Tietojen analysoinnissa tarkastellaan dataa usein moniulotteisesti (kuva 2.7) ja esimerkiksi vakuutusten myyntiä voidaan analysoida ajallisesti, vakuutuslajeittain ja organisaatio-yksiköittäin. Tällöin voidaan saada heti tieto siitä, paljonko oli autovakuutuksen myynti toukokuussa 1-alueella. Tämä voidaan toki hoitaa myös relaatiokannoillakin, ja sitä tarkoitusta varten on viime aikoina kehitetty ns. *moniulotteisia MDD-kantoja (Multidimensional Databases)*, jotka perustuvat OLAP-ajatteluun (Hovi 1997). Tämä moniulotteinen tietorakenne voi olla kaksi- tai useampiulotteinen. Kuitenkaan näitä kantoja ei

voida käyttää operatiiviseen tietojenkäsittelyyn (Hovi 1997). Tästä on kehitetty relaatiotietokannan ja OLAP:n yhdessä muodostama *ROLAP* (*Relational OLAP*), jossa tieto on tallennettuna relaatiotietokannan tauluna.

Moniulotteisten tietojen käsittelyssä tarkastellaan operatiivisista järjestelmistä tuotettua tietoa, joka on jo yhdistettyä ja summattua. Tämän seurauksena tietovarastointi ja OLAP liittyvät saumattomasti toisiinsa. Tietovarastointi sisältää tietojen siirtämisen, muokkaamisen ja varastoinnin operatiivisista järjestelmistä, jolloin OLAP-analyysityökalujen tehtäväksi jää tietojen poimiminen tietovarastosta. OLAP yhdistetään usein ainoastaan liiketoiminnan tietojen analysointiin, jolloin tavalliset analysointitehtävät liittyvät myös markkinointiin ja myyntiin sekä asiakas- ja taloustietojen käsittelyyn.



Kuva 2.7: Moniulotteinen kuutiomalli, jossa myynti voidaan esittää ajan (kuukausi, vuosi), tuotteiden ja toimipisteiden mukaisesti.

3 Asiakkuudenhallintajärjestelmät

Markkinaympäristön ja kilpailutilanteen muuttumisen seurauksena yritykset ovat olleet pakotettuja siirtymään vanhanaikaisesta tuotantokeskeisestä toiminnasta yhä enemmän asiakaskeskeisemmän toimintamallin suuntaan (Luomala & al. 2001). Tällöin asiakassuhteet ja kokonaisvaltainen asiakkuudenhallinta ovat kehittyneet yritysten organisaatioiden tärkeimmiksi painopisteiksi.

Asiakkuudenhallintajärjestelmän eli *CRM:n* (*Customer Relationship Management)* avulla yrityksillä on mahdollisuus seurata kanta-asiakkaidensa asiakassuhteiden hoitamista markkinoinnista ja myynnistä alkaen aina käyttäjätukeen sekä loppupalautteiden analysointiin (Luomala & al. 2001). Tällöin voidaan todeta CRM-termin tarkoittavan yleisesti asiakaskeskeistä liiketoiminnan ajattelutapaa, jolla yritys tavoittelee itselleen ensisijaisesti entistä parempaa taloudellista kannattavuutta sekä asiakastyytyväisyyttä. Asiakkuuden-hallintajärjestelmä sisältää yleensä asiakkaiden ryhmittelyn (*segmentointi*) kuten myös niiden kohdistamisen (*profilointi*) (Bounsaythip & Rinta-Runsala 2001). Asiakkaiden segmentoinnissa asiakkaat jaetaan ja ryhmitellään toisistaan erottuviin pienempiin keskenään samanlaisiin osiin yleisten ominaisuuksien mukaan, joita voivat olla mm. taloudelliset tekijät, alueelliset tekijät ja käyttäytymiserot. Profiloinnissa asiakkaat kuvataan heidän yksittäisten ominaisuuksiensa perusteella, kuten elämäntyyli, ikä, talous, kulttuuri ja maantieteellinen sijainti.

3.1 Asiakkuuksienhallinta

Monilla toimialoilla kilpailu on vain entisestään kiristynyt tarjonnan lisääntyessä, jolloin tuotteiden sekä palvelujen nopea liikkuminen maiden sisällä ja maanrajojen ylikin on tullut mahdolliseksi. Tämän seurauksesta yritysten toiminta ei välttämättä sitoudu enää aina maa- tai aluekohtaiseksi (esim. Amazon-kirjakauppa ja Ebay-huutokauppa Internetissä). Asiakas voi hankkia samoja palveluita ja tuotteita useiden erilaisten jakelukanavien kautta, eikä siksi yhden ostokerran perusteella ole vielä mahdollisuutta pitää asiakassuhdetta tulevaisuudessa itsestään selvänä. Tämän johdosta yrityksen pitää pystyä tarjoamaan asiakkaalleen hänen tarvitsemiaan tuotteita elinkaaren eri tilanteissa.

Asiakkuudenhallinnan ensimmäisenä tavoitteena on yhdistää kaksi erillistä osa-aluetta: markkinoinnin ja myynnin sekä palvelujen järjestelmät. Näiden yhdistämisen perusteluna on varmasti monia erilaisia syitä, mutta yksi merkittävä peruste on se, että yritykset tulevat fuusioiden ja laajentumisten seurauksena yhä suuremmiksi ja asiakkaiden nykyiset tarpeet sekä niiden muuttuminen eivät enää ole niin helposti huomattavissa. Ongelmaa on myös lisännyt tarjolla olevan tietomäärän kasvaminen, jolloin sen oikean ja tarvittavan tiedon poimiminen tietovirrasta on työlästä. Tällöin yrityksen vaarana on etääntyä asiakkaista, jolloin muut asiat tulevat tärkeimmäksi eli yrityksellä ei ole silloin enää selvää näkemystä asiakkaiden nykyisistä ja tulevista tarpeista.

Jokaiseen talouteen osoitettu laaja massamarkkinointi tavoittaa kylläkin monia mahdollisia asiakkaita, mutta se on kallista sekä huonosti kannattavaa ja perusteltua ainoastaan silloin, kun yritys haluaa lisätä tunnettavuuttaan tai myydä päivittäistavaroita. Joskus yrityksille muodostuu kilpailuetua käyttämällä juuri vain niille mahdollisille ostajille kohdennettua markkinointia. Massamarkkinoinnista ei ole hetkessä päästy tähän yksilöityyn markkinointiin (kuva 3.1), vaan juuri siihen tarvitaan asiakkaiden yksilöityjä tietoja ostokäyttäytymisestä ja markkinoista sekä myös heidän omasta taloudestaan.

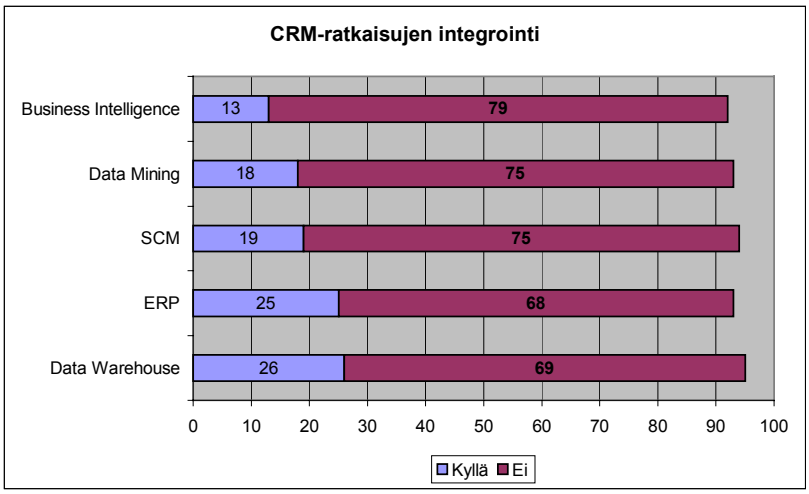
Kuva 3.1: Markkinointimenetelmien tavoitettavuus (PWHC 1999).

Yritykset pyrkivät asiakkuudenhallintajärjestelmien avulla tuntemaan asiakkaansa entistä paremmin, jolloin niiden on helpompi etsiä asiakaskunnastaan kannattavat ja säilyttämisen arvoiset asiakkaat. Näiden tietojen avulla pyritään kohdistetulla markkinoinnilla kasvattamaan muista asiakkaista näiden hyvien asiakkaiden kaltaisia asiakkaita. Toimiva järjestelmä paljastaa myös helposti ne kannattamattomat asiakkaat, joista yritykset eivät välttämättä ole kiinnostuneita. Taulukosta 3.1 havaitaan markkinoinnin kehitys aina massamarkkinoinnista asiakkuudenhallintaan sekä niiden väliset eroavaisuudet.

Taulukko 3.1: Markkinoinnin kehitys (Dyché 2002).

	Massamarkkinointi =>	**Segmenttimarkkinointi** =>	**Asiakkuudenhallinta** =>
Keskeisyys	- tuotekeskeinen	- ryhmäkeskeinen	- asiakaskeskeinen
Kohderyhmä	- kaikki	- ryhmä	- yksilö
Kampanjat	- harvoja	- monia	- paljon
Analysointi	- vähän tai ei yhtään	- ryhmäanalyysi	- asiakkaan käyttäytyminen ja profilointianalyysi
Kesto	- lyhytaikainen	- lyhytaikainen	- pitkäaikainen

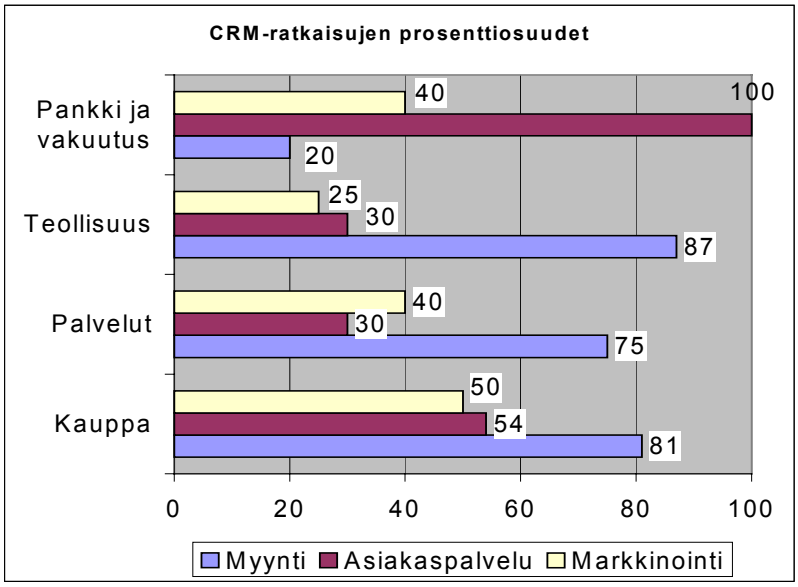
Asiakkuudenhallintajärjestelmät ovat suuria kokonaisuuksia ja tämän johdosta yritykset hankkivat yleensä aluksi vain järjestelmien osaratkaisuja esim. myynninohjaukseen (Toivanen 2000). CRM-ratkaisujen integrointi yritysten muihin järjestelmiin on ollut vielä varsin vähäistä. Vuonna 2001 tehdyssä tutkimuksessa oli mukana 130 yritystä (kuva 3.2), joista lähes 25 % oli integroinut CRM-ohjelmistonsa johonkin tietovarastoratkaisuun ja vain 13 % oli yhdistänyt sen Business Intelligence -työkaluun (Erkkilä 2001).



Kuva 3.2: CRM-ratkaisujen integrointi (N=130) muihin yrityksen sovelluksiin (Erkkilä 2001).

Business Intelligence = tietovaraston analysointia yrityksen tarpeita varten.
SCM (Suply Chain Management) = toimitus- ja kysyntäketjun hallinta.
ERP (Enterprise Resource Planning) = yrityksen integroitu informaatiojärjestelmä.

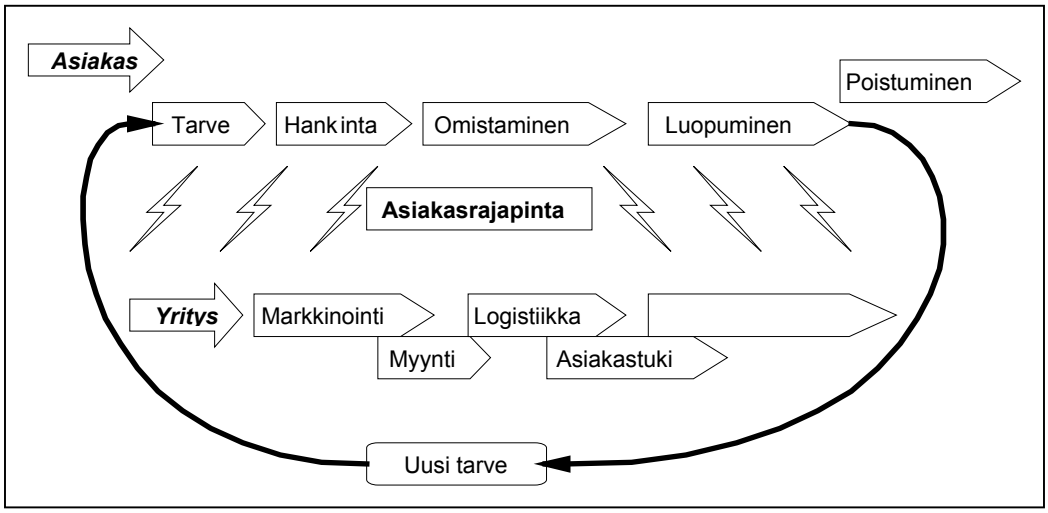
Samassa tutkimuksessa tarkasteltiin myös toteutettuja CRM-ratkaisuja myynnin, asiakaspalvelun ja markkinoinnin osalta. Kuvasta 3.3 voidaan huomata, että pankki- ja vakuutusalalla myynnissä käytetään CRM-pohjaisia ratkaisuja vain 20 %, kun se on muilla aloilla jopa 80 %. Vastapainoksi asiakassuhteen jatkuvuudesta ja pitkäaikaisuudesta johtuen pankki- sekä vakuutusalalla asiakaspalveluratkaisuissa CRM-ratkaisujen osuus on täydet 100 %, kun taas muilla toimialoilla sen käyttö on huomattavasti vähäisempää.



Kuva 3.3: CRM-ratkaisujen toteutus myynnin, asiakaspalvelun ja markkinoinnin ohjelmistoissa (Erkkilä 2001).

Asiakkaalle muodostuu elämänsä aikana paljon pieniä ja muutamia isoja tärkeitä tilanteita (täysi-ikäisyys, avioliitto, lapsen syntyminen tai asunnonosto), joissa hän tekee merkittäviä päätöksiä. Näissä tilanteissa myös yrityksen asiakassuhde on aina vaarassa. Asiakkaan elinkaari alkaa tarpeen ilmenemisellä ja loppuu aina uuteen tarpeeseen tai vanhasta luopumiseen (Luomala & al. 2001). Asiakkaalle muodostuu siis eri prosessin vaiheessa erilaisia tarpeita ja odotuksia, joihin yrityksen on pystyttävä vastaamaan omalla tarjonnallaan. Yrityksillä on käytössään erilaisia toimintoja (kuva 3.4), mutta näiden välinen vuorovaikutus usein puuttuu tai on puutteellinen. Tämä johtuu siitä, että asiakkuuden elinkaaren eri vaiheissa asiakas tekee ratkaisuja useiden eri henkilöiden kanssa ja nämä henkilöt eivät välttämättä tiedä mitä aikaisemmat henkilöt ovat asiakkaan kanssa keskustelleet (Luomala & al. 2001). Asiakkaita koskevat tiedot on tallennettu yritysten järjestelmiin, mutta usein juuri näiden tietojen yhdistäminen muodostuu ongelmaksi.

Yksi CRM-ajattelun perusta on näiden tietoaukkojen poistaminen, jolloin asiakkaan saama kokonaisvaltainen palvelu paranee ja yritys saa lisää katetta nykyisestä asiakassuhteesta. Tässä vaiheessa pyritään luomaan *proaktiivinen asiakassuhde*, jolloin asiakkaan ongelma-tarpeet pystytään ennustamaan jo etukäteen (Luomala & al. 2001). Vastaavasti *reaktiivisessa asiakassuhteessa* havaitaan asiakkaan kiinnostuksen alkaminen vasta silloin, kun hän itse tunnistaa tarpeensa ja kertoo siitä yritykselle (Luomala & al. 2001).



Kuva 3.4: Asiakkuuden elinkaari (Luomala & al. 2001).

3.2 Markkina-analyysi

Jokaisella kaupparyhmällä on nykyisin käytössään jonkinlainen kanta-asiakas- tai bonuskortti, joita on asiakkaille jaettu vuosien aikana miljoonia kappaleita. Puolta näistä korteista käytetään aktiivisesti, ja niiden käytöstä kauppaketjut maksoivat kanta-asiakkailleen vuonna 2000 erilaisina bonuksina yli 170 miljoona euroa (Kärkkäinen 2001). Kuitenkaan nämä kortit eivät ole pelkästään tarkoitettu asiakkaiden iloksi ja hyödyksi, vaikka asiakkaita houkutellaan käyttämään niitä erilaisilla raha- tai tavarapalkinnoilla. Asiakkaina emme aina ensimmäiseksi ajattele sitä, että kaupan kassakuitti sisältää suuren määrän tietoa, joka on kauppiaalle tai kauppaketjulle erittäin arvokasta. Muutaman prosentin ostohyvitystä vastaan asiakkaina luovutamme tärkeitä tietoja itsestämme: tiedot omasta kulutuksesta ja tottumuksista. Myöhemmin kuittitiedoista saadaan selville jopa tuotteiden tarkkuudella, milloin ja mitä olemme ostaneet, millaisia ne ostokset olivat ja mitä ne maksoivat tai jopa kuinka kaupan erilaiset mainoskampanjat toimivat.

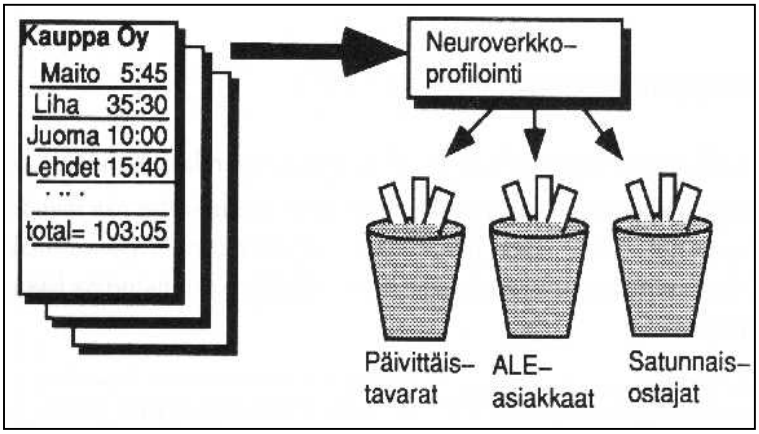
Tulevaisuudessa kauppaliikkeet ovat myös varmasti kiinnostuneita tiettyjen tuotteiden kohdistamisesta suoraan ennalta määrätyille asiakasryhmälle esim. kirjatarjoukset niitä tarvitseville. Yritykset hyötyvät nyt suuresta profiloidusta asiakastietorekisteristään ja tarjonnan kohdistaminen asiakkaisiin on tällöin mahdollista. Tämä edellyttää, että kauppiaat tietävät todellisuudessa, millaisia ryhmiä asiakkaat edustavat. Tällöin kaupan asiakassuhdehallinta voi olla taitavasti kohdennettua markkinointia, mutta pelkästään asiakkaan nimen lisääminen mainoskirjeeseen ei enää riitä. Asiakkaan perustietojen ja

ostoskäyttäytymisen perusteella yritykset eivät voi vielä vetää kovin pitkälle meneviä tarkkoja johtopäätöksiä, koska emme voi varmasti sanoa, ostaako asiakas tavaran itselleen vai lahjaksi.

3.2.1 Toimintatapa

Asiakasprofiloinnin tarkoitus on jakaa kuittien sisältämät ostoskerrat tyypillisiin ryhmiin, joiden asiakkaiden ostoskäyttäytyminen on samankaltaista. Analyysin yksi toiminta-menetelmä on *itseorganisoituva kartta* eli *SOM (Self-Organizing Map*) (Koikkalainen 1994). SOM-menetelmässä algoritmi sijoittaa neuronit (karttayksikkö) kaksiulotteiselle tasolle siten, että sisällöllisesti lähellä toisiaan olevat kuitit sijoittuvat vierekkäisille neuroneille ja kaukana toisistaan olevat ovat tasollakin kauimpana (Kohonen 1997).

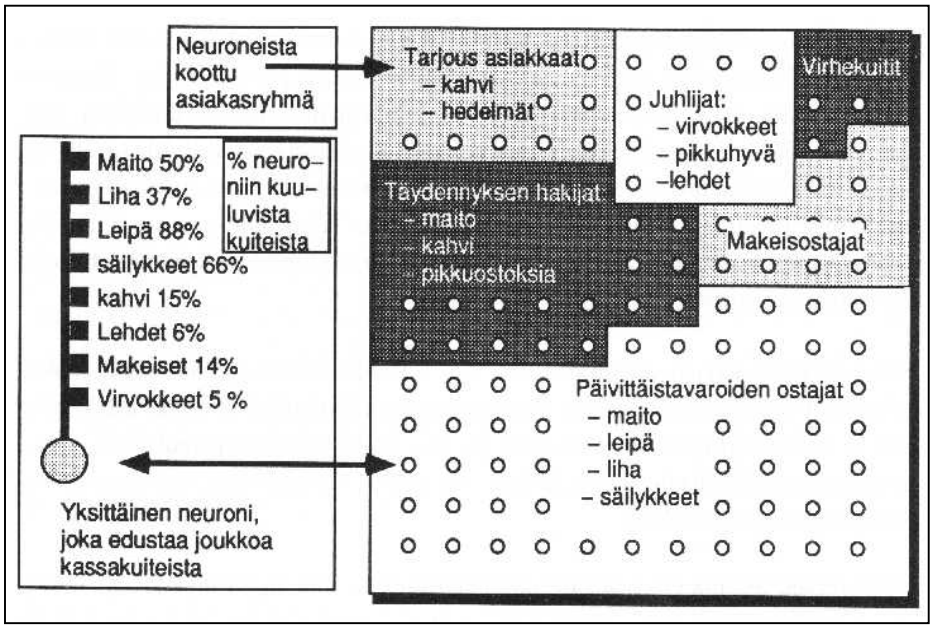
Asiakkaan ostoskoriin keräämien tavaroiden viivakoodit rekisteröityvät kaupan kassakoneessa tiedostoksi. Myöhemmin analysointiohjelma käy läpi näitä miljoonien kuittien tietoja tuhansia ja taas tuhansia kertoja oppien suoritusprosessien aikana. Lopuksi ohjelma ryhmittelee kuitit ostokoreiksi (kuva 3.5), jolloin niistä saadaan esille asiakkaiden tarkemmat ostotottumukset ja -käyttäytymiset. Ryhmittelyn tarkoituksena on jakaa kuitit siten, että jokainen kori edustaa jotakin tyypillistä käyttäytymismallia. Luokittelumalli on lopulta valmis otettavaksi käyttöön, kun verkko on käynyt opetusaineiston läpi riittävän useasti. Tällöin aineistosta on löytynyt kaikki löydettävissä olevat asiakasryhmät ja niiden painotukset sekä ryhmien väliset tärkeimmät muuttujat (Koikkalainen 1994).



Kuva 3.5: Kassakuittien ryhmittely (Koikkalainen 1994).

Tarkemmin kuvattuna algoritmin toiminnassa tapahtuu aluksi itseorganisoituvan kartan alustus, jolloin karttavektoreille annetaan alkuarvot satunnaisesti syötevektoreista tai valitaan ne satunnaisesti arvoalueelta. Tämän jälkeen yhtä kuittia verrataan aina yksitellen kaikkiin karttavektoreihin, jolloin jokaisessa neuronissa lasketaan syötevektorin (laskun) ja painovektorin samankaltaisuus. Samalla tavoin käydään läpi kaikki neuronit ja valitaan voittajaneuroni, jonka arvo on samankaltaisin. Tämän jälkeen korjataan voittajaneuronin ja kartalla sen ympärillä olevien neuronien karttavektoria opetuskertoimella. Alussa opetuskertoimet ovat suuria, jolloin kartta ”elää” hyvin voimakkaasti. Myöhemmin opetuskertoimia pienennetään opetuksen edistyessä, jolloin loppuvaiheessa kartan muuttuminen on hyvin pientä, lähinnä hienosäätöä. Korjauksen jälkeen valitaan seuraava syötevektori ja toistetaan opetus. Ajan kuluessa juuri tämän naapuristo-opetuksen ansiosta voittajaneuronin ympärille muodostuu samankaltaisten neuronien joukko.

Kaksiulotteisella kartalla voidaan havaita samanlaisten neuronien muodostavia ryhmiä, joiden ostoskäyttäytyminen on helposti tutkittavissa (kuva 3.6). Ryhmistä voidaan todeta esimerkiksi ostoskorin keskihinta ja moniko asiakkaista on ostanut tuotteita (Koikkalainen 1994). Kuitenkin jokaisen ryhmän sisältä on nähtävissä myös pientä hajontaa, sillä eiväthän kaikki ostajat osta juuri samoja tavaroita. Jokaisesta yksittäisestä neuronista pystytään näkemään, kuinka monta prosenttia sen asiakkaista on mitäkin tuotetta ostanut.



Kuva 3.6: Ryhmittelyn antama lopputulos (Koikkalainen 1994).

Ostoskorianalyysin perusteella kauppaliikkeet pystyvät muuttamaan tuotevalikoimaansa siihen suuntaan, joihin heidän asiakkaansa haluavat painottaa ostoskäyttäytymistään (Toivanen 2001). Tällöin erilaisia kampanjoita pystytään kohdistamaan suoraan niitä käyttäville asiakkaille. Näiden lisäksi kaupat käyttävät myös ryhmittelyjen lopputuloksia apuna suunnitellessaan kaupan hyllyjärjestystä ja tuotteiden sijoittamista. Tästä hyvänä esimerkkinä ovat markettien maitohyllyt, jotka sijaitsevat melkein aina sisääntulon vastakkaisessa nurkassa. Tällöin asiakas kävelee koko kaupan läpi tehden ehkä muitakin heräteostoksia kuin vain hakeakseen sen tarjouksessa olevan maidon.

3.2.2 Mahdollisuudet ja uhat

Kauppaketjut suorittavat nykyisin jatkuvaa ostoskorianalyysia, mutta menetelmäähän voidaan kuitenkin soveltaa melkein samanlaisena kauppojen ulkopuolellekin (Antikainen 1999). Mikäli yrityksellä on suuri määrä erilaisia tuotteita ja asiakkaita, voi toimiala olla pankki- tai vakuutustoiminta, postimyynti tai tukkukauppa, joille todellisten asiakasryhmien hahmottaminen on muuten vaikeata ilman neuroverkkoa. Kauppaketjujen lisäksi tiedämme suuria profiloituja asiakastietorekistereitä olevan varmasti myös vakuutusyhtiöillä, puhelinyhtiöillä, pankeilla ja lehtitaloilla. Näille kaikille on yhteistä yrityksen toiminnan perustuminen sopimussuhteiseen liiketoimintaan, jolloin asiakastietoja on automaattisesti käytettävissä.

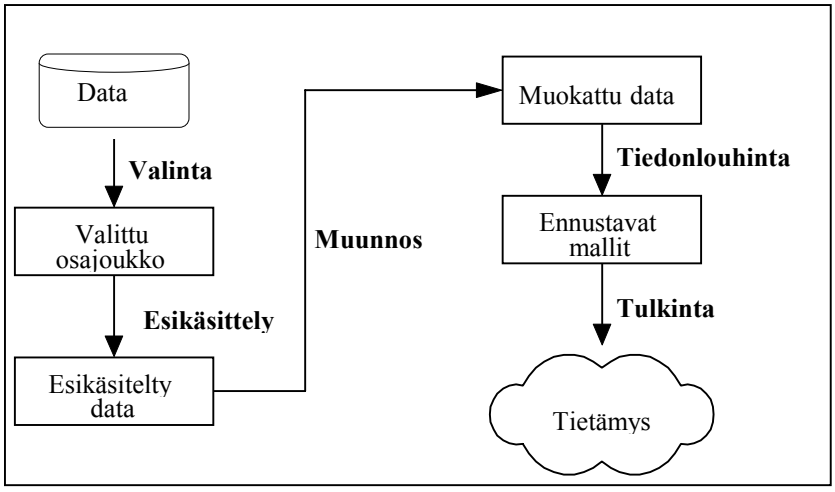
Kauppaketjuilla on suuret tietovarastot, joissa tietoa ei luokitella heidän kertomansa mukaan jokaisesta asiakkaastaan yksilöllisesti eli tietokannasta ei voi nähdä tuotekohtaisesti, mitä kukin asiakas on ostanut (Erkkilä 2001). Kanta-asiakastiedon hyödyntämisessä olennaista on erilaisten asiakastyyppien seulominen ja siksi erilaisten tietojen avulla pyritään luokittelemaan asiakkuuksia eikä yksittäisiä asiakkaita. Kuitenkin koneiden laskentatehon kasvaessa ja tietovarastotekniikan kehittyessä kauppaliikkeet voivat jopa ruveta yksilöimään tietoa tarkemmin. Tällöin siitä voi aiheutua ongelmia tiedon joutuessa vieraisiin käsiin, tai bonuskortin yhteistyöyritykset käyttävät niitä tietoja omia tarpeitaan varten. Tällöin voidaan miettiä vakuutustilannetta, jossa vakuutusyhtiö pystyy luokittelemaan tietokannastaan yksilöitynä ne ihmiset, jotka elämäntavoiltaan (ostavat mm. rasvaisia tuotteita) ovat riskihenkilöitä. Tällöin kynnys myöntää sairausvakuutus tällaiselle asiakkaille voi nousta,

koska vakuutusyhtiön mukaan tällaisella asiakkaalla on suuri sairastumisen riski, tai asiakkaan tietoja käytetään perusteena evätä korvattava vahinko (ostoskori sisältää terveydelle riskialttiita ruokia). Tämä johtaa siihen, että asiakas on itse jo omalla kanta-asiakaskortin käytöllä antanut perusteet korvauksen tai vakuutuksen myöntämisen hylkäämiselle.

4 Tiedonlouhinta

Tiedonlouhinta (*Data Mining)* on uusimpia menetelmiä tietovarastojen analysointiin. Tätä termiä on käytetty kirjallisuudessa yleisesti viittaamaan periaatteisiin ja tekniikoihin, joita käytetään suurten tietomäärien tutkimiseen. Tämän menetelmän avulla on mahdollistaa löytää raakatiedosta tai tietovarastosta sinne kätkeytyvää yllättävää ja ei-itsestäänselvää informaatiota kuten ryhmiä ja yhteyksiä eri asioiden välillä.

Tiedonlouhintaa pidetään yhtenä osa-alueena *tietämyksen muodostamista tietokannoista* (*Knowledge Discovery in Databases, KDD*). Tietämyksen muodostamisessa rakennetaan tiedoista uusia käyttökelpoisia ja ymmärrettäviä malleja, joiden rakentamisessa käytetään apuna vain tiettyä osaa koko tietovarastosta. Menetelmän aineistot jaetaan kolmeen osaan: *opetus*-, *validointi*- ja *testiaineisto* (Parr Rud 2001). Opetusaineistolla opetetaan aineistoa tunnistamaan esim. riskialttiit asiakkaat ja validointiaineistolla päätetään mallin opetuksen lopettaminen, silloin kun oppimista ei enää tapahdu. Lopuksi testiaineistolla varmistetaan, että mallin toiminta on halutunlaista. Tämä toimintamenetelmä on sekä interaktiivinen että iteratiivinen (Roiger & Geatz 2003). Tiedonlouhinnan toteuttaminen etenee selkeinä kokonaisuuksia kuvan 4.1 mukaan, josta voidaan havaita prosessin muita työvaiheita olevan tiedon valinta, esikäsittely ja muunnos sekä tulosten tulkinta.

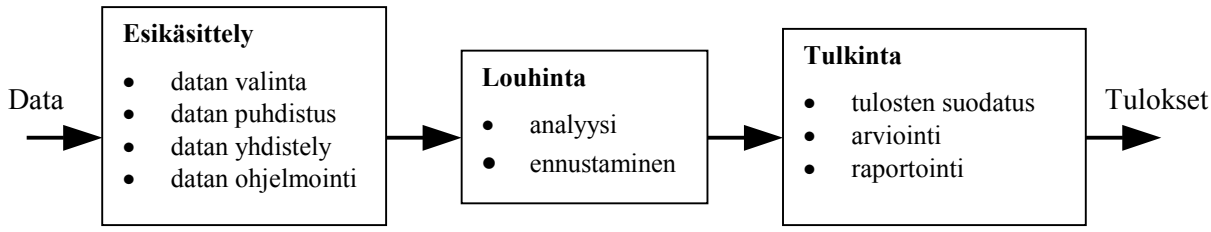


Kuva 4.1: Tietämyksen muodostamisprosessi.

Tiedonlouhintatekniikat perustuvat algoritmeihin, jotka yhdistelevät tekniikoita tietokantojen käsittelystä, tilastotieteestä ja koneoppimisesta. Kehiteltyjä metodeja voidaan käyttää millä tahansa sovellusalalla, kuten kaupan, televiestinnän ja lääketieteen alalla. Hyvinä esimerkkeinä ovat käyttäjien mukaan personoidut verkkosivustot ja suoramarkkinointi-asiakkaiden tunnistaminen ja niihin liittyvät kampanjat. Tietovarastojen analysoiminen tavanomaisilla tilastomenetelmillä olisi muuten hidasta, epäkäytännöllistä ja jopa mahdotonta toteuttaa niiden sisältämän suuren tietomäärän johdosta. Tiedonlouhinnan tekniikoiden paremmuus näkyy siinä, että ne louhivat ja järjestelevät valtavia tietomääriä ryhmiksi yrittäen löytää niistä samankaltaisuuksia käyttäjille vastaukseksi. Tämän johdosta analysoinnin perusteella saadut ratkaisut saattavat poiketa aikaisemmista matemaattisten funktioitten (tilastolaskenta) avulla saaduista vastauksista, joiden tueksi ja selvyydeksi aineistosta yritetään löytää uusia yhtäläisyyksiä.

4.1 Tietojen esikäsittely

Yrityksen tuotantokäytössä oleviin järjestelmiin tiedot tulevat monista eri tiedostoista, joiden tietorakenteet ja rajapinnat ovat hyvinkin erilaisia. Nämä tietokannat ovat kooltaan suuria, jolloin niiden sisältöön jää huomaamatta puutteellisia sekä yhteensopimattomia tietoja. Nämä ongelmat voivat olla tietojen puuttumista, arvoalueen ulkopuolisia arvoja tai ristiriitaisuuksia. *Tiedon esikäsittelyllä* (*data preprocessing*) voidaan parantaa tiedon laatua ja lopulta louhinnan lopputulosta (kuva 4.2). Tiedon esikäsittelyyn on olemassa useita erilaisia tekniikoita: *tiedon puhdistus* (*data cleaning*), *tiedon yhdistäminen* (*data integration*), *tiedon muutos* (*data transformation*) ja *tiedon vähentäminen* (*data reduction*) (Han & Kamber 2001).



Kuva 4.2: Tiedonlouhinnan vaiheet (Karanta 2002).

4.1.1 Tietojen puhdistaminen

Tietoaineisto ei aina ole ”puhdas”, koska se voi sisältää vanhentunutta tietoa, puuttuvia arvoja tai kirjoitusvirheitä. Nämä virheet pitää korjata ymmärrettävään muotoon ennen tiedon jatkokäsittelyä. Tietokannan puhdistus voidaan suorittaa useammalla eri menetelmällä, joista puuttuvien arvojen korjaamisessa on helpointa vain poistaa tai sivuuttaa puutteelliset tietueet. Tämä on mahdollista vain silloin kun nämä edustavat vain pientä määrää koko aineistosta (Roiger & Geatz 2003). Oikeiden tietojen etsiminen ja lisääminen käsin puuttuvien arvojen paikalle on aikaa vievää sekä suuremmilla tietomäärillä jopa suorastaan mahdotonta toteuttaa. Helpompi tapa on korvata puuttuva tieto erikoismerkillä (esim. ääretön), joka ei esiinny tietokannan tietueissa (Han & Kamber 2001). Tällöin nämä tietueet voidaan myöhemmin helposti yhdistää puuttuviksi arvoiksi tai jättää ne jopa huomioimatta. Näiden lisäksi yhtenä vaihtoehtona voidaan puuttuvan tiedon paikalle antaa arvoksi muiden sen luokan tietueiden keskiarvo (Han & Kamber 2001). Vastaavasti silloin kun halutaan puuttuva arvo korvata mahdollisimman todennäköisellä arvolla käytetään menetelmänä päättelyyn perustuvia työkaluja, kuten Bayesin-menetelmää tai päätöspuita (Han & Kamber 2001). Päättelyn apuna näissä menetelmissä käytetään toisten ominaisuuksien arvoja ennustamaan juuri näitä puuttuvia arvoja.

Satunnaisten virheiden ja vaihteluvirheiden korjaaminen voidaan tehdä käyttäen apuna erilaisia tiedon pehmennystekniikoita. *Sitomisessa* (*binning*) pehmennetään numeeristen arvojen erilaisuutta muuttaen niiden arvoja naapuruston kesken. Tällöin esijärjestetyt muuttujien arvot jaetaan useisiin eri ryhmiin, joissa niiden arvoa muutetaan ryhmän sisällä keskiarvon, mediaanin tai ääriarvojen mukaisesti (Han & Kamber 2001). Vastaavasti ryhmittelymenetelmässä samanlaiset arvot sijoittuvat omiin ryhmiin, jolloin on helpompi tunnistaa poikkeavat ryhmän ulkopuoliset arvot. Lisäksi tietoaineisto voidaan käydä läpi tietokonetta apuna käyttäen tunnistaen ja lajitellen ulkopuoliset arvot erilliseen tiedostoon, josta analysoija voi käydä ne käsin läpi ja tarvittaessa korjata sekä poistaa tarpeettomat. Tämä menetelmän toteuttaminen on huomattavasti helpompaa kuin vastaavasti koko tietoaineiston läpikäyminen (Han & Kamber 2001). Yleisesti tiedetään tietojen puhdistamisen ja korjaamisen olevan iso ongelma, josta muodostuu jopa 70–80 % tietovarastoinnin lopullisista käyttökustannuksista (Newell 2000).

4.1.2 Tietojen yhdistäminen

Tiedonlouhintaa varten joudutaan useimmiten rakentamaan yhtenäinen tietokanta yhdistäen tiedot useista erilaisista kannoista kuten tietovarastotekniikassa on aikaisemmin mainittu. Tietojen yhdistämisessä on tärkeätä tunnistaa ja ratkaista tietojen ristiriitaisuudet. Tällöin ongelmiksi muodostuvat eri tietokannoissa eri tavalla nimetyt samaa tietoa kuvaavat tietueet, jolloin ohjelman on mahdotonta päätellä niiden edustavan samaa tietoa vaikkakin ne olisivat johdonmukaisesti nimettyjä. Esimerkiksi *vaknro* ja *vakuutusnumero* voivat tarkoittaa samaa asiaa, mutta eri tietokannoissa ne on vain nimetty eri tavalla johtuen huolimattomuudesta, eri ohjelmatoimittajasta tai ainoastaan vain toimintatavasta. Tiedonsisällön osalta myös eri mittajärjestelmien käyttäminen vaikeuttaa yhdistämistä, esim. jos vakuutusmaksut ovat eri rahayksikössä, jolloin maksu voi sisältää erilaisia maakohtaisia veroja ja valtion perittämäksi määrättyjä piiloveroja (palosuojelumaksu, työttömyysvakuutusmaksu ja sairasvakuutus-maksu), jotka joudutaan poistamaan maksusta tapauskohtaisesti.

Saman tiedon toistaminen useammassa paikassa on turhaa silloin, kun tieto on jo valmiina saatavilla yhdestä paikasta. Jotkut näistä tiedon toistamisen aiheuttamista tarpeettomuuksista havaitaan käyttäen *korrelaatioanalyysia* (*correlation anlysis*) kaavan 1 mukaisesti (Han & Kamber 2001). Tämän menetelmän toimintatapana on mitata kahden lineaarisesti riippuvan muuttujan arvon yhtäläisyyttä toisiinsa, jolloin analyysin vastauksena antamat arvot vaihtelevat +1:n ja –1:n välillä.

$$r_{A,B} = \frac{\sum (A - \overline{A})(B - \overline{B})}{(n-1)\delta_A \delta_B} \qquad (1)$$

Kaavassa *n* edustaa ominaisuuksien lukumäärää ja $\overline{A}$ sekä $\overline{B}$ ovat keskiarvoja *A*:sta ja *B*:stä kaavan 2 mukaan. Kun taas δ_A edustaa *A*-arvon keskihajontaa ja δ_B vastaavasti B:stä kaavan 3 mukaan.

$$\overline{A} = \frac{\sum A}{n} \qquad \overline{B} = \frac{\sum B}{n} \qquad (2)$$

$$\delta_A = \sqrt{\frac{\sum (A - \overline{A})^2}{n-1}} \qquad (3)$$

Tämän analyysin lopputuloksena voidaan tehdä johtopäätöksiä muuttujan $r_{A,B}$ arvosta (Han & Kamber 2001). Lopputuloksen ollessa suurempi kuin nolla korreloivat *A* ja *B* positiivisesti tarkoittaen, että mikäli *A*:n arvo kasvaa niin todennäköisesti kasvaa myös *B*:n arvo. Mitä suuremmaksi arvo $r_{A,B}$ muodostuu, sitä varmemmin se ennustaa samankaltaisuutta, jolloin toinen arvoista (*A* tai *B)* voidaan poistaa tarpeettomana. Arvon ollessa tasan tai lähellä nollaa ovat *A* ja *B* itsenäisiä, eikä näiden arvojen välillä ole mitään riippuvuutta. Negatiivisessa korrelaatiossa arvo on alle nollan, jolloin toisen arvon noustessa toinen laskee. Tällöin arvon vastakohta korreloi toisen arvon kanssa, jolloin mitä negatiivisemmaksi arvo muodostuu, sitä enemmän arvon vastakohdalla on yhtäläisyyttä tutkittavan kohteen kanssa.

4.1.3 Tietojen muuntaminen

Tiedon muuntamisella tarkoitetaan tiedon korjaamista tai yhdistämistä yhtenäiseksi kokonaisuudeksi louhintaa varten. Ennen louhintaa tietoaineisto tulee käydä läpi ja muuntaa se osittain helpommin ymmärrettävään muotoon, jotta louhinnan algoritmit toimisivat riittävän tehokkaasti (Berry & Linoff 2000). Tämä on suuri haaste tiedonlouhinnan toteutuksessa ennen kaikkea sen suuren ajankäytön ja hyvän lopputuloksen löytymisen osalta. Menettelytapoja on useita: *pehmennys* (*smoothing*), *summaaminen* (*aggregation*), *yleistäminen* (*generalization*) ja normalisointi (*normalization*) (Han & Kamber 2001).

Pehmennyksen toimintaperiaate on esitelty kohdassa 4.1.1. Vastaavasti summaaminen menetelmänä kertoo jo nimellään, että tiedot kerätään yhteen paikkaan jo valmiiksi laskettuna. Tiedon summaaminen on tärkeätä tehdä etukäteen, mikäli emme ole kiinnostuneita osatiedosta vaan ainoastaan niiden loppusummasta (Han & Kamber 2001). Esimerkiksi vakuutuksen hinta voi olla tallennettuna kohteittain, jolloin valmiiksi laskettu vakuutuksen kokonaismaksu puuttuu aineistosta. Tällöin tämä joudutaan joka kerta laskemaan, mutta mikäli se olisi jo valmiina, se jouduttaisi merkittävästi tiedon käsittelyä. Yleistämisellä tarkoitetaan tarkan tiedon korvaamista karkeamman tason käsitteellä. Esimerkiksi vakuutetun asuinrakennuksen vakuutusmäärä voidaan korvata termeillä ali-, täysarvo- tai ylivakuutus.

Tällöin näiden valittujen käsitteiden avulla poistetaan mahdollisuus ristiriitaisen tiedon esittämiseen.

Normalisoinnissa tieto suhteutetaan pienemmälle arvovälille kuten esimerkiksi 0–1. Tällöin arvoista suurin saa maksimiarvon ja pienin minimiarvon, kun taas muut asettuvat näiden väliin säilyttäen kuitenkin alkuperäiset suhteelliset eronsa. Tämä normalisointi on tarpeen siksi, että alkuperäisten arvojen arvoalueet ovat hyvin erilaiset eri ominaisuuksilla kuten asiakkaiden ikä 1–100 vuotta tai korvausten suuruus 0–10 milj. euroa. Seuraavaksi käsittelen kolmea erilaista tietojen arvoasteikon muutosmenetelmää: *min-max, z-piste* ja *desimaaliasteikolla normalisointi* (Roiger & Geatz 2003).

Min-max-normalisointi vastaa suoraviivaista muutosta alkuperäisestä tiedosta ja on käyttökelpoinen silloin kun tunnetaan aineiston minimi- ja maksimiarvot. Nämä muunnokset tehdään jokaiselle arvolle erikseen, jolloin tietoaineiston suurimman arvon max_A ja pienimmän arvon min_A väliin jäävät muut v arvot (Han & Kamber 2001). Arvoalueen rajaksi määritellään yleensä 0–1, jolloin new_min_A saa suoraan arvon nolla ja new_max_A saa arvokseen yksi. Kaavan 4 mukaan lasketut v' arvot asettuvat siten arvoalueelle new_max_A – new_min_A, väliin.

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \qquad (4)$$

Z-piste-normalisointi perustuu arvojen ominaisuuksien normalisointiin keskihajonnasta (kaava 5). Tätä normalisointia käytetään silloin, kun ei varmasti tiedetä todellisen arvoalueen maksimi- ja minimiarvoja tai arvoalueen ulkopuolelle jäävät arvot sekoittavat min-max-normalisoinnin (Han & Kamber 2001). Keskihajonta δ_A voidaan laskea kaavan 3 mukaan ja keskiarvo $\overline{A}$ kaavan 2 mukaan.

$$v' = \frac{v - \overline{A}}{\delta_A} \qquad (5)$$

Desimaaliasteikolla normalisointia voidaan käyttää silloin, kun tiedetään arvoalueen maksimi- ja minimiarvot (Han & Kamber 2001). Tällöin itseisarvoltaan suurin arvo otetaan huomioiduksi, minkä perusteella desimaaliasteikko muodostetaan. Arvojoukon luvut jaetaan 10^j arvolla, jolloin periaatteessa vain siirretään desimaalipistettä vasemmalta oikealle (kaava 6). Pilkun siirtämisen suuruus riippuu maksimin suuruudesta. Tässä kaavassa *j* on pienin kokonaisluku, jolla toteutuu lopputuloksen ehto *Max(|v'|)<1*.

$$v' = \frac{v}{10^j} \tag{6}$$

4.1.4 Tietojen vähentäminen

Tietovarastoissa oleva suuri tietomäärä voi jopa hidastaa tiedonlouhintaprosessia tai tehdä sen toteuttamisen mahdottomaksi kohtuullisessa ajassa. Tämän takia tiedon vähentäminen on tarpeellista, mutta se on kuitenkin tehtävä vaarantamatta itse louhinnan lopputulosta. Tiedon vähentämisessä on käytössä useita eri menetelmiä, kuten *tiedon summaaminen* (*data aggregation*) ja *mittasuhteen pienentäminen* (*dimension reduction*) (Han & Kamber 2001).

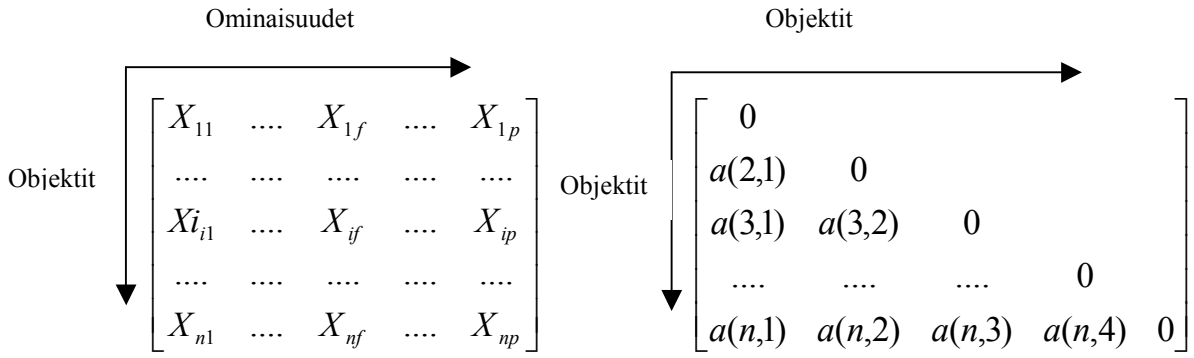
Tiedon summaamisen toimintaperiaate on esitelty kohdassa 4.1.3. Mittasuhteiden pienentäminen tulee tarpeelliseksi tietoaineiston sisältäessä satoja muuttujia, joista monet ovat asiaankuulumattomia tai tarpeettomia tai joilla ei ole mitään tekemistä itse tämän louhinnan kanssa (puhelinnumerot ja pankkitilin numero) (Han & Kamber 2001). Tietojen manuaalinen poistaminen on vaikeaa ja aikaa kuluttavaa erityisesti silloin, kun ei tunneta tiedon sisältöä riittävän tarkasti. Poistamalla olennaiset muuttujat tai pitämällä juuri ne epäolennaiset muuttujat mukana on lopputulos epäedullinen, jolloin tiedonlouhinnan täydellinen sekoittuminen on mahdollista.

4.2 Tietojen esittäminen

Tietotyyppeinä käytetään usein tiedon esittämisessä tieto- tai erilaisuusmatriisia kuvan 4.3 mukaisesti (Han & Kamber 2001). Yleensä tieto sijaitsee alkuvaiheessa kaksiulotteisessa tietomatriisissa, josta se muutetaan erilaisuusmatriisin muotoon ennen ryhmittelyn

toteuttamista. Tietomatriisi kuvaa tiedon olemassaoloa kaksiulotteisena taulukkona, jossa jokaisella vaakarivillä sijaitsee kohde eli *objekti* ja sarakkeissa ovat sen muuttujat eli *attribuutit*. Esimerkiksi autovahingoista kohteina ovat yksittäiset vahingot ja ominaisuuksina ovat niiden vahinkotiedot (esim. kuljettajan ikä, olosuhteet, ajoneuvo, korvausmäärä).

Erilaisuusmatriisissa tieto kuvataan taulukkona, jossa taulukon rivit ja sarakkeet edustavat objekteja. Taulukon arvo *a(i,j)* sisältää objektien *i* ja *j* ominaisuuksien erilaisuuden arvon. Erilaisuudella tarkoitetaan sitä, kuinka eri objektien ominaisuudet eroavat toisistaan ja kuinka tämä erilaisuus pystytään mittaamaan erilaisten laskentakaavojen perusteella. Erilaisuutta voidaan myös kutsua jossakin tapauksessa kahden objektin etäisyydeksi. Käytännössä taulukon arvo *a(i,j)* saa vain positiivisia arvoja. Arvon lähestyessä nollaa ovat *i* ja *j* lähellä toisiaan, mitä suurempi arvo niin sitä enemmän ne ovat eroavat toisistaan.



Kuva 4.3: Kaksiulotteiset matriisit (kaksiulotteinen ja erilaisuusmatriisi).

4.2.1 Etäisyyksien mittaaminen

Havaintojen erilaisuutta mitataan niiden etäisyydellä, joista suosituimpana menetelmänä käytetään *Euclidean etäisyyttä* (Han & Kamber 2001), missä $(x_{i1}, x_{i2,..,} x_{ip})$ ja $(x_{j1}, x_{j2,..,} x_{jp})$ ovat p-ulotteisen tietojoukon objekteja (kaava 7).

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2} \qquad (7)$$

Toinen tunnettu menetelmä etäisyyksien mittaukseen on myös *Manhattan etäisyys* kaavan 8 mukaisesti (Han & Kamber 2001).

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}| \quad (8)$$

Molemmat etäisyydet ovat metrikkoja, koska ne täyttävät aina kaavan 9 mukaiset vaatimukset (Han & Kamber 2001).

- *d(i,j) 0:* *Etäisyys ei voi olla negatiivinen.*
- *d(i,i) = 0:* *Etäisyys itseensä on aina nolla.* *(9)*
- *d(i,j) = d(j,i):* *Etäisyys on symmetrinen.*
- *d(i,j) <= d(i,h) + d(h,j):* *Etäisyys i:stä j:hen on aina lyhyempi suoraan kuin kolmannen pisteen kautta kiertäen.*

Etäisyyksien laskemisen yhteydessä saattaa esiintyä tarvetta painottaa jotakin tiettyä ominaisuutta enemmän kuin toisia ominaisuuksia. Tästä hyvänä esimerkkinä voidaan todeta, että asiakkaan irtisanomasta vakuutuksesta korvatulle tai evätylle vahingolle pitäisi ehkä antaa enemmän painoa kuin muille ominaisuuksille, koska huonosti hoidettu vahinko voi usein ärsyttää asiakkaan tekemään vakuutusyhtiön vaihtamispäätöksen. Molempien edellä mainittujen etäisyyskaavojen yhteydessä voidaan käyttää jokaisen ominaisuuden kohdalla omaa painokerrointa, mikäli sen käyttäminen nähdään tarpeellisena (Han & Kamber 2001). Euclidean etäisyyden painotettu laskentakaava voidaan esittää kaavan 10 mukaisesti.

$$d(i,j) = \sqrt{(w_1 |x_{i1} - x_{j1}|^2 + w_2 |x_{i2} - x_{j2}|^2 + ... + w_p |x_{ip} - x_{jp}|^2)} \quad (10)$$

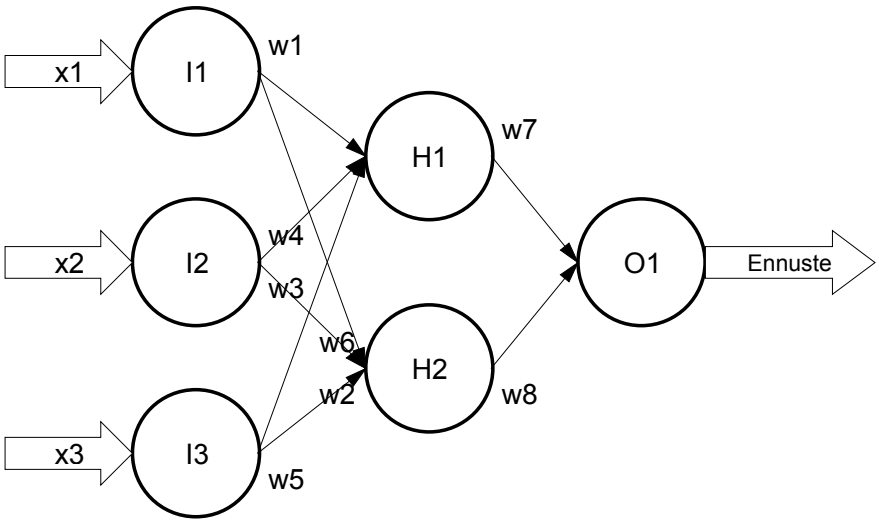
Edellä mainitut menetelmät toimivat ainoastaan tietyllä arvovälillä oleville numeromuuttujille. Tietomatriisi voi myös sisältää binäärisiä muuttujia, joiden etäisyyksiä ei edellä mainitulla laskentakaavoilla pystytä järkevästi laskemaan. Vaihtoehtoisesti nämä voidaan muuttaa numeeriseen muotoon tai käyttää laskennassa *Jaccard'n kerrointa*, jolloin etäisyyksien laskeminen perustuu kahden esiintyvän olotilan lukumäärien esiintymisen laskemiseen (Han & Kamber 2001).

4.3 Neuroverkko

Tiedonlouhinnan perustekniikat perustuvat tällä hetkellä tekoälyn puolella kehitettyihin algoritmeihin, joiden avulla paljastetaan muuten vaikeasti tulkittavia ja löydettäviä tietomassoissa olevia riippuvuuksia. *Neuroverkot* (*neural networks*) jäljittelevät perustoiminnaltaan ihmisaivojen rakennetta, mutta ainoastaan aivot pystyvät monimutkaiseen samanaikaiseen prosessointiin, joista nykyisillä laskentavälineillä voidaan vain haaveilla.

Neuroni muodostaa verkon toiminnan perustan suorittaen tiedonkäsittelyn, joka toiminnaltaan vastaa ihmisillä lyhytkestoista muistia (Törmä 1997). Ihmisten aivoissa sijaitsee miljoonia *neuroneita* eli hermosoluja ja vastaavasti neuroverkoissa on vain satoja solmuja (Roiger & Geatz 2003). Ihmisten aivoilla on mahdollisuus oppia ja käsitellä uutta tietoa sekä soveltaa niitä uusissa ratkaisuissa. Neuroverkkojen tavoitteena on ollut kehittää aivoja jäljitteleviä menetelmiä, mutta kuitenkin näiden kahden toiminnan vertaaminen keskenään on melko pinnallista.

Neuronien toiminta perustuu vuorovaikutukseen *syötetason* (*I*), *piilotettujen tasojen* (*H*) ja *tulostason* (*O*) neuronien kesken kuvan 4.4 mukaisesti. Piilotettujen tasojen lukumäärä voi vaihdella syötteiden lukumäärän ja ongelman monimutkaisuuden mukaan (Bounsaythip & Rinta-Runsala 2001). Neuroverkon jokaisen tason kaikista neuroneista on kytkentä aina seuraavan tason kaikkiin neuroneihin, jolloin verkkoa kutsutaan *täysin kytketyksi* (Roiger & Geatz 2003). Samalla tasolla olevilla neuroneilla ei kuitenkaan ole kytkentää keskenään. Verkon opetusvaiheessa jokaista syötettä muutetaan painokertoimella (*W*), joka esittää saapuvan syötteen voimakkuutta. Tämän seurauksesta syötteiden vaikutusta verkon käyttäytymiseen vahvistetaan tai heikennetään. Myös kytkentöjen painoja voidaan muuttaa niiden oppimislain perusteella, jolloin verkko suoriutuu tehtävästään parhaiten (Törmä 1997). Neuroverkon opetus vaatii suuren aineiston sekä aikaa, mutta kerran opetettua aineistoa voidaan myöhemmin käyttää uusiin analysoitaviin aineistoihin (Bounsaythip & Rinta-Runsala 2001).



Kuva 4.4: Neuroverkon eräs mahdollinen rakenne (Bounsaythip & Rinta-Runsala 2001).

Neuroverkon toiminta havaitaan matemaattisesti kaavasta 11, jossa neuroniin (*i*) tuleva syöte (*x*) kerrotaan sitä vastaavalla neuronista (*j*) saapuvalla painokertoimella (*w*). Neuronin kokonaissyötteen (u_i) arvo saadaan yhteenlaskettaessa neuroniin saapuvat painokertoimella korjatut syötteet, joiden maksimilukumäärää kuvaa kaavassa *n*-muuttuja.

$$u_i = \sum_{j=1}^{n} w_{ij} x_j \qquad (11)$$

Neuroverkon laskentamalleja voidaan luokitella monin eri tavoin, joista eräitä tapoja on luokitella ne tehtävätyypin tai verkon rakenteen mukaisesti. Tehtävätyypin mukaan ne jakautuvat neljään luokkaan: *luokittelu-, assosiaatio-, optimointi-* ja *itseorganisoituvat mallit* (Fu 1994). Vastaavasti rakenteen perusteella verkot jakautuvat neuronin ja verkon rakenteeseen sekä oppimisalgoritmin rakenteeseen (Rojas 1996).

Oppimisalgoritmin rakenteen perusteella neuroverkot luokitellaan vielä toiminnan mukaan, koska aluksi verkot eivät osaa tehdä mitään ja niitä on opetettava. Opettamiseen käytetään opetusalgoritmia, jolloin neuroverkolle annetaan joukko syötteitä niiden ominaisuuksien oppimista varten. Opetusalgoritmeja on useita eri neuroverkon rakenteen ja käyttötarkoituksen mukaisesti. Opetusalgoritmit jaetaan yleisesti niiden toiminnan perusteella ennustavaan data-analyysiin eli *ohjattuun* (*supervised learning*) ja tutkivaan data-analyysiin eli *ohjaamattomaan opetukseen* (*unsupervised learning*).

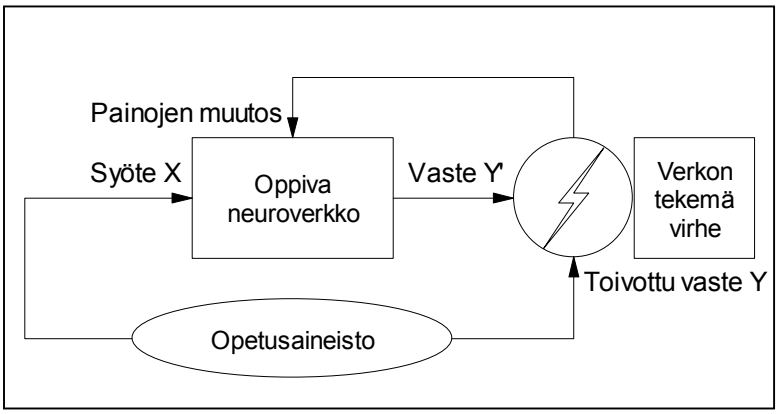
Ohjattuja menetelmiä ovat *luokittelu* (*classification*), *ennustaminen* (*regression*), *päätöspuut* (*decision trees*) ja *arvioiminen* (estimation). Ohjaamattomista menetelmistä voidaan mainita *itseorganisoituminen* (self organization) ja *ryhmittely* (clustering), joiden toimintaa tarkastellaan tarkemmin ryhmittelyanalyysin yhteydessä luvussa 5.

Näiden lisäksi käytetään myös tilastollisia menetelmiä kuten *lineaarinen* (*linear regression*) ja *logistinen ennustaminen* (*logistic regression*) (Parr Rud 2001) sekä *kiinteäpainoisia tilansiirtoverkkoja* (Koikkalainen 1994). Tilansiirtoverkoissa ei tapahdu suoranaista oppimista, vaan neuronien painot määritellään käyttäjän toimesta etukäteen valmiiksi. Painojen määrittelyssä käytetään matemaattisia menetelmiä, josta tunnetuin on energiafunktion käyttö (Koikkalainen 1994).

4.3.1 Ohjattu opettaminen

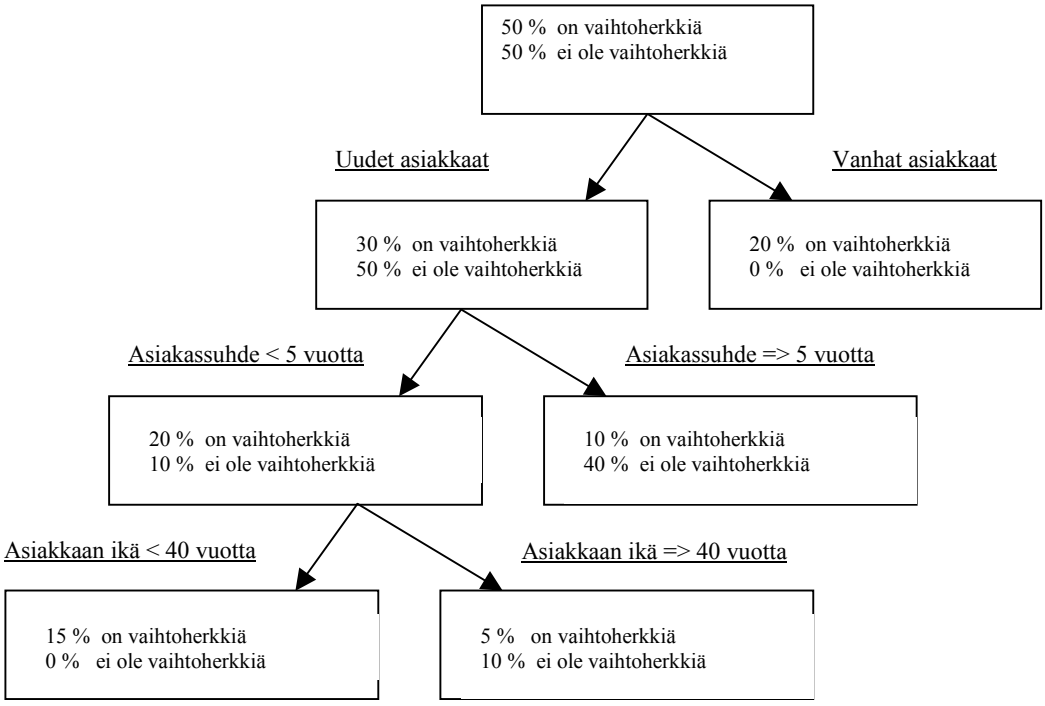
Opetettaessa neuroverkkoa esimerkiksi tunnistamaan kukkia ohjelmalle annetaan syötteenä paljon erilaisia kukkien piirteitä (kuten tuoksu, koko, väri), joita painotetaan niiden yleisyyden mukaan. Tällöin ei kuitenkaan opeteta sitä, mikä tekee ruususta ruusun vaan sitä, kuinka ohjelma oppii yhdistämään ruusun piirteet sen oikean merkitykseen.

Yksi neurolaskennan keskeinen periaate on juuri tämä esimerkeistä oppiminen, jolloin neurolaskennan teho perustuu kykyyn oppia tehtävien ratkaisut sekä soveltaa tehtyjä ratkaistuja ennakkotapauksina. Kuvan 4.5 oppivaverkko toimii myös järkevästi sellaisille syötteille, joille ei löydy opetusaineistoa. Laskennallisesti opetus on kuitenkin aikaa vievää, mutta opetusvaiheen jälkeen tapahtuva käyttö on tehokasta. Tunnetuin opetusalgoritmi on virheen takaisinkorjausmenetelmä eli *backpropagation* (Koikkalainen 1994). Tämä algoritmi perustuu esimerkkien käyttöön, jolloin lasketaan verkon vastaus ja sitä verrataan haluttuun vastaukseen. Tarvittaessa kytkentöjen painoja muutetaan, jolloin seuraavilla kerroilla verkon toiminta aiheuttaa nykyistä pienemmän virheen.



Kuva 4.5: Ohjatun opettamisen periaate (Koikkalainen 1994).

Luokittelu kuluu ennustaviin menetelmiin, jolloin luokittelussa opitaan aikaisemmin rakennetun opetusjoukon pohjalta. Tällöin esimerkiksi ihminen luokittelee jatkuvasti tapahtuvia asioita aikaisempien havaintojen perusteella (esim. kylmä, kuuma). Eräs yleisimmin käytetty sopiva luokittelutekniikka on päätöspuut. Tämä toiminta perustuu ennustemalliin, jossa apuna käytetään helposti ymmärrettävää puurakennetta. Päätöspuun toiminta perustuu tiedon löytymiseen haarautumisten kautta, eli jokaiselta oksalta haaraudutaan aina kahdelle uudelle oksalle (kaava 4.6). Tällöin lapsioksan havaintojen yhteenlaskettu summa on aina isätason vastaavan havaintojen lukumäärä.

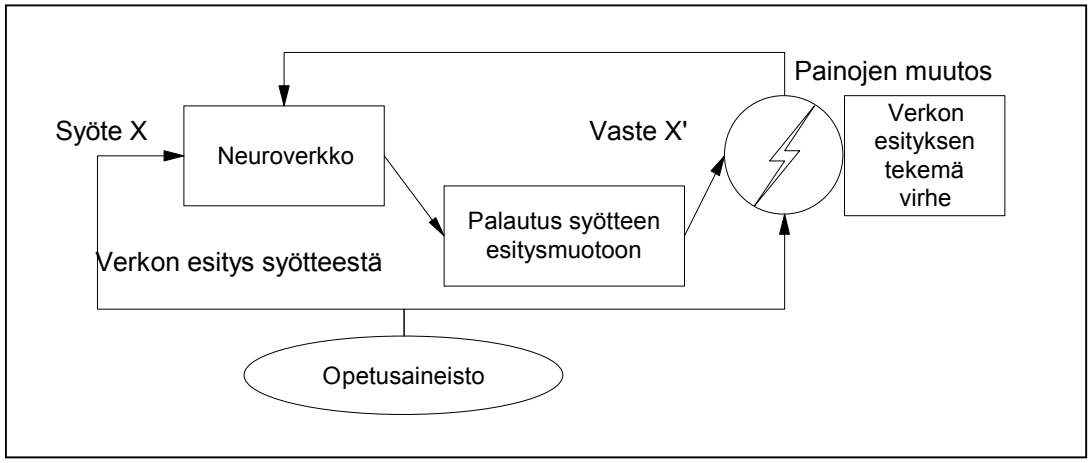


Kuva 4.6: Esimerkki asiakkaan vaihtoherkkyydestä päätöspuuta käyttäen.

Luokittelun lopputuloksena saadaan vastauksena erilaisia ilmaisuja (esim. täysi/tyhjä, kyllä/ei), joille on vaikeata ja suorastaan mahdotonta löytää tarkkoja raja-arvoja. Esimerkkinä voidaan todeta, että on helppoa rajata täyden ja tyhjän lasin raja-alue puolikkaaksi, mutta tällöin ongelman tarkastelu siirtyy seuraavaksi puolikkaan ja täyden raja-alueelle. Tämän jälkeen ongelmana on määritellä, milloin lasi on täynnä, puolillaan tai melkein täysi. Arvioimisessa voidaan huomata joukkoonkuulumisen olevan asteittainen liukuva käsite, jolloin arvoksi saadaan lasin ollessa täysi (1) ja puolillaan (0,5) (Roiger & Geatz 2003). Tällöin arvioiminen voidaan määritellä reaalilukuna nollan ja yhden väliltä, eikä ainoastaan ääriarvona kyllä tai ei. Vastaavasti ennustaminen on lähes sama menetelmä kuin luokittelu ja arvioiminen, kuitenkin erotuksena näistä siinä perustetaan ennustaminen tulevaan käyttäytymiseen pohjautuvaksi (Roiger & Geatz 2003).

4.3.2 Ohjaamaton opettaminen

Ohjaamattomassa opetuksessa eli kilpailuoppimisessa tunnetaan syöte, mutta ei sen vastetta. Oppimisen tehtävänä on löytää syötteistä yhtäläisyyksiä, joita hyödynnetään informaation tulkinnassa. Tätä menetelmää voidaan havainnollistaa siten, että syöteinformaation käyttäytymisestä muodostetaan malli neurolaskennalla. Opetusalgoritmissa neuronit kilpailevat keskenään siitä, mikä edustaa syötettä parhaimmin (Koikkalainen 1994). Tunnetuin ohjaamattoman opetuksen menetelmistä on Teuvo Kohosen kehittämä itseorganisoituvat kartat (Kohonen 1997). Itseoppimisen tarkoitus on löytää samanlaisuuksia, koska näitä voidaan hyödyntää informaation tulkinnassa. Tässä menetelmässä neurolaskennalla muodostetaan syöteinformaation käyttäytymisestä malli kuvan 4.7 mukaisesti. Opetusalgoritmissa neuronit kilpailevat keskenään siitä, mikä vastaa sisällöltään syötettä parhaimmin (Koikkalainen 1994).



Kuva 4.7: Ohjaamattoman opetuksen periaate (Koikkalainen 1994).

Ryhmittelyä kutsutaan yleisesti ohjaamattomaksi oppimiseksi, koska käyttäjän ei tarvitse määritellä louhinnassa käytettäviä luokkia, vaan algoritmin on tarkoituksena löytää ja määritellä ne sekä niiden lukumäärä olemassa olevan tiedon pohjalta. Ryhmittelyssä aineisto jaetaan sisällön perusteella automaattisesti erilaisiin ryhmiin niiden samankaltaisuuden perusteella. Aineistosta etsitään siinä olevia kasautumia pyrkien löytämään ne arvot, jotka ovat lähellä toisiaan ja liittäen ne samankaltaisiin ryhmiin. Tällöin tietoalkiot saman ryhmän sisällä muistuttavat toisiaan enemmän kuin muissa ryhmissä sijaitsevat alkiot ja vastaavasti kaukana toisistaan sijaitsevat ryhmät ovat sisällöltään erilaisempia.

5 Ryhmittelyanalyysi

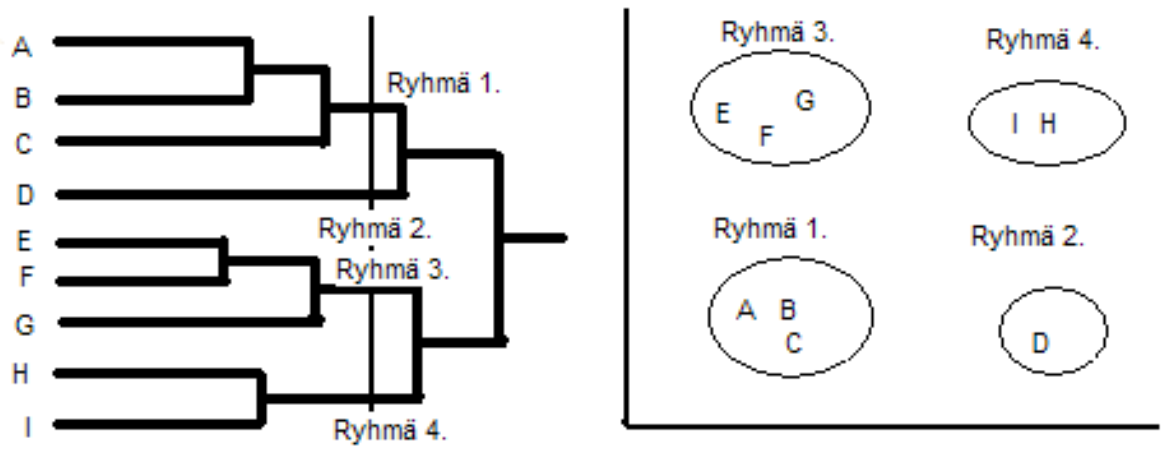
Ryhmittelyanalyysin tarkoituksena on löytää tietomäärästä viitteitä niiden keskinäisestä samankaltaisuudesta, jonka perusteella tiedot jaetaan samansisältöisiin ryhmiin tai luokitellaan havainnot jo aikaisemmin löydettyihin ryhmiin. Aineistot esitetään tavallisimmin matriisina, jolloin eri riveillä olevat vektorit ovat aineistossa olevia pisteitä. Tällöin sama tieto voidaan esittää myös pisteenä moniulotteisessa avaruudessa, jolloin avaruuden piste vastaa yhtä aineiston riviä.

Ryhmittelyanalyysin toiminta perustuu useisiin eri työvaiheisiin. Aluksi valitaan aineistosta ryhmiteltävät muuttujat ja muokataan niitä sopivammaksi ryhmittelyä sekä myös lopputuloksen parantamista varten. Seuraavaksi valitaan käytettävä etäisyysmitta, joka voi olla esimerkiksi euklidinen etäisyys tai korrelaatiokerroin. Kaikilla ryhmittelymenetelmillä saadaankin jonkinlaisia ryhmittelyjä lopputuloksiksi, mutta oikean algoritmin valinnalla varmistetaan tulosten oikeellisuus. Lopputuloksen esittäminen visuaalisesti parantaa ryhmittelyn ymmärtämistä, ja tulosten arviointi eli *validointi* on lopuksi tärkein toimenpide. Analyysin oikeellisuuden määrittäminen on tarpeellista suorittaa kriittisesti, koska ohjelmallisesti tuotettua ryhmittelyn tulosta ei voida aina pitää absoluuttisen varmana ja oikeana vastauksena.

Ryhmittelymenetelmät voidaan jakaa *hierarkkisiin* (*hierarchical*) ja *osittaviin* (*partition-based*) menetelmiin (Han & Kamber 2001). Suurin ero näiden kahden menetelmän välillä on siinä, että hierarkkisissa menetelmissä ei päätetä ryhmien lukumäärää etukäteen. Vastaavasti osittavissa menetelmissä alustetaan aina lähtökohdaksi jokin ehdotusratkaisu, jota sitten algoritmilla pyritään parantamaan.

Hierarkkiset menetelmät toimivat puurakenteen mukaisesti ja ne jaetaan kahteen eri ryhmään: *kokoaviin* (*aggolomerative*) ja *jakaviin* (*divise*) (Bearson & al. 2003). Kokoavat algoritmit muodostavat ratkaisun yhdistelemällä ryhmiä. Aluksi jokainen havainto vastaa yhtä ryhmää, jolloin ensimmäisellä askeleella etsitään samankaltaiset havainnot ja yhdistetään ne yhdeksi ryhmäksi. Seuraavaksi uusitaan menettely, jolloin löydetään jokin kahden havainnon ryhmä tai yhdistetään jokin havainto edelliseen ryhmään (kuva 5.1). Algoritmin edetessä ryhmien lukumäärä vähenee jokaisella askeleella yhdellä, kunnes kaikki havainnot kuuluvat yhteen ryhmään. Kuitenkin ryhmittely pitäisi keskeyttää silloin, kun ryhmittelyn kriteereissä tapahtuu

selvä muutos huonompaan. Vastaavasti jakavien algoritmien toiminta on päinvastaista, jolloin alun yksi ryhmä jakautuu kahtia siihen asti, kunnes jokaisessa ryhmässä on yksi havainto.



Kuva 5.1: Hierarkkisen ryhmittelyn kokoavan algoritmin toteutus (Jain & al. 1999).

Hierarkkisten menetelmien aikavaatimus on $\theta(n^2 log\ n)$ ja tilavaatimus $\theta(n^2)$, missä n on alkioiden lukumäärä (Jain & al. 1999). Näiden molempien menetelmien käytön aikana tapahtuneet valinnat ovat myöhemmin peruuttamattomia, jolloin alussa tehdyt huonot valinnat vaikuttavat myös loppuun asti. Käytettyjä menetelmiä ovat *painopistemenetelmä* (*centroid method*), *keskiarvomenetelmä* (*average linkage method*) ja *lähimmän naapurin menetelmä* (*nearest neighbour*).

Osittavassa ryhmittelyssä oletetaan ryhmien lukumäärästä olevan esitietoa, jonka perusteella havainnot sijoitetaan ensimmäisellä kerralla. Alkuryhmitystä parannetaan siirtämällä havaintoja ryhmästä toiseen samankaltaisuuden perusteella, joka perustuu havaintojen iteratiiviseen ryhmittelyyn. Lopullinen ratkaisu saavutetaan silloin, kun siirrot eivät paranna enää ryhmien sisältöä. Hierarkkisista menetelmistä poiketen havainnon paikkaa voidaan tarvittaessa muuttaa eri ryhmien välillä, mutta ryhmien lukumäärä on sidottu. Näistä menetelmistä hyvänä esimerkkinä ovat K-means ja RLS-algoritmi. Osittavien menetelmien aikavaatimus on yleensä lineaarinen ja tilavaatimus $\theta(n)$, missä n on alkioiden lukumäärä (Jain & al. 1999).

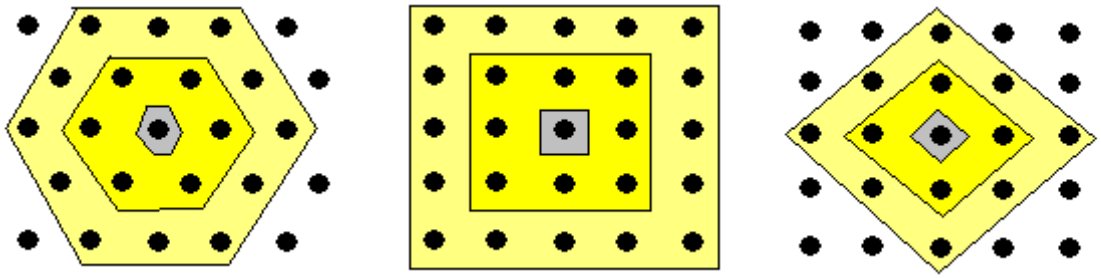
5.1 Itseorganisoituva kartta

Tunnetuin *ohjaamattoman* (unsupervised) oppimisen menetelmistä on Kohosen 1980-luvulla kehittämä *itseorganisoituva kartta* eli *SOM* (Self-Organizing Map) (Kohonen 1997). Verkon oppimistapana on kilpaileva oppiminen, jolloin vierekkäiset solut kilpailevat syötteistä ja vaikuttavat toisiinsa oppien vähitellen paremmin tunnistamaan samankaltaisia syötteitä. Aikaisemmin SOM-algoritmin toimintaa on kuvattu karkealla tasolla jo markkina-analyysin tarkastelun yhteydessä kohdassa 3.2.1.

Itseorganisoituva kartta on neuroverkkoalgoritmi, jossa ei käytetä piilotettuja tasoja vaan ainoastaan lähtö- ja tulostasoa. Kartan avulla on helppo visualisoida moniulotteisesta tiedosta kaksiulotteisia karttoja (Bounsaythip & Rinta-Runsala 2001). Kartta tiivistää alkuperäiset tiedot säilyttäen sen aineiston topologiset mittasuhteet, jolloin ne vektorit, jotka ovat lähekkäin moniulotteisessa avaruudessa ovat lähekkäin myös kartalla. Karttayksiköt eli *neuronit* muuttavat painokertoimiaan kohti tilaa, jossa jokainen neuroni on keskimääräisesti yhtä usein aktiivinen ja aktiivisuudet ovat jakautuneet mahdollisimman tasaisesti erilaisten syötteiden joukkoon. Kartalla olevat neuronit ovat vuorovaikutuksessa ympärillä sijaitsevien neuronien kanssa. Tällöin lähellä toisiaan olevat neuronit reagoivat samantyyppisiin syötteisiin, kun taas kauempana olevat neuronit reagoivat erityyppisiin syötteisiin.

5.1.1 Kartan rakenne

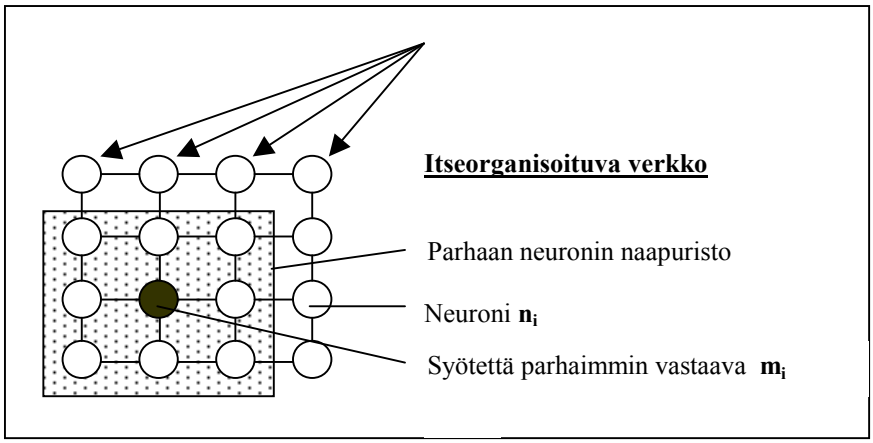
Algoritmin käyttäjä valitsee aluksi kartan topologian, joka voi olla kuusikulmainen, nelikulmainen tai suunnikas (kuva 5.2). Tällöin kartan valinnan yhteydessä määritellään myös neuronien kytkentätapa muihin naapurineuroneihin (kuva 5.3). Näiden lisäksi käyttäjä arvioi tarvittavan kartan mittasuhteet sekä tietoaineistossa esiintyvien ryhmien lukumäärän. Virheellisesti voidaan ajatella, että kartan koko ei merkitse mitään, mutta se on kuitenkin tärkein tekijä järjestymisen onnistumisessa (Myllyniemi & Sarjakoski 1996). Mikäli alue on pieni, tällöin ei muodostu ryhmiä vaan kaikki ovat yhtä samaa ryhmää. Jos alue on liian iso, tällöin on vaarana syntyä liian sulkeutuneita ryhmiä. Ryhmien lukumäärän valinta on algoritmissa tehtävä etukäteen, mikä johtaa siihen, että ryhmittelyä pitää kokeilla useilla eri ryhmien lukumäärällä. Tämän lisäksi jokaisella samasta aineistosta tehtävällä eri suorituskerralla muodostuu aina lopputuloksena hieman erilaisesti ryhmitelty kartta.



Kuva 5.2: Itseorganisoituvan kartan topologia (kuusikulmainen, neliömäinen, suunnikas).

Kuva 5.3: Neuronien kytkentätavat (kuusikulmainen, neliömäinen, suunnikas).

Itseorganisoituva kartta on yleensä yksi- tai kaksiulotteinen neuroverkko (kuva 5.4), jossa jokainen solmu on neuroni n_i ja jokaiseen karttayksikköön i liittyy aina painovektori $m_i=(y_{i1}, y_{i2}, \ldots y_{in})$. Nämä painovektorin muuttujien arvot kuvaavat neuronin ominaisuuksia. Karttayksikön ja syötettävän datan (x) välille määritellään etäisyysmitta $||x- m_i||$ (Kohonen 1997), tällöin karttayksiköiden välisellä etäisyydellä tarkoitetaan yksiköiden välistä etäisyyttä tila-avaruudessa.



Kuva 5.4: Itseorganisoituvan kartan periaate.

5.1.2 Algoritmin toiminta

Ennen opetuksen aloittamista karttayksiköt alustetaan, jolloin taulukolle annetaan alkuarvot. Alustaminen suoritetaan yksinkertaisimmin täysin satunnaisilla arvoilla tai valitun opetusjoukon arvoalueen arvoilla. Kartan alustamistapa voi vaikuttaa järjestymisnopeuteen opetuksen edetessä, jolloin kuitenkin karkean alustuksen jäljiltä tapahtunut opetus johtaa usein myös järkevään lopputulokseen. Kartan opetus tehdään lukemalla tiedostosta arvoja, jotka syötetään opetusalgoritmille arvo kerrallaan. Kartan jokaiselle neuronille syötetään sama syöte eli syötevektori, jolloin opetusdatan etäisyys (samankaltaisuus) lasketaan kartan jokaiseen karttayksikköön (kaava 12). Samankaltaisuuden määrittelyssä käytetään usein euklidista etäisyyttä syöte- ja painovektorin välillä.

$$\left|x - m_c\right| = \min_i \left|x - m_i\right| \qquad (12)$$

Jokaisella algoritmin suorituskerralla verrataan yhtä syötevektoria kerrallaan kartan kaikkien solujen painovektoreihin. Kun kartan jokainen karttayksikkö on käyty läpi, on löytynyt pienin etäisyyden arvo syötevektorin ja kartan neuronin välillä. Tätä lähimpänä opetusdatan arvoa sillä hetkellä olevaa kartan yksikköä kutsutaan *BMU-voittajaneuroniksi* (*Best Matching Unit*). Tämän jälkeen kartan neuronin ja sitä ympäröivien neuronien arvoa muutetaan α-kertoimella lähemmäksi opetusdataa kaavan 13 mukaisesti.

$$m_i(t+1) = m_i(t) + \alpha(t)\left[x(t) - m_i(t)\right] \qquad (13)$$

Korjauksen jälkeen valitaan uusi syötevektori ja toistetaan opetusta. Opetuksen tavoitteena on saada kartta järjestymään ja pienentää topologinen virhe mahdollisimman pieneksi. Algoritmia suoritetaan niin monta kertaa kuin tarvitaan tarpeeksi tarkan lopputuloksen saamiseksi. Suorituskertoihin ei vaikuta itse syötevektorien lukumäärä vaan ryhmityksen muodostuminen. Tällöin algoritmia voidaan soveltaa hyvinkin suurten aineistojen analyysiin, mutta pientä aineistoa joudutaan kierrättämään algoritmissa useita kertoja halutun oppimistuloksen saavuttamiseksi (Kohonen 1997).

Kartan järjestymisen nopeuteen vaikutetaan opetusparametrien oikealla valinnalla. Alustavana sääntönä Kohonen mainitsee kirjassaan ensimmäisen tuhannen opetusaskeleen olevan karkean opetuksen vaiheen, jonka aikana naapuruston koko voi alussa olla jopa yli puolet kartan koosta pienentyen aina yhteen yksikköön. Tällöin α-oppimisnopeuskertoimen alkuarvon on lähellä yhtä (Kohonen 1997). Kartassa tapahtuu voimakasta järjestäytymistä, jolloin suurella naapurustolla tasoitetaan voittajaneuronin ympäristöä samankaltaisemmaksi. Toisen vaiheen alussa naapuruston koko voi olla pienempi, jolloin myös α-kertoimen arvo pienenee opetuksen edetessä lineaarisesti kohti nollaa. Tällöin kertoimen arvon lähestyessä nollaa muutokset pienevät ja muuttuvat ainoastaan hienosäädöksi. Oppimiskerroin on yleensä lineaarisesti ajan suhteen vähenevä tai funktion (kaava 14) mukaisesti ajan suhteen vähenevä.

$$\alpha(t)=\frac{A}{t+B},\left(0<\alpha<1\right) \qquad (14)$$

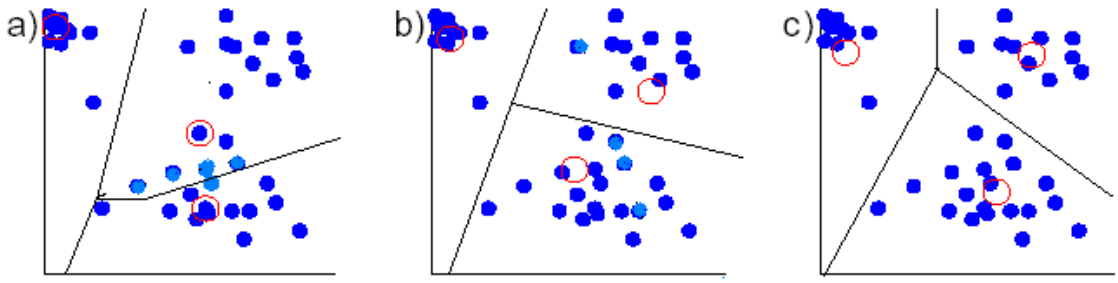
Kohosen itseorganisoituva kartta soveltuu hyvin ryhmittelyanalyysin toteutukseen, mutta sitä voidaan myös käyttää luokittelussa. Tällöin kartan opettaminen tapahtuu normaalisesti, jonka jälkeen jokaiselle neuronille annetaan luokka opetusjoukon avulla. Luokittelussa painovektorille etsitään lähin neuroni ja luokitellaan se neuronia vastaavaan luokkaan.

Itseoraganisoituvat kartat välttävät hierarkkisten menetelmien ongelmat, koska ne eivät ole hierarkkisten menetelmien tavoin kovin herkkiä pienille tietojen muutoksille. Niiden ongelmana on käyttäjän vaikeus osata valita klustereiden oikea lukumäärä etukäteen. Vastaavasti karttojen etuna voidaan pitää helppoa ymmärrettävyyttä ja lopputulosten kuvien visualisoitujen tulkintaa.

5.2 K-means algoritmi

K-means algoritmi on yksinkertainen ja kohtalaisen tehokaskin ryhmittelymenetelmä, jolloin sen käyttöönottaminen on ollut helppoa. K-means on toimintatapana osittava menetelmä, joka perustuu havaintojen iteratiiviseen jakamiseen toistuvasti ryhmiin pienentämällä Euclidean etäisyyttä havainnon ja ryhmän välillä (Ahola & Rinta-Runsala 2001). Tämän perusteella

ryhmien paikat muuttuvat algoritmien edetessä kuitenkin niiden määrän pysyessä vakiona (kuva 5.5).



Kuva 5.5: K-means algoritmin karkea toiminta-ajatus (Ahola & Rinta-Runsala 2001).

a) Ympyrät ovat alkuperäisiä keskuspisteitä ja pisteet ovat havaintoja.

b) Muuttuneet keskuspisteet ja rajat ensimmäisen toiston jälkeen.

c) Lopulliset keskuspisteet ja rajat.

Aluksi käyttäjä sijoittaa alkiot tietoavaruuteen paikoilleen ja määrittelee tarvittavien ryhmien lukumäärän (k). Seuraavaksi valitaan näistä kartalla olevista pisteistä satunnaisesti (t=1) k kappaletta pisteitä ryhmien keskipisteiksi (Jain & al. 1999). Kaavan 15 mukaan X on algoritmissa aineiston sisältävä matriisi, jonka riviin j viitataan X_j:llä. Vastaavasti k on haluttu ryhmien lukumäärä ja w on matriisi, jossa ovat ryhmien keskipisteet.

$$w_i^{(t)} = X_j, j = rand(1,...,n), i = (1,...,k) \tag{15}$$

Seuraavaksi samankaltaisuutta mitataan etäisyydellä, jolloin K-means algoritmi laskee jokaisesta ryhmän keskipisteestä etäisyydet aineiston muihin pisteisiin. Tämän jälkeen verrataan ryhmäkeskipisteiden etäisyyksiä pisteistä siten, että kuhunkin ryhmään sijoitetaan kuulumaan ne pisteet, jotka ovat sen ryhmän keskipistettä lähimpänä. Tällöin alkuvaiheessa voi tapahtua ryhmien välillä suuriakin ryhmäjaon muuttumisia. S-matriisi sisältää tiedon siitä, mihin ryppääseen kukin data-aineiston alkio kuuluu. Saman matriisin rivinumero kertoo ryhmän keskipisteen rivin w-matriisissa ja alkion arvo ryppääseen kuuluvan alkion rivin X-matriisissa (kaava 16).

$$S_i = \{l \mid 1 <= l <= n, d(X_l, w_i^{(t)}) \le d(X_l, w_j^{(t)}) \forall j = 1,...,k, j \neq i\}, \tag{16}$$

$i=1, \ldots, k$

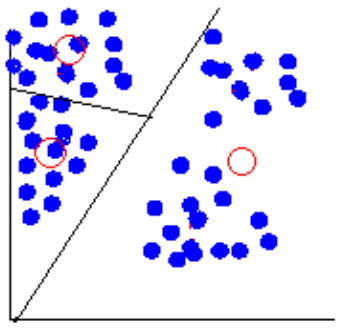
Tämän uuden ryhmäjaon jälkeen lasketaan keskipisteet kullekin ryhmälle uudelleen tämän uuden ryhmittelyn perusteella ja verrataan uusia keskipisteitä uudelleen taas muihin pisteisiin. Tätä kahta askelta toistetaan uudelleen ja uudelleen, kunnes keskipisteet ovat vakiintuneet paikoilleen ja paikallinen minimi on saavutettu (Fränti & Kivijärvi 2000) tai keskipisteiden paikka ei enää huomattavasti vaihdu (kaava 17).

$$w_i^{(t+1)} = \frac{1}{\#\, S_i} \sum_{j=1}^{n} X_l, l \in S_{i,j}, i = (1,...,k) \tag{17}$$

Algoritmin suorittaminen lopetetaan, jos ehto kaavan 18 mukaan täyttyy, muuten $t = t+1$ ja jatketaan kaavan 16 mukaan. Koko algoritmin laskennallinen aikavaatimus on $\theta(tKn)$, jossa t on iteraatioiden lukumäärä, K on ryhmien lukumäärä ja n on datapisteiden lukumäärä.

$$\forall i \,||\, w_i^{(t+1)} - w_i^{(t)} \,|| \leq \varepsilon, i = (1,...,k) \tag{18}$$

K-means algoritmin toimintapa on yksinkertainen ja sen käyttäminen on helppoa (Fränti & Kivijärvi 2000). Algoritmi soveltuu käytettäväksi, kun tiedetään tai ollaan suhteellisen varmoja siitä, kuinka monta ryhmää aineistossa on luonnostaan. Valitettavasti tätä ei yleensä pystytä tietämään aineistosta, joten kokeilun kautta päädytään usein sopivaan ryhmien lukumäärään. Valitettavasti algoritmi tekee paikallisia muutoksia alkuperäiseen ryhmittelyyn, jolloin se voi antaa väärän tuloksen ja jää jumiin paikalliseen minimiin (Fränti & Kivijärvi 2000). Virhe voi tapahtua alussa ryhmien keskipisteiden satunnaisen valinnan epäonnistuttua kuvan 5.6 mukaisesti, jolloin K-means algoritmi ei löydä välttämättä optimia ryhmäjakoa. Tämän takia tarkastelemme seuraavaksi RLS-algoritmia, jonka pitäisi pystyä suoriutumaan paremmin tällaisesta ongelmasta.



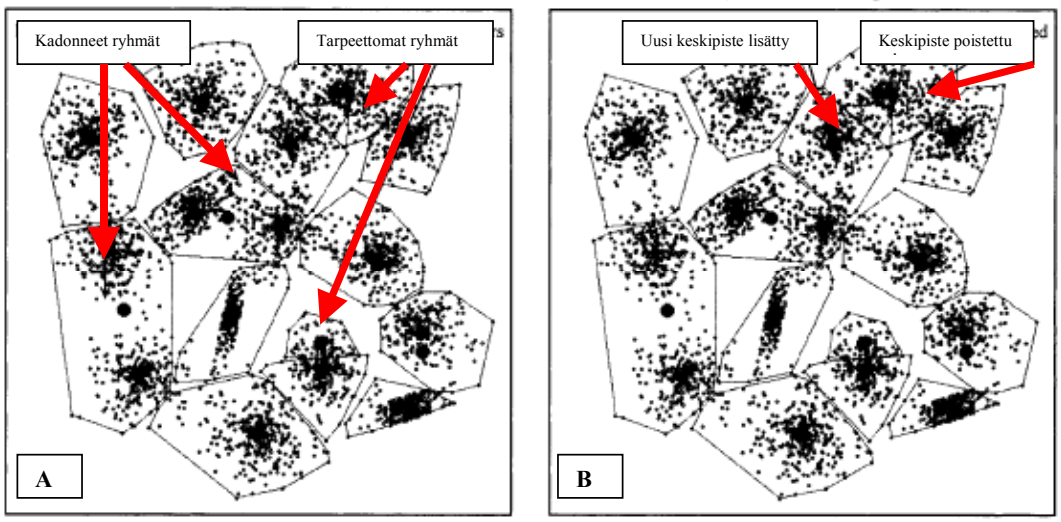
Kuva 5.6: K-means algoritmin alkutilanteen ongelma.

5.3 Satunnaistettu paikallishaku -algoritmi

Randomized Local Search (RLS) on uudempi ryhmittelyalgoritmi, joka perustuu perinteiseen local search -optimointitekniikkaan. Tärkein RLS-algoritmin suunnittelun periaate on ollut saada aikaiseksi korkea ryhmittelyn laatu sekä säilyttää yksinkertainen toteutus että laskennallinen helppous (Fränti & Kivijärvi 2000).

Algoritmin toiminnan perusajatus on, että olemassa olevaa ratkaisua parannetaan vaiheittain. Algoritmin toiminta alkaa alkuratkaisusta, jossa sijoitetaan alkiot tietoavaruuteen paikoilleen ja seuraavaksi algoritmi valitsee käyttäjän määräämän määrän satunnaisia alkioita ryhmien edustajiksi. Nyt suoritetaan ryhmäjako, jolloin algoritmi laskee etäisyydet jokaisesta ryhmän keskipisteestä kaikkiin aineiston muihin pisteisiin. Tämän jälkeen verrataan ryhmäkeskipisteiden etäisyyksiä pisteistä siten, että kuhunkin ryhmään sijoitetaan kuulumaan ne pisteet, jotka ovat sen ryhmän keskipistettä lähimpänä. Tähän asti algoritmin toiminta on samanlaista kuin jo aikaisemmin esitellyssä K-means algoritmissa.

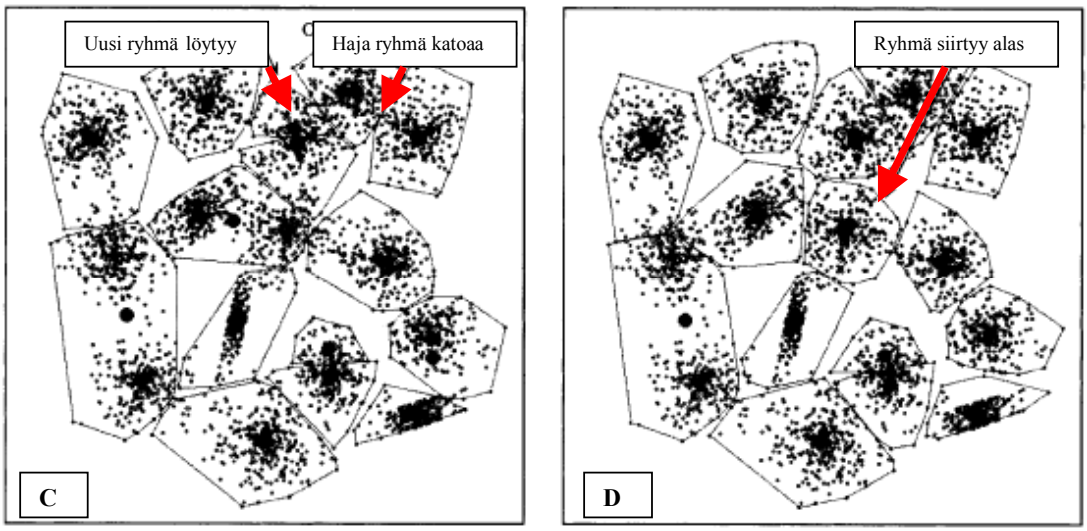
Alun ryhmittelyn tuloksena voidaan havaita kuvasta 5.7 aineistossa olevan kadonneita ryhmiä (kasautumat erillään) ja täysin turhia tai väärin sijoitettuja ryhmiä (pieniä kasautumia isojen ryhmien läheisyydessä). Näistä väärin sijoitetuista ryhmistä poistetaan yhdestä kerrallaan keskipiste ja luodaan uusi ryhmän keskipiste satunnaiseen paikkaan tietoavaruuteen, joka ei kuitenkaan ole välttämättä kovinkaan kaukana sen aikaisemmasta sijainnista. Käsiteltävä ryhmä voidaan myös valita determistisesti, jolloin ohjelma valitsee ryhmän, jonka puuttuminen huonontaisi ryhmittelyn lopputulosta vähiten (Fränti & Kivijärvi 2000). Vastaavasti uusi ryhmä lisätään sen ryhmän ympäristöön, missä ryhmässä on suurin vääristymä.



Kuva 5.7: RLS-algoritmin toiminta (Fränti & Kivijärvi 2000).
A) alkuperäinen ratkaisu ja B) satunnainen vaihto

Tarkoituksena on muodostaa näistä uusista valinnoista lokaalisti optimaalisia ratkaisuja naapuristofunktion (*neighborhood function*) avulla (Fränti & Kivijärvi 2000). Ratkaisua pidetään yleisesti lokaalisti optimaalisena silloin, kun se on lopputulokseltaan parempi kuin kaikki muut ratkaisut kyseisessä naapurustossa.

Aineiston uudelleen jakamisessa tämä väärin sijoitettu vanha ryhmä hävitetään ja uusi ryhmä sijoittuu tämän uuden keskipisteen ympäristöön. Tällöin myös naapuriryhmien kesken tapahtuu uudelleen järjestymistä ja keskipisteiden siirtymisiä sekä ryhmäjaon muuttumisia (kuva 5.8). Algoritmin edetessä aineistosta parannetaan vain yhtä ryhmää kerrallaan, jolloin uusi ryhmä hyväksytään sen parantaessa ratkaisua tai muussa tapauksessa se hylätään. Tämä satunnainen vaihto voi jo osittain ratkaista ongelman, mutta se voi myös sijoittaa keskipisteen myös väärään tai huonoon paikkaan mm. toisen keskipisteen läheisyyteen. Tällöin tätä virhettä voidaan korjata käyttämällä K-means algoritmiä lopulliseen hienosäätöön keskipisteen sijoittamiseksi sen oikealle paikalle (Fränti & Kivijärvi 2000). Artikkelin mukaan jo K-means algoritmin suorittaminen kahteen kertaan riittää tavallisesti esittämään näkemyksen, paraneeko ryhmittely tällä tavalla vai ei. Näistä uusista ratkaisuista valitaan paras uudelleen käsiteltäväksi ja tätä parannetaan iteratiivisesti, kunnes jokin lopettamisen tavoite on saavutettu tai lopetetaan ajallisesti suoritus.



Kuva 5.8: RLS-algoritmin toiminta (Fränti & Kivijärvi 2000).
C) uudelleen järjestely ja D) k-means iteraatiot

Lokaalisen haun menetelmillä on kyky saavuttaa nopeasti suurissakin ongelmissa ratkaisuja, vaikka ne eivät välttämättä ole optimaalisia. Lokaalin haun suurimpana ongelmana pidetään sen jäämistä kiinni usein lokaaliin optimiin, mikä johtuu siitä, että menetelmät eivät pysty hahmottamaan lokaalia ympäristöä kauempana olevia parempia ratkaisuja (Fränti & Kivijärvi 2000). Kuitenkin tutkimuksissa on vahvasti esitetty, että RLS-algoritmi on yksi tämän hetken parhaimmista ryhmittelyalgoritmeista niin laadullisesti kuin myös ajallisesti mitattuna.

Seuraavassa luvussa tutustutaan tarkemmin vakuutustietoaineiston tiedonlouhintaan ja sen vastauksena saatujen ryhmien hyväksikäyttöön asiakkuudenhallinnan toteutuksessa. Tiedonlouhinta suoritetaan käyttäen RLS-ryhmittelyalgoritmia.

6 Tiedonlouhinta vakuutusalalla

Viime vuosina on yrityksille alkanut vähitellen selvitä, että yrityksen tärkeimmät asiat eivät välttämättä ole työntekijät, brändit, tehtaat tai taseet, vaan ennen kaikkea niiden uskolliset asiakkaat (Billington 1999). Tämä tarkoittaa sitä, että useimmilla toimialoilla uusien asiakkaiden hankkiminen on käynyt entistä kalliimmaksi ja vaikeammaksi. Kun yritykset tietävät, minkälaiset asiakkaat ovat heille hyviä ja uskollisia, yrityksillä pitäisi olla mahdollisuus tutkia niiden asiakkaiden käyttäytymistä, jotka ovat jo siirtyneet muiden yritysten asiakkaiksi. Valitettavan usein olen myös itse huomannut, että yritykset lyövät laimin tämän vaiheen, jolloin markkinointi ja myynti keskittyvät enemmänkin uusiin asiakkaisiin kuin ilmeisiin menetyksiin. Kuitenkin menetetyistä asiakassuhteista voidaan löytää vihjeitä, kuinka menetykset saadaan tulevaisuudessa takaisin asiakkaiksi tai estetään vastaisuudessa samanlaisten asiakkaiden menettäminen. Joskus jopa pelkkä yhteydenotto menetettyihin asiakkaisiin voi palauttaa 30 %:ssa tapauksista asiakassuhteen takaisin (Billington 1999).

Tiedonlouhinnan yleisimpiä sovelluskohteita vakuutusalalla ovat mm. riskien ja markkinointikampanjoiden ennustaminen sekä asiakasryhmien tunnistamisen ja käyttäytymisen analyysi (Karanta 2002). Riskien ennustamisessa pyritään löytämään eri asiakasryhmien riskiprofiilit, jolloin vakuutuksenottajan ja hänen vakuuttamansa kohteen riskit pystytään mittamaan vakuutuksenantajan kannalta katsottuna kannattavaksi. Asiakasprofilointi, jota tässä tutkielmassa nimitettäköön *asiakaskorianalyysiksi,* on lähinnä tarkoitettu asiakassuhteen ylläpitoon ja asiakkaan käyttäytymisen ennustamiseen.

Näiden lisäksi tiedonlouhintaa voidaan soveltaa rikollisten käyttäytymisen analyysiin, vakuutuspetosten tunnistamiseen, kannattavuusanalyysiin, sairauksien diagnosointiin potilasdatan perusteella ja jälleenvakuutustason määrittämiseen (Karanta 2002). Vakuutusrikollisuuden tunnistamisessa haetaan petollisten, sekä vakuutusvilpillisten korvaushakemusten että itse vakuutuksenottajien yhdistäviä profiileja. Kannattavuus-analyysissä seurataan olemassa olevien tuotteiden kannattavuutta sekä määritetään uusien tuotteiden hintatasoja ja sopivuutta markkinoille. Sairauksien ennustamisella pyritään rajoittamaan sairausvakuutusten korvausmenon kasvua ja tulevia hinnan korotuksia, jolloin vakuutuksen myöntämisen yhteydessä (analysoinnin jälkeen) voitaisiin jo käyttää

rajoitusehtoa tiettyjen perinnöllisten tai piilevien sairauksien varalta. Jälleenvakuutuksessa tiedonlouhintaa voidaan käyttää ennustamaan tulevaa keskimääräistä vahinkosuhdetta, jolloin arvioidaan jälleenvakuutuksen suhdetta omalla riskillä olevaan vahinkomeno-olettamaan.

6.1 Vakuutusyhtiön asiakasvaihtuvuuden ongelma

Vahinkovakuutusta[1)] ei enää pidetä kasvavana alana, vaan viime vuosina erilaiset sijoitus-, eläke- ja henkivakuutukset ovat kiinnostaneet vakuutusyhtiöitä huomattavasti enemmän, koska näiden markkinat ovat olleet vielä osittain jakamatta. Kun täysin uusia vahinkovakuuttajia on vaikeata löytää, vahinkovakuutusmaksutulon kasvu pitää saadaan aikaan kilpailemalla olemassa olevista toisten yhtiöiden asiakkaista. Samanlainen kylläinen markkinatilanne on myös mm. matkapuhelinoperaattoreilla ja pankeilla, koska lähes jokaisella on jo matkapuhelinliittymä tai asiakassuhde jossakin pankissa. Näiden kilpailijoiden ja vähäisten uusien asiakkaiden tavoittelu aiheuttaa hintakilpailua, josta on seurauksena myös liikekulujen kasvaminen markkinointitoimenpiteiden johdosta. Hintojen laskeminen on aina kuluttajan kannalta hyvä asia, mutta vakuutusalalla sen vaikutus voi heijastua tulokseen vasta vuosien päästä yhdistetyn kulusuhteen[2)] huononemisena.

Kauppaketju tietää aina myydessään tuotteen kannattavan myyntihinnan, joka on sen tuotteen hankintahinta lisättynä yrityksen myyntikatteella. Vakuutusyhtiön vahinkomenon tarkka ennustaminen on sen sijaan vaikeaa, koska siinä on mukana myös ripaus tuuriakin. Joidenkin tuotteiden osalta on erittäin vaikea arvioida vahinkomenoa, hankalia mm. luonnonilmiöistä ovat myrskyt ja tulvat. Vakuutusyhtiö pystyy useiden vuosien historiatiedoista määrittelemään tilastollisesti laskennallisen vahinkosuhteen arvon, mutta tämäkään ei aina riitä. Tämän arvion epäonnistumisen seurauksena vakuutusyhtiö voi myöhemmin joutua korottamaan vuosittain vakuutusmaksujaan, jotta liiketoiminnan kannattavuus pystyttäisiin säilyttämään. Kuitenkin itse vahinkovakuuttaminen on usein liiketoimena tappiollista eli asiakkailta saatava vakuutusmaksutulo ei riitä kattamaan vahinkomenoja ja niiden hoidosta aiheutuvia liikekuluja, jolloin sijoitustoiminnan tuotot parantavat liiketoiminnan lopulta kannattavaksi.

[1)] Vahinkovakuutus sisältää palo-, metsä-, koti-, maatila-, kiinteistö-, yritys-, eläin- ja autovakuutukset.

[2)] Yhdistetty kulusuhde (%) saadaan laskemalla yhteen vahinkosuhde (vahingot per omalla vastuulla oleva maksutulo) ja liikekulusuhde (liikekulut per omalla vastuulla oleva maksutulo).

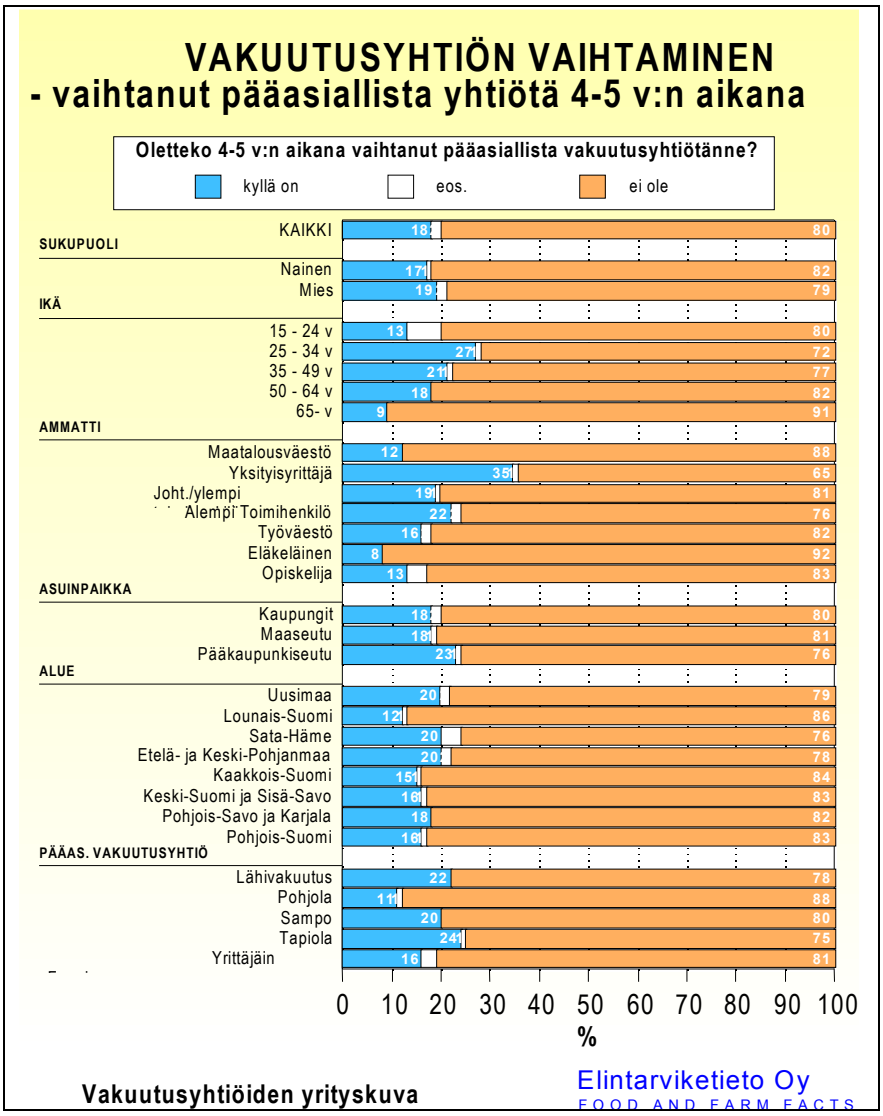
Asiakaspoistuma eli vanhojen asiakkaiden siirtyminen toisen yhtiön asiakkaiksi on periaatteessa aina huono asia, mikäli tämä ei johdu itse vakuutusyhtiön tekemistä toimenpiteistä irtisanoa riskialttiit ja maksamattomat asiakkaat. Tämä samanlainen asiakasvaihtuvuuden ongelma korostuu sellaisissa yrityksissä, joissa toiminta perustuu toistuvuuden kautta jatkuvuuteen sekä pitkäaikaisiin asiakassuhteisiin. Näistä mainituista yrityksistä hyvänä esimerkkinä ovat pankit, puhelin-, lehti- ja vakuutusyhtiöt. Asiakaspoistuman ennustaminen etukäteen antaa yritykselle mahdollisuuden kohdistaa yhteydenotot tai markkinointitoimenpiteet juuri näihin mahdollisesti lähteviin asiakkaisiin. Tällöin vanhan asiakassuhteen säilyttäminen olisi vielä mahdollista ilman suurempia toimenpiteitä, ainoastaan sisällyttämällä se asiakkuudenhallinnan yhdeksi toimintatavaksi.

Monesti kuitenkin mielikuva vakuutusyhtiöiden ominaisuuksien arvostuksesta voi asiakkaan mielestä olla ja onkin aivan erilainen kuin vakuutusyhtiön oma kuva. Tämä havaitaan myös tutkimuksesta (taulukko 6.1), jonka aineisto on kerätty vuonna 1999 kuluttajatutkimuksena 1 000 henkilökohtaisessa haastattelussa otannan edustaen koko Suomen 15 vuotta täyttänyttä väestöä. Tässä tutkimuksessa peräti 40 % haastatelluista on maininnut hinnan olevan tärkein asia vakuutuksia valittaessa. Tämä on mielestäni aika yllättävää, koska vakuutuksen sisällöissä ja ehdoissa on aina eroja vakuutusyhtiöiden välillä. Kuitenkin 32 % haastatelluista painottaa juuri tätä vakuutusmaksun edullisuutta suhteessa vakuutusturvaan. Vahingonkorvauksen merkitys on suuri, koska sehän on se turva, jonka asiakas on ostanut itselleen. Tämä väite on mielenkiintoinen, koska jokaisella niistä 30 % asiakkaista ei voi olla omakohtaista kokemusta korvauskäytännöstä. Tällöin tämä väite perustuu ympäristön, tuttavien ja mainonnan luomaan mielikuvaan vakuutusyhtiöstä. Vakuutusyhtiön joustava palvelu (12 %) sekä vakuutusten aktiivinen tarjonta (3 %) ei juurikaan näytä kiinnostavan asiakkaita, koska vain murto-osa asiakkaista painottaa niitä valinnassaan. Lisäksi suuri vakuutusyhtiö (3 %) ja näkyvä mainonta (1 %) ovat yllättäen asiakkaita kiinnostamattomia valintakriteerejä.

Taulukko 6.1: Vakuutusyhtiön valinnankriteerit (Lähivakuutus 1999).

Vakuutusyhtiön valintakriteerit asiakkaan mielestä (yleensä mainittu)	%
Vakuutusten hinnat	40
Luotettava vakuutusyhtiö	38
Edulliset vakuutusmaksut suhteessa vakuutusturvaan	32
Korvaa vahingon sattuessa nopeasti, joustavasti ja sovitulla tavalla	29
Oikeudenmukainen korvauksissa	20
Joustava palvelu	12
Vakavarainen vakuutusyhtiö	10
Henkilökunta on palvelunhaluista	10
Henkilökunta on asiantuntevaa	9
Hyvä vakuutuspalveluiden valikoima	9
Kuuntelee asiakasta ja huomioi hänen tarpeensa	8
Arvostettu vakuutusyhtiö	6
Keskittämis-, ammattijärjestö- tms. alennukset vakuutuksista	4
Aikaansa seuraava vakuutusyhtiö	4
Saan samasta yhtiöstä kaikki tarvitsemani vakuutukset	3
Suuri vakuutusyhtiö	3
Ystävien/sukulaisten suositukset tai kokemukset	3
Tiedottaa ja tarjoaa vakuutuksia aktiivisesti	3
Yhtiö on suomalainen/kotimainen	2
Näkyvä mainonta	1

Samassa tutkimuksessa tiedusteltiin myös vakuutusyhtiön vaihtamisherkkyyttä (kuva 6.1). Tutkimuksen mukaan pääasiallista vakuutusyhtiötään oli vaihtanut viimeisen 4–5 vuoden aikana lähes joka viides vastaaja (18 %). Näistä useimmin vakuutusyhtiötä vaihtoivat juuri 25–34-vuotiaat henkilöt (27 %), jolloin näihin päätöksiin ovat voineet vaikuttaa elämässä tapahtuneet muutokset (mm. opiskelu, perhe, asunto, auto, lapsi). Yllätyksenä todetaan kuitenkin, ettei sukupuolella ollut suurempaa merkitystä vaihtamistiheyteen, jolloin naiset eivät ole vakuutusasioissa paljoakaan asiakasuskollisempia kuin miehet. Ammattiryhmistä yksityisyrittäjät ovat vaihtaneet eniten vakuutusyhtiötä (35 %), mikä selittyy yrittäjien halusta etsiä edullisinta kokonaispakettia ja vakuutusyhtiöiden kiinnostuksesta yrityksiä kohtaan (lakisääteiset vakuutukset ja eläkevakuutukset). Eläkeläisten vaihtamistiheys on ollut pienin (8 %), mistä päätellään vanhempien ihmisten olevan asiakkaina tyytyväisiä ja asiakasuskollisimpia. Pääkaupunkiseudulla näyttää vakuutusyhtiön vaihtamistiheys olevan suurempi (23 %) kuin maaseudulla (17 %) tai muissa pienissä kaupungeissa. Haastattelussa selviää myös koulutuksen ja palkan vaikutus vaihtamisherkkyyteen, sillä mitä suuremmat tulot asiakkaalla on, sitä todennäköisemmin hän on vaihtanut vakuutusyhtiötään. Vakuutuksenottajista joka kymmenes (10 %) on vaihtanut yhtiötä, kun talouden vuositulot ovat alle 15 000 euroa vuodessa, mutta yli 30 000 euron vuositulojen talouksista vaihtajia on jopa joka neljäs (24 %).



Kuva 6.1: Vakuutusyhtiön vaihtamisherkkyys (Lähivakuutus 1999).

Tutkimuksessa haastatelluista asiakkaista (taulukko 6.2), jotka olivat vaihtaneet vakuutusyhtiötään muutaman viime vuoden aikana, useimmat kertoivat vaihdon syyksi uuden vakuutusyhtiön paremman tarjouksen tai paremmat vakuutusehdot (47 %). Osa vastaajista vaihtoi yhtiötä myös sen takia, että he olivat tyytymättömiä entisen vakuutusyhtiön palveluun (14 %) tai korvauksiin (10 %). Vakuutusten keskittämistäkään ei kannata väheksyä, sillä jopa 8 % piti sitä yhtenä vakuutusyhtiön vaihdon perusteena. Kuitenkin tuttavien suosituksesta oli vaihtanut vakuutusyhtiötään vain 3 % vastaajista. Tällöin tämä tutkimus tukee myös omaa näkemystäni niistä syistä, jolloin asiakas harkitsee varmasti vakuutusyhtiön vaihtamista. Nämä tärkeimmät syyt ovat vakuutuksen hinta, huono palvelu ja kielteinen korvauspäätös.

Taulukko 6.2: Syy vakuutusyhtiön vaihtamiseen (Lähivakuutus 1999).

Vaihtoi vakuutusyhtiötä, koska...	%
sain paremman tarjoukset/paremmat ehdot uudesta vakuutusyhtiöstä	47
en ollut tyytyväinen entisen vakuutusyhtiöni palveluun	14
petyin entisen vakuutusyhtiön korvauksiin	10
halusin keskittää vakuutukseni yhteen yhtiöön	8
tuttava/sukulainen suositteli	3
vakuutusyhtiö lopetti toimintansa/meni konkurssiin	3
uusi vakuutusyhtiö herätti mielenkiintoa	5
muu syy	10

6.2 Vahinkovakuutusyhtiön tietokanta

Tässä tutkielmassa tarkastellaan vakuutusyhtiön päättyneiden vakuutusten eli asiakkaan itsensä ja vakuutusyhtiön irtisanomien vakuutusten ominaisuuksia. Aineistossa on mukana kaikkia vapaaehtoisia vahinkovakuutuksia, kuten palo-, metsä-, koti-, maatila-, eläin-, kiinteistö-, yritys- ja autovakuutuksia. Kuitenkin tästä aineistosta on jätetty huomioimatta vapaaehtoiset henkilövakuutukset (sairaus- ja tapaturmavakuutukset) ja lakisääteiset vakuutukset (liikenne-, eläke- ja tapaturmavakuutukset), koska niiden erilaisuuden vuoksi niitä pitäisi mielestäni tarkastella omina ryhminään. Tämän lisäksi myös tiedoston luonnin fyysiset ongelmat estivät näiden aineistojen saamisen juuri nyt tähän tarkoitukseen. Vahinkovakuutukset ovat eri tietojärjestelmässä kuin lakisääteiset vakuutukset, jolloin näiden tietojen yhdistäminen olisi ollut hankalaa niiden erilaisten tietokantojen takia. Tällä vahinkovakuutusten esinevakuutuksiin kohdistuvalla aihealueen rajauksella haluttiin myös helpottaa aineiston ryhmittelyn lopputuloksena saatujen ryhmien tulkintaa.

Tutkielmassa aineiston ryhmittelyn edistymistä lähestytään ”top-down” mukaisesti (taulukko 6.3) aineiston ja muuttujien lukumäärää vaiheittain pienentäen. Ensimmäinen ryhmittely suoritetaan koko aineistolle, jolloin ryhmittelyn annetaan itse etsiä ryhmien lukumäärä (vähintään kaksi ryhmää). Toisessa vaiheessa pienennetään aineistoa siten, että siihen jäävät ainoastaan autovakuutukset kuten kasko-[1)], osakasko-[2)] ja palovakuutukset[3)], ja lisäksi muuttujien määrää myös pienennetään. Lisäksi samaa aineistoa tarkastellaan vielä 11 ja 5

1) Törmäys-, ilkivalta-, palo-, hirvi-, varkaus-, keskeytys-, uusarvolunastus-, autopalvelu- ja oikeusturvavakuutus (laaja).
2) Palo-, hirvi-, varkaus-, ilkivalta-, autopalvelu- ja oikeusturvavakuutus (rajoitettu).
3) Palo-, hirvi- ja oikeusturvavakuutus (suppea).

muuttujalla. Ryhmittelyohjelman annetaan aluksi etsiä ryhmien lukumäärä (vähintään kaksi), mutta myöhemmin pakotetaan aineisto ryhmittymään 3 ja 10 ryhmään. Kolmannessa vaiheessa pienennetään aineistoa vielä siten, että siihen jäävät autovakuutuksista enää vain kaskovakuutukset muuttujien lukumäärän pysyessä viidessä. Ryhmittelyn annetaan etsiä sopiva ryhmien lukumäärä (vähintään kaksi), mutta myöhemmin pakotetaan ryhmittely vielä viiteen ryhmään. Toimintasuunnitelman aineistot ja ryhmien lukumäärät on valittu alustavasti, jolloin niihin tehdään työn edistyessä tarvittavia muutoksia. Kaikkien näiden ryhmittelyjen lopputuloksena on tarkoitus löytää vakuutuksien tai asiakkaiden piirteitä eli *asiakasprofiileja*, joihin vakuutusyhtiön kannattaisi kiinnittää huomiota tulevaisuudessa. Nämä asiakasprofiilit muodostuvat useista muuttujista (mm. ikä, sukupuoli, kuntakoodi ja vakuutuslaji) jopa kymmenistä eikä ainoastaan vain muutamasta muuttujasta.

Taulukko 6.3: Toimintasuunnitelma aineiston ryhmittelyjen toteutukseen.

Ryhmittely	Aineisto lkm	**Muuttujia**	Ryhmiä	**Tavoite**
1.	3396	21	?/2 vähintään	Koko kanta
2.1	937	11	?/2 vähintään	Autovakuutukset
2.2	937	11	3	Autovakuutukset
2.3	937	5	10	Autovakuutukset
3.1	400	5	?/2 vähintään	Kaskovakuutukset
3.2	400	5	5	Kaskovakuutukset

6.2.1 Tietojen valinta

Tietojen valinta aineistoa varten on vaikeaa, sillä operatiiviseen tietokantaan on syötetty vakuutusta tehtäessä paljon asiakkaan ja vakuutuksen tietoja. Tässä vaiheessa on mahdollista, että aineistosta jää pois ryhmittelyn lopputuloksen kannalta hyvinkin tärkeää tietoa tai siihen tulee mukaan epäoleellista ja ryhmittelyä sekoittavaa tietoa. Tämän tutkielman muuttujat valittiin kirjoittajan vuosien vakuutusalan kokemuksen perusteella mahdollisimman laaja-alaisesti edustaen vakuutuksen sekä asiakkaan tietoja (liite 1 ja liite 2).

Seuraavana tarkastellaan muuttujien valinnan perusteita niiden tärkeyksien ja mahdollisten yhteyksien mukaan (taulukko 6.4). Vakuutuksen tiedoista *vakuutuslaji* (muuttuja 1) kertoo päättyneen vakuutuksen lajin tunnuksen, joista esimerkkinä 225 arvo tarkoittaa moottoriajoneuvon palovakuutusta (suppea turvataso). Jollakin vakuutuslajilla voi olla suurempi irtisanomisprosentti kuin toisella lajilla, joten vakuutusten irtisanominen voi johtua vakuutuksen hinnasta, tuotteesta tai sen vakuutusehdoista. Kuitenkin nykyisin vakuutusyhtiöt

pyrkivät saamaan asiakkaan kaikki vakuutukset samaan yhtiöön, jolloin päättyneiden vakuutuslajien jakauma voi olla hyvinkin tasainen eri vakuutuslajien kesken (esim. auto-, koti-, tapaturma-, ja koiravakuutus). Vakuutuksen *kohdetyyppistä* (2) saattaa löytyä jotakin yhtäläisyyttä vakuutuslajin kanssa. Kohdetyyppi on yhteinen nimittäjä samanlaisille vakuutuksille, jolloin esimerkiksi autovakuutuksia on kolmenlaisia ja niitä kutsutaan yleisesti vain autovakuutuksiksi.

Taulukko 6.4: Aineiston muuttujat, arvoalueet, selitteet sekä niiden käyttö ryhmittelyissä.

	Muuttuja	Arvoalue	Selite	1	2.1	2.2	2.3	3
1.	Vakuutuslaji	12-225	Vakuutuksen vakuutuslajin tunnus	X	X	X		
2.	Kohdetyyppi	1-5	Vakuutuksen kohdetyyppi	X				
3.	Asiakkaan syntymävuosi	0-83	Vakuutuksenottajan syntymävuosi	X	X	X	X	X
4.	Kuntakoodi	0-927	Vakuutuksenottajan asuinkunnan kuntakoodi.	X	X	X		
5.	Sukupuoli	1-5	Vakuutuksenottajan sukupuoli	X	X	X	X	X
6.	Asiakasrooli	0-8	Vakuutuksenottajan asiakasrooli	X				
7.	Suoraveloitus	0-1	Vakuutuksen maksutapana suoraveloitus	X				
8.	Voimassa oloaika	0-34	Vakuutuksen voimassaoloaika vuosina	X	X	X	X	X
9.	Omavastuu	0-1674	Vakuutuksen omavastuu vahingoissa	X				
10.	Alennusprosentti	0-71	Vakuutuksen alennusprosentti	X				
11.	Indeksityyppi	0-4	Vakuutuksen indeksityyppi	X				
12.	Vakuutuksen päättymissyy	0-9	Vakuutuksen päättymissyy	X	X	X		
13.	Asiamiestunnus	0-40000	Vakuutuksen tehneen asiamiehen tunnus	X				
14.	Vahinkokorvaukset	0-22004	Vakuutuksen voimassaoloaikana suoritetut korvaukset	X	X	X		
15.	Vanha yhtiö	0-99	Mistä yhtiöstä vakuutus on siirtynyt?	X	X	X	X	X
16.	Vakuutusmäärä euroina	0-2129784	Vakuutuksen kohteiden yhteenlaskettu vakuutusmäärä	X				
17.	Vuosimaksu euroina	0-3321	Vakuutuksen vakuutusmaksu	X	X	X		
18.	Viivästyskorko	0-12	Vakuutuksen maksuun sisältynyt viivästyskorko	X				
19.	Laskun tila	0-13	Vakuutuksen laskun tila	X	X	X		
20.	Alennus vuosimaksusta	0-2635	Vakuutuksen maksun alennus	X				
21.	Uusi yhtiö	0-99	Mihin yhtiöön asiakas on siirtynyt?	X	X	X	X	X

Vakuutuksenottajan tiedoista *asiakkaan syntymävuoden* (3) perusteella pyritään saamaan selville vakuutuksenottajan iän vaikutus vakuutusten päättymiseen. Aikaisemmin esitellyn tutkimuksen mukaan (kuva 6.1) asiakkaan iällä on merkittävä vaikutus vakuutusten vaihtamiseen, koska vanhemmat asiakkaat ovat todennäköisesti asiakasuskollisempia kuin nuoremmat asiakkaat. Vakuutuksenottajan ikääntymisen vaikutus *vahinkokorvauksiin* (14) sekä myös vakuutuksen *alennusprosenttiin* (10) voi myös olla mielenkiintoinen tarkastelukohde. Asiakkaan asuinkunnan *kuntakoodin* (4) perusteella voidaan päätellä, mitä vaikutusta asuinpaikkakunnalla on irtisanomistiheyteen. Ovatko asiakkaat kaupungeissa vai maaseudulla uskollisempia? Tällöin yrityksen oman ja kilpailijoiden palveluverkoston tiheys

saattaa korostua, mutta tämä ominaisuus kuitenkin sekoittuu myyntiin ja palveluun sekä puhelimen että Internetin kautta.

Vakuutuksenottajan s*ukupuoli* (5) voi olla hyvin yllätyksellinen naisten osalta, sillä yleensähän perheen vakuutukset ovat miehen nimissä. Näkemykseni mukaan naisten rooli vakuutuksen vaihtamistilanteessa on kuitenkin merkittävä, koska perheet tekevät usein yhteiset päätökset ja tällöin naisten merkitys korostuu enemmän. Mutta valitettavasti sitä ei tässä aineistossa pystytä näyttämään toteen. *Asiakasrooli* (6) on suurimmaksi osaksi yksityinen, ja mielenkiintoista on tarkastella sen yhteyttä sukupuoleen, kohdelajiin sekä vakuutuslajiin. *Suoraveloitus* (7) on yleistynyt maksutapana, jolloin näiden asiakkaiden yhteys *viivästyskorkoihin* (18) voi olla mielenkiintoinen. Suoraveloitusvaltuutuksen antanut asiakas saattaa hoitaa maksunsa ajallaan ja on vakuutusyhtiölle tämän kriteerin perusteella hyvä asiakas.

Vakuutuksen tiedoissa vakuutuksen *voimassaoloaika vuosina* (8) ei kerro aikaisempaa asiakashistoriaa vaan ainoastaan tämän kyseisen vakuutuksen voimassaoloajan. Lyhytaikaiset vakuutukset kertovat ehkä vakuutuksenottajien hakevan edullisinta vakuutusta, mutta pitkäaikaisten asiakkaiden vakuutusten päättyminen kertoo jostakin muusta ongelmasta. Kuitenkin vakuutusten päättymisen taustalla voi olla myös vakuutuksenottajan omassa elämässä tapahtuneet muutokset, joista vakuutuksenantajalla ei ole aina etukäteen mahdollisuutta saada tietoa (mm. perheen perustaminen, avioero tai kuolema). Mielenkiintoista on myös tarkastella vakuutuksen voimassaoloajan vaikutusta vanhempien tai nuorempien asiakkaiden irtisanomistiheyteen.

Vakuutuksen *omavastuu* (9) vaihtelee eri vakuutuslajeissa. Kuitenkin mitä suurempi omavastuu on, sitä pienempi on vakuutusmaksu. Asiakkailla, joilla on suurempi omavastuu, on vähemmän korvattavia vahinkoja, koska osa näistä vahingoista jää vakuutuksenottajan omalle vastuulle. Näkemykseni mukaan omavastuun suuruudella ei vakuutusta tehdessä ole kovinkaan paljon merkitystä, sillä asiakkaan mielestä vahinkohan sattuu sitten joskus tulevaisuudessa tai jos se nyt sattuu olleenkaan.

Vakuutuksen *alennusprosentti* (10) kertoo vakuutuksen ylimääräisen alennusprosentin. Tämä ei kerro koko totuutta, sillä vakuutuksen eri kohteille (mm. rakennus, irtaimisto) on jo annettu erilaisia alennuksia (laajuus-, bonus-, ikä-, turvalukko-, pinta-ala- ja vuotovahinkoalennus).

Suuret ylimääräiset alennusprosentit kertovat kovasta kilpailusta tai tuotteen alihinnoittelusta, jolloin markkinat aina määräävät tuotteen lopullisen hintatason. Alennusprosentit voivat vaihdella eri vakuutuslajeissa ja keskittämällä vakuutukset samaan yhtiöön asiakas voi saada suuremmat alennusprosentit kuin pelkästään yhden vakuutuksen perusteella.

Vakuutuksen *indeksityypin* (11) merkitystä voidaan tarkastella ainoastaan vakuutusmäärien osalta. Mikäli päättyneissä vakuutuksissa on paljon sellaisia vakuutuksia, joita ei ole sidottu indeksiin, todennäköistä on, että näiden kohteiden vakuutusmäärät ovat jääneet jälkeen rahanarvon muutoksen (hintojen nousu ja inflaatio) seurauksesta. 1980-luvulla inflaation vaikutus oli vuosittain hyvinkin suuri (jopa 8 %), jolloin monet saattoivat poistaa vakuutuksen indeksistä. Vakuutuksenottajien mielestä vakuutusmäärän noustessa vakuutusmaksu nousi aina samassa suhteessa ja jopa hieman vielä enemmän ottaen huomioon vakuutusyhtiön tariffikorotukset, jolloin vakuutusmäärät saattoivat nousta yli senhetkisen vakuutettavan omaisuuden arvon (ylivakuutus). Tämä tulee esille niissä vanhemmissa vakuutuksissa, jotka ovat sidottuja elinkustannusindeksiin. Vastaavasti nykyiset täysarvovakuutukset sidotaan hinnan muutosten osalta rakennuskustannusindeksiin.

Vakuutuksen *päättymissyy* (12) voi olla moniperusteinen, mutta lähtökohtana vakuutusyhtiö tai asiakas voi irtisanoa vakuutuksen. Mikäli vakuutus on päättynyt uutta vakuutusta vastaan, silloin vakuutukset on uudistettu nykyisen tilanteen mukaiseksi. Mikäli asiakas on siirtynyt toiseen yhtiöön, silloin uuden yhtiön tunnus ja päättynyt vakuutuslaji voivat kertoa jotakin irtisanomisliikenteestä. *Asiamiestunnus* (13) kertoo päättyneen vakuutuksen tehneen asiamiehen tunnuksen. Joillakin asiamiehillä saattaa olla enemmän päättyneitä vakuutuksia kuin toisella, mikä voi johtua asiamiehen toimintatavasta tai vähäisestä kiinnostuksesta vanhoja asiakkaita kohtaan. Vakuutuksen voimassaoloaikana *maksetut korvaukset* (14) voivat kertovat korvauksien vaikutuksen vakuutusten päättymiseen. Yleensä vahinkotilanteissa kokemukseni mukaan tulee melkein ensimmäisenä esille toteama: ”Vakuutusta on maksettu kymmenen vuotta ja aikaisemmin ei ole korvausta haettu. Mikäli nyt tätä vahinkoa ei korvata kokonaan, niin kyllä vakuutukset vaihtuu toiseen yhtiöön”. Tällöin maksetut korvaukset ovat erittäin kiinnostava tarkastelukohde. Vastaavasti hylätyistä korvauspäätöksistä ei ole merkintää aineistossa, mutta kuitenkin joku asiakas on todennäköisesti irtisanonut vakuutuksensa juuri hylätyn korvauspäätöksen seurauksesta.

Asiakkaan *entisen yhtiön tunnus* (15) kertoo asiakkaan aikaisemman vakuutusyhtiön. Tämän perusteella voidaan havaita, erottuuko jostakin muusta yhtiöstä suurempi positiivinen asiakasvirta tai palaako asiakas takaisin entiseen yhtiöön. Vakuutettavien kohteiden v*akuutusmäärät* (16) ovat saattaneet jäädä jälkeen todellisesta arvostaan, mikäli vakuutusta ei ole sidottu indeksiin. Onko tällaisissa vakuutuksissa suurempi asiakaspoistuma? Vakuutuksen v*uosimaksun* (17) suuruudesta voidaan päätellä, onko kyseessä vakuutusmaksultaan suuri, normaali vai pieni asiakas. Tällä asiakasprofiililla voi myös olla vaikutusta vakuutuksen alennusprosenttiin. V*iivästyskorko* (18) kertoo asiakkaan vakuutusmaksun maksutavasta, koska maksujen viivästyminen on aina huolestuttava tekijä. *Laskun tilasta* (19) on havaittavissa maksetun laskun tila. Laskun tilasta ja viivästyskoron määrästä voi löytyä yhdistäviä tekijöitä tarkemmassa tarkastelussa. Usein laskut on maksettu normaalisti, mutta laskulla voi olla merkintöjä avoin, ulosmittaus tai muistutus. *Alennus vuosimaksusta* (20) kertoo saman minkä, alennusprosentti kertoi jo aikaisemmin, mutta nyt se on muutettu euroiksi. *Siirtyneen asiakkaan uudesta yhtiöstä (21)* voidaan havaita, erottuuko jokin yhtiö suurempana hyötyjänä kuin toiset yhtiöt kaikissa vakuutuslajeissa vai jossakin tietyssä lajissa.

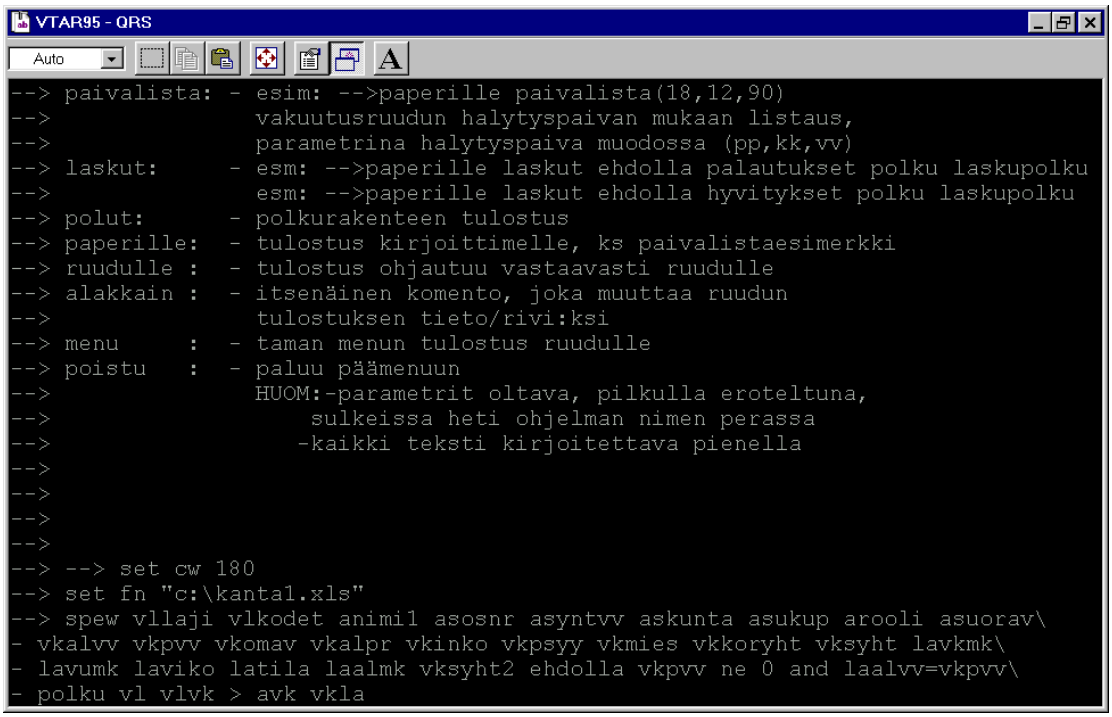
6.2.2 Tietokannan poiminta

Aineisto poimittiin operatiivisesta tietokannasta *QRS-kyselykielen* (*Query Retrival System*) avulla omaksi tiedostoksi. QRS on tietokannan tarkasteluun tarkoitettu yksilauseinen kyselykieli, jota käytetään MDBS IV -tietokantojen yhteydessä. Kyselykielen avulla kannasta voidaan tulostaa raportteja ja tilastoja näytölle, tiedostoon tai kirjoittimelle. Aineistoon poimittiin kannasta ne vakuutukset, joiden päättymisvuosi oli muu kuin nolla (päättymisvuosi on voimassa olevilla vakuutuksilla nolla) ja joiden laskutuskauden vuosi oli sama kuin päättymisvuosi (viimeisen laskun tiedot). Aineistoa poimittaessa edettiin vakuutuksen tietojen kautta viimeisen laskun tietoihin, joista poimittiin valittujen muuttujien tiedot tiedostoon (kuva 6.2).

```
set cw 180                    {tietueen pituus}
set fn "c:\kanta1.xls"        {tiedoston nimi}
spew vllaji vlkodet animi1 asosnr asyntvv askunta asukup arooli asuorav\
vkalvv vkpvv vkomav vkalpr vkinko vkpsyy vkmies vkkoryht vksyht lavkmk\
lavumk laviko latila laalmk vksyht2 ehdolla vkpvv ne 0 and laalvv=vkpvv\
polku vl vlvk > avk vkla
```

Kuva 6.2: QRS-kysely operatiivisesta tietokannasta.

Testikannan poiminta tehtiin kyselyikkunassa (kuva 6.3), jonka poiminnan lopputuloksena saatiin Excel-taulukko. Ensimmäinen ongelma esiintyi jo tietokannan poiminnan yhteydessä, sillä monien yritysten jälkeen käytetty kysely keskeytyi aina heti alkuunsa. Tähän selvisi ratkaisuna, että QRS-kyselyn pituus sai olla maksimissaan 255 merkkiä ja rivillään voi olla 128 merkkiä (MS-DOS-ohjelma). Tämän johdosta kysely pilkottiin kolmeen osaan muodostaen jokaisella kerralla osatiedosto, jotka lopuksi yhdistettiin yhdeksi kokonaisuudeksi. Varmistukseksi jokaisen tiedoston ensimmäisenä muuttujina olivat asiakkaan nimi ja henkilötunnus, joten kolmen tiedoston yhdistämisessä voitiin varmistua, että jokaisella rivillä oli varmasti saman asiakkaan tiedot.



Kuva 6.3: QRS-kyselykielen käyttöliittymä.

6.2.3 Tietokannan valmistelu louhintaa varten

Yleisesti luullaan, että niillä, joilla on saatavilla paljon erilaista tietoa, on se myös helposti käytettävissä. Valitettavasti tämä ei kyllä aina pidä paikkaansa. Tiedot voivat olla tallennettuna, mutta ne voivat olla väärässä muodossa tai paikassa tai ne on jätetty kokonaan tallentamatta epäolennaisena tai työntekoa hidastavana asiana. Lisäksi tietojen tallentamisessa eri työntekijät käyttävät erilaisia toimintatapoja samoissa tilanteissa, ja tähän arkipäivän ongelmaan törmättiin myös tämänkin aineiston käsittelyssä.

Kolmen tiedoston yhdistämisen jälkeen aineistoa käsiteltiin Excel-taulukkolaskentaohjelman avulla käsin sekä taulukkolaskentaohjelman muuntotyökaluin. Aineistossa (taulukot 6.5, 6.6 ja 6.7) havaittiin olevan paljon puuttuvia ja jopa virheellisiä arvoja. Pääsääntöisesti puuttuvat arvot olivat suurin ongelma ja samoin myös arvoalueen ulkopuoliset arvot, joita ei olisi pitänyt olla kanta-aineistossa. Alkuvaiheessa vakuutuksenottajan henkilötunnus oli asiakkaan syntymävuoden tarkistamista varten, mutta se poistettiin pois lopullisesta aineistosta.

Ensimmäiseksi havaittiin aineistossa olevan paljon yritysasiakkaita, joilla oli henkilötunnuksen paikalla y-tunnus. Mikäli yritysvakuutukset olisivat mukana, vakuutuksen-ottajan ikää ei olisi mahdollista määrittää yritysten yhteydessä. Tässä tapauksessa huomattiin, että yritysasiakkaista olisi pitänyt olla enemmän tietoa talletettuna (mm. yrityksen historia, luottotiedot, omavaraisuus, liikevaihto, omistajat) ryhmittelyä varten kuin mitä nyt oli. Usein osa tästä tiedosta on selvillä vakuutuksen tekovaiheessa, mutta sille kaikelle ei ole paikkaa järjestelmässä eikä kaikkea sitä tietoa tarvita enää vakuutuksen myöntämispäätöksen jälkeen. Näiden tietojen puuttumisen takia aineistosta poistettiin kaikki yritysvakuutukset (oy, ky ja tmi) sekä lisäksi myös kiinteistövakuutukset (as oy ja kiint. oy). Tämän välttämättömän rajauksen johdosta alkuperäinen 3527 vakuutuksen aineisto supistui nyt lopulliseen 3396 vakuutuksen aineistoksi.

Aineiston sisällön tarkemmassa tarkastelussa havaittiin ensimmäiseksi, että siellä oli paljon puuttuvia sekä myös virheellisiä tietoja (taulukko 6.5). Henkilötunnus puuttui monilta vakuutuksenottajilta, mutta sillä ei ollut merkitystä, mikäli asiakkaan syntymävuosi oli tiedossa. Mikäli syntymävuosi puuttui ja asiakkaan henkilötunnus oli käytettävissä, syntymäaika saatiin selville henkilötunnuksesta poimimalla. Näiden molempien tietojen puuttuessa tiedot jouduttiin etsimään jostakin muualta ja täydentämään se käsin. Samoin myös

sukupuolen osalta puuttui arvoja, jolloin nämä puuttuvat arvot käytiin läpi manuaalisesti. Aineistossa oli paljon merkkijonomuuttujia (sukupuoli, asiakkaan rooli, suoraveloitus), jotka muutettiin taulukkolaskentaohjelman korjaustyökaluin numeromuuttujiksi. Esimerkiksi sukupuolimuuttujalle annettiin arvoksi miehelle yksi (M=1) ja naiselle nolla (N=0). Tämän lisäksi monen muuttujan arvot (mm. asiakkaan rooli ja suoraveloitus) sisälsivät numeroita tai kirjaimia, jolloin nämä kirjaimet muutettiin vastaamaan numeroita.

Taulukko 6.5: Otos päättyneiden vakuutuksien korjaamattomasta aineistosta (kentät 1–9). (Nimet, henkilö- ja Y-tunnukset on muutettu.)

VLLAJ (vakuutuslaji)	**VLKOHDET** (vakuutus-kohde)	**ANIMI1** (asiakkaan nimi)	**ASOSNR** (Henkilö- tai Y-tunnus)	**ASYNTVV** (asiakkaan syntymävuosi)	**AKUNTA** (asuinkunta)	**ASUKUP** (sukupuoli)	**AROOLI** (asiakkaan rooli)	**ASUORAV** (suoraveloitus)
12	1	KARHUNEN EINO	010160-233W	60	564	M	Y	
12	1	TAHVANAINEN JARI	010137-141S	37	146	M	Y	X
13	1	HILTUNEN MARI KA	010114-5702	0	146	N	Y	
13	1	HASSINEN KIMMO		0	146		Y	
42	1	AS OY RIVITALO	7306553	0	146		F	
205	4	KETTURI VÄINÖ	010167-023D	67	146	M	Y	
225	4	PALTTURI TIMO	010158-0254	58	146	M	Y	
225	4	PENTTILÄ OIVA	010143-022J	43	146	M	Y	

Vakuutuksen voimassaoloaika laskettiin vakuutuksen päättymisvuoden ja alkamisvuoden erotuksena (taulukko 6.6). Tietokannassa tämä vuosiluku oli ilmoitettu kahdella desimaalilla (vuoden 2000 ongelma korjattu ohjelmallisesti). Kahden desimaalin merkintä 99 tarkoitti vuotta 1999 ja 01 tarkoitti vuotta 2001. Nyt vähennyslaskuna toteutettu laskelma ei toiminut täydellisesti vaan aiheutti ongelmia, koska kun 0 vuodesta (2000) vähennettiin 90 (1990), lopputulos oli –90 vuotta, kun sen olisi pitänyt olla 10 vuotta. Tähän ongelmaan jouduttiin rakentamaan laskentakaava. Lopuksi poistettiin vakuutuksen alkamisvuoden ja päättymisvuoden kentät tarpeettomina. Lisäksi puuttuvat arvot kentissä tuottivat edelleen ongelmia ja veivät aikaa.

Aineiston laadinnassa käytettiin päättymisvuotena vuotta 2001, jolloin käsiteltävänä rahayksikkönä oli vielä silloinen markka. Tämän johdosta muutettiin rahayksiköksi euro ja pyöristettiin arvot täysille euroille (taulukko 6.6). Tämä muunnos tehtiin vakuutuksen omavastuuseen, alennukseen ja maksettuihin korvauksiin. Lisäksi havaittiin joissakin vakuutuslajeissa ongelmia omavastuun kanssa, sillä poiminnan yhteydessä omavastuu oli muuttunut nollaksi. Kuitenkin vain muutamissa vakuutuslajeissa voi omavastuu olla nolla, mutta yleensä vakuutuslajeilla on jokin tätä suurempi omavastuu. Tämä korjaus tehtiin käsin käymällä nämä vakuutuslajit läpi. Lisäksi ongelmia tuli myös alennusprosenttien yhteydessä,

sillä muutamat poiminnan yhteydessä desimaaleja sisältäneet alennusprosentit muuttuivat päiväyksen muotoon ja nämä virheellisyydet korjattiin käsin muokkaamalla.

Taulukko 6.6: Otos päättyneiden vakuutuksien korjaamattomasta aineistosta (kentät 10–19). (Asiamiestunnukset on muutettu.)

VKALVV (vakuutuksen alkamisvuosi)	**VKPVV** (vakuutuksen päättymis-vuosi)	**VAKIKA** (vakuutuksen voimassaolo aika vuosina)	**VKOMAV** (vakuutuksen omavastuu mk)	**VKALPR** (vakuutuksen alennuspro-sentti)	**VKINKO** (indeksityyppi)	**VKPSYY** (vakuutuksen päättymissyy)	**VKMIES** (asiamies-tunnus)	**VKKORYHT** (vakuutuksesta maksetut kor-vaukset mk)	**VKSYHT** (yhtiökoodi siirtyneelle vakuutukselle)
95	01	=-94 !!!!!!!!!	600	0	0		32100000	0	
90	0	=-90 !!!!!!!!!	600	0	1		32100000	3456	
82	99	=17	600	1,5,1950	1		32100011	0	
79	99	=20	600	0	1	1	32100000	0	
0	0	=0 !!!!!!!!!	1000	47,15	2		32150014	0	36
98	01	=-97 !!!!!!!!!	0	29,1			32140000	0	
94	99	=5	750	0	0		32100000	0	55
1	0	=-1 !!!!!!!!!	0	0			32120000	0	

Viimeisestä osatiedostosta muutettiin markkamäärät euroiksi (taulukko 6.7), jolloin tämä muutos tehtiin vakuutuksen vakuutusmäärään, vuosimaksuun, viivästyskorkoon ja vuosimaksun alennukseen. Laskun tila -muuttujassa oli merkkejä, jotka korvattiin numeroarvoilla.

Taulukko 6.7: Korjaamaton aineisto päättyneistä vakuutuksista (kentät 20–25).

LAVKMK (vakuutusmäärä mk)	**LAVUMK** (vuosimaksu mk)	**LAVIKO** (viivästyskorko mk)	**LATILA** (laskun tila)	**LAALMK** (alennus mk vuosimaksusta)	**VKSYHT2** (yhtiökoodi päät-tyneet vakuu-tukset)
93000	351	0	1	0	
139728	463	0	1	40	
98493	388	10	1	33	
105154	428	0	f	0	
0	1366	0	3	1505	
0	2050	3	3	1970	
0	135	0	1	23	
0	128	0	3	10	

Kun kaikki edellä mainitut muutokset ja korjaukset oli tehty, tiedostot yhdistettiin yhdeksi tiedostoksi. Tällöin havaitsin aineiston valmisteluun kuluneen ennakoitua enemmän aikaa. Arvioin etukäteen käyttäväni aikaa tiedon valmisteluun (haku ja puhdistus) noin 50 tuntia, mutta todellisuudessa käytin siihen lähes 200 tuntia, vaikka tiedot olivat operatiivisesta toimivasta järjestelmästä peräisin.

Seuraavaksi aineiston käsittelyä jatkettiin normalisoinnilla ja ryhmittelyllä, jolloin törmättiin taas uusiin ongelmiin. Aineistossa havaittiin olevan vielä puutteita (puuttuvia arvoja,

virheellisyyksiä), sekä konvertointiohjelmassa epäiltiin olevan myöskin jokin ongelma. Virheet sotkivat lopputuloksen siten, että rivit ja sarakkeet menivät sekaisin. Tarkemmassa tarkastelussa huomattiin, että virheen aiheutti taulukkolaskentaohjelman tiedoston muuttaminen lopuksi tekstitiedostoksi (ASCII), jolloin siinä muodossa siihen erehdyttiin tekemään vielä muutamia viime hetken korjauksia (käytetty välimerkkeinä tabulaattoria ja välilyöntiä) ennen normalisointia. Normalisoinnin jälkeen tiedot muutettiin takaisin taulukkolaskentaohjelman muotoon, jolloin rivit ja sarakkeet menivät sekaisin. Tämän seurauksesta aineisto jouduttiin vielä kerran käymään läpi manuaalisesti etsien kaikki tyhjät ja virheelliset arvot, joita löytyi muutamia kymmeniä kappaleita. Lisäksi tässä vaiheessa epäiltiin, että asiamiestunnuksen lukuarvo (esim. 32110000) on liian suuri konvertointiohjelmalle ylivuodon johdosta. Tämän seurauksesta asiamiehen tunnuslukua pienennettiin poistamalla kolme ensimmäistä lukuarvoa (esim. 10000), jolloin tällä ei ollut merkitystä asiamiehen myöhempää tunnistamista ajatellen.

6.2.4 Muuttujien normalisointi ja skaalaus

Tietokannan korjausten jälkeen jatkettiin aineiston tarkastelua ennen ryhmittelyä Excel-taulukkolaskentaohjelmassa, jossa laskettiin kunkin muuttujan minimi-, maksimi- ja keskiarvot sekä keskihajonnat. Tällöin tarkistettiin vielä kerran minimi- ja maksimiarvojen avulla, oliko muuttujissa vielä oletettujen arvoalueiden ulkopuolisia arvoja sekä korjattiin mahdolliset virheellisyydet.

Tämän jälkeen suoritettiin muuttujien normalisointi Excel-taulukkolaskentaohjelmassa, jolloin jokainen muuttuja muutettiin pienemmälle arvoalueelle *z-piste normalisoinnin* laskentakaavaa (kaava 5) käyttäen. Normalisointi tehtiin kullekin muuttujalle erikseen, jolloin suurin arvoista sai maksimiarvon ja pienin minimiarvon. Muut arvot asettuivat näiden arvojen väliin säilyttäen kuitenkin alkuperäiset suhteelliset eronsa. Tämän jälkeen jokaisen muuttujasarakkeen keskiarvo oli nolla ja keskihajonta oli yksi. Lopuksi skaalattiin muuttujien arvot Min-Max-normalisointia käyttäen (kaava 4) koko aineiston osalta. Muuttujien arvojen ollessa desimaalilukuja erittäin pienellä alueella (0–1), jouduttiin aineisto muuttamaan kokonaisluvuiksi isommalla arvoalueelle. Tämä tehtiin kertomalla skaalauksen lopputuloksena saadut arvot miljoonalla.

6.2.5 Ryhmittelyn toteutus

Normalisoinnin jälkeen aineisto tallennettiin takaisin ASCII-muotoon, jonka jälkeen aineisto muutettiin *TS-tiedostoksi* (*Training Set*) RAW2CB-ohjelmalla. Tämän jälkeen aineisto oli valmiina tiedonlouhintaa varten ja louhinta toteutettiin Professori Pasi Fräntin kehittämällä RLS-ryhmittelyanalyysiohjelmistolla. Tämä ohjelma ryhmitteli aineiston ryhmiin, joko käyttäjän antaman ryhmien lukumäärän mukaisesti tai päätellen ryhmien lukumäärän täysin itsenäisesti. Lopputulokset (liite 4) kirjoitettiin koodikirjaan käyttäjän määräämään *CB-tiedostoon* (*Codebook*). Tiedostossa jokainen vakuutus on indeksoitu ryhmittelyn mukaisiin ryhmiin. Lopuksi tämä tiedosto yhdistettiin takaisin alkuperäiseen aineistoon Excelissä, jonka jälkeen jatkettiin aineiston ryhmien tarkastelua taulukkolaskentaohjelmassa korrelaatioanalyysillä.

6.3 Aineiston tutkiskelu

Jokaisen uuden ryhmiteltävän aineiston aluksi tarkasteltiin koko aineiston sisältöä ennen ryhmittelyä ja vastaavasti ryhmittelyn jälkeen tarkasteltiin lopputuloksena saatujen ryhmien sisältämien vakuutusten piirteitä. Aineiston tarkastelussa käytettiin apuna Excel-taulukkolaskentaohjelmassa olevaa *korrelaatiofunktiota*. Tämä analyysityökalu mittaa kahden tietosarjan välistä riippuvuutta, jotka on skaalattu mittayksiköstä riippumattomiksi. Lopputuloksena korrelaation laskenta palauttaa kahden tietosarjan solualueiden korrelaatiokertoimen. Korrelaatiotyökalulla voidaan määrittää, kuinka kaksi tietoaluetta liittyvät yhteen eli liittyvätkö toisen sarjan suuret arvot toisen sarjan suuriin arvoihin (positiivinen korrelaatio), liittyvätkö toisen sarjan pienet arvot toisen sarjan suuriin arvoihin (negatiivinen korrelaatio) vai liittyvätkö kummankaan sarjan arvot toisiinsa (korrelaatio lähellä nollaa). Korrelaatio on sitä suurempi mitä lähempänä korrelaatiokerroin on 1:tä tai –1:tä, kun taas sen ollessa lähellä nollaa tai nolla korrelaatiota ei esiinny joukkojen välillä.

Korrelaatiofunktion käyttämien solujen sisällön tulee olla lukuja tai lukuja sisältäviä nimiä, matriiseja tai viittauksia. Funktio ohittaa kaikki matriisien tai viittausten argumentit, jotka sisältävät tekstiä, totuusarvoja tai tyhjiä soluja. Kuitenkin nollan sisältävät solut lasketaan mukaan. Jossakin tapauksessa matriisien sisältämien arvojen keskihajonta voi nolla, jolloin funktio palauttaa virhearvon (#JAKO/0).

Keskihajonta (Sd) on tärkein ja käytetyin hajonnan mitta. Keskihajonta kuvaa havaintoarvojen keskimääräistä etäisyyttä keskiarvosta. Keskihajonnan neliötä kutsutaan *varianssiksi*. Mitä pienempiä ovat keskihajonnat ja varianssi, sitä tiiviimmin havaintoaineisto on keskittynyt keskiarvon ympärille. Keskihajonnaksi saatu tulos tarkoittaa, että yksittäiset havaintoarvot sijaitsevat keskimääräisesti tämän yksikön päässä havaintoarvojen keskiarvosta (Av).

6.4 Koko aineiston ryhmittely

Alkuperäinen aineisto sisälsi 3396 päättyneen vakuutuksen tiedot, ja jokaisesta vakuutuksesta oli 21 muuttujan verran tietoa vakuutuksenottajasta ja vakuutuksesta. Taulukosta 6.8 havaitaan ennen ryhmittelyä tämän koko aineiston minimi-, maksimi- ja keskiarvot sekä keskihajonnan arvot. Suurimmassa osassa skaalaamattomien muuttujien minimiarvot olivat nollia ja maksimiarvot olivat laajalla arvoalueella. Pienin keskihajonta löytyi kohdetyypistä (muuttuja 2), sukupuolesta (5), asiakasroolista (6), suoraveloitusaineistosta (7), indeksityypistä (11) ja viivästyskorosta euroina (18), jolloin näiden muuttujien arvot olivat ryhmittyneet tiiviisti keskiarvon ympärille. Aineistossa havaittiin olevan paljon palokohteita, joiden vakuutuksenottajana oli keskiarvon mukaan 62-vuotias yksityinen mieshenkilö. Vakuutukset olivat olleet voimassa keskimäärin 5 vuotta ja niiden omavastuu oli 90 euroa. Lisäksi keskiarvon mukaan jokaiselle vakuutukselle oli suoritettu korvauksia keskimäärin vain 49 euroa ja vastaavasti vakuutusmaksu oli ollut 194 euroa. Asiakkaat eivät yleensä ole käyttäneet suoraveloitusta maksutapana, mutta maksut oli kuitenkin maksettu ajallaan ilman viivästyskorkoja.

Taulukko 6.8: Koko aineiston muuttujien minimi-, maksimi- ja keskiarvot (Av) sekä keskihajonta-arvot (Sd) ennen ryhmittelyä.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Min	12	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Max	225	5	83	927	5	8	4	100	1674	71	8	20000	40000	22004	99	2129783	3321	12	537	2635	99
Av	88	2	39	278	2	2	0	5	90	1	1	11	3964	49	5	34126	194	0	7	85	2
Sd	67,86	1,39	24,63	269,44	1,09	1,93	0,42	4,98	96,91	3,37	0,54	383,62	9628,38	737,19	14,13	93545,26	248,21	0,45	12,86	204,22	10,47

Aineistoa tarkasteltiin aluksi korrelaatioanalyysilla, jossa tutkittiin kunkin muuttujan välistä riippuvuutta toisiin sarjoihin. Aluksi ensimmäisen muuttujan arvoja verrattiin kaikkiin muihin muuttujien arvoihin laskemalla korrelaatiokertoimet, ja tätä jatkettiin niin kauan kunnes jokaiselle muuttujaparille oli laskettu korrelaatiokertoimet. Lopputuloksena saatiin kaksiulotteinen matriisitaulukko (liite 3), jossa vaakarivin samat korrelaatioarvot ovat myös samalla numerolla olevalla pystyrivillä. Kertoimen ollessa tasan yksi oli vertailu tapahtunut samaan matriisiin. Taulukoissa on esitetty ainoastaan itseisarvoltaan 0,20 ylittävät korrelaatioarvot, koska tätä pienemmät luvut edustavat kovin vähäistä tai olematonta korrelaatiota. Tämä esitystapa myös selkeyttää taulukon esitystä ja helpottaa sen tulkitsemista.

Ryhmittelemättömässä tietokannassa havaittiin olevan muutamia arvoja, joiden korrelaatioarvo oli itseisarvoltaan suurempi kuin 0,5, jolloin näitä korrelaatioita voitiin pitää jo merkittävänä. Lajin (muuttuja 1) ja kohdetyypin (2) korrelaatio oli 0,49. Tämän lisäksi kohdetyyppi (2) korreloi –0,68 indeksityypin (11) kanssa. Ensimmäiseen korrelaatioon löytyi perusteeksi se, että samalla vakuutuslajilla oli aina tietty sama kohdetyyppi (palovakuutus = palokohde). Jälkimmäiseen korrelaatioon ei löydy muuta perustetta, kuin näiden arvoalueet olivat lähes identtiset. Asiakkaan syntymävuosi (3) korreloi –0,58 sukupuolen (5) sekä indeksityypin (11) kanssa –0,65. Syntymävuoden vaikutus sukupuoleen voi olla merkityksellinen, sillä vanhemmat asiakkaat olivat useammin miehiä (vakuutukset miehen nimissä). Vastaavasti sukupuolen korrelaatiolle indeksityypin (11) kanssa sekä vakuutuksen päättymissyyn (12) korrelaatiolle 0,65 laskuntilan (19) kanssa ei löydy selitystä. Näiden lisäksi voitiin todeta koko aineistossa esiintyvän lievää korrelaatiota (yli 0,3) useiden muuttujien kesken, mutta näiden korrelaatioiden perusteella ei vielä voitu tehdä kovinkaan merkittäviä johtopäätöksiä. Näiden havaintojen perusteella aineistosta ei vielä ennen ryhmittelyä pystytä huomaamaan selviä ryhmiä tai muita yhtäläisyyksiä.

6.4.1 Aineiston ryhmittely kahteen ryhmään

Ensimmäisessä ryhmittelyssä aineiston annettiin vapaasti ryhmittyä, jolloin aineisto jakaantui kahteen eri ryhmään. Ryhmittelyn lopputuloksena saatu tulostiedosto osoitti (liite 5), että f-kerroin pienenee, mitä vähemmän on ryhmiä antaen minimin viimeisessä kohdassa (2 ryhmää). Tästä voidaan päätellä, että datassa ei ehkä ole luontevia ryhmiä tai siinä on juuri vain nämä kaksi ryhmää. Mutta on myös mahdollista, että f-kerroin ei anna täysin luotettavaa

tulosta monien prosessointien seurauksesta. Seuraavana tarkastelemme lähemmin näitä lopputuloksena syntyneitä ryhmiä. Ryhmään 1 ryhmittyi 4,8 % koko aineiston vakuutuksista eli vain 141 vakuutuksen tiedot ja vastaavasti ryhmään 2 jäivät loppujen 3255 vakuutuksen tiedot. Ryhmä 2 sisälsi kaikkia muuttujia tasapuolisesti alkuperäisen aineiston mukaan ja tiedoston analysointi oli silmämääräisesti vaikeaa vakuutusten lukumäärän johdosta.

Ryhmässä 1 havaittiin olevan useiden muuttujien kesken voimakasta korrelaatiota (liite 6). Vakuutuksen kohdetyyppi (2) korreloi indeksityypin (11) kanssa –0,51 ja asiakkaan syntymävuosi (3) korreloi asiakkaan roolin (6) kanssa –0,68. Kumpaankaan korrelaatioon ei löydy yksinkertaista selitystä. Lisäksi suoraveloitusaineisto (7) korreloi indeksityypin (11) kanssa 0,65, joita yhdistävänä asiana havaittiin olevan yhtenevä arvoalue. Vakuutuksen omavastuu (9) korreloi vakuutuksen päättymissyyn (12) kanssa –0,58 sekä vakuutuksen vuosimaksun (17) kanssa 0,72. Myöskin alennusprosentti (10) korreloi vakuutuksen päättymissyyn (12) kanssa –0,51, vakuutusmäärän (16) 0,70 ja alennus euroina vuosimaksusta (20) 0,86 kanssa. Lisäksi korrelaatiota esiintyy vakuutuksen päättymissyyn (12) ja vakuutusmäärän (16) –0,68 sekä vakuutusmaksun (17) –0,64 kanssa. Sen sijaan ryhmässä 2 korrelaatiota esiintyy muuttujien kesken vähemmän. Vakuutuksen kohdetyyppi (2) korreloi indeksityypin (11) kanssa –0,69 ja asiakkaan syntymävuosi (3) korreloi sukupuolen (5) –0,59 sekä asiakkaan roolin (6) kanssa –0,64. Lisäksi korrelaatiota esiintyy vakuutuksen päättymissyyn (12) ja laskun tilan (19) kesken 0,65.

Tarkastelemme seuraavaksi muuttujittain ryhmien sisältämiä vakuutusten tietoja. Molemmissa ryhmissä vakuutuslajina (1) olivat edustettuna kaikki vakuutuslajit tasapuolisesti, eikä mikään lajeista erottunut hallitsevana lukumääränä ryhmässään. Kohdetyypin (2) perusteella ryhmän 1 sisältö jakaantui palokohteisiin 85,1 %, autokohteisiin 14,2 %, venekohteisiin 0,7 % ja muihin 0,6 %. Vastaavasti ryhmän 2 sisältö jakaantui palokohteisiin 72,2 %, autokohteisiin 26,2 % ja venekohteisiin 1,0 %. Kohdetyypin mukaan vakuutukset olivat jakautuneet näiden kahden ryhmän kesken tasaisesti. Vakuutuksenottajien syntymävuoden (3) keskiarvo ryhmässä 1 oli 1934 ja ryhmässä 2 se oli 1939 (taulukko 6.9). Tämän mukaan ryhmässä 2 oli hieman nuorempia vakuutuksenottajia, mutta sillä ei ollut suurempaa merkitystä. Vakuutuksenottajien keski-ikä oli aika korkea, mikä johtuu siitä, että mukana oli jo useita vuosia sitten päättyneitä vakuutuksia.

Taulukko 6.9: Ryhmittelyn tuloksena syntyneiden ryhmien muuttujien keskiarvot.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*
Ryhmä1	74	1	34	265	2	3	0	6	140	3	1	7	4045	24	7	77754	303	0	3	85	45
Ryhmä2	88	2	39	278	2	2	0	5	88	1	1	11	3961	50	5	32236	189	0	8	85	0
Koko	88	2	39	278	2	2	0	5	90	1	1	11	3964	49	5	34126	194	0	7	85	2

Vakuutuksenottajien asuinkunnat (4) jakautuivat molemmissa ryhmissä melko tarkasti koko kannan mukaisesti, eikä edes pienemmässä ryhmittymässä (ryhmä 1) mikään kunta tai alue erottunut hallitsevana. Eri asiamiestunnukset (13) olivat jakautuneena tasaisesti molempien ryhmien kesken, eikä mikään niistä erottunut lukumäärällisesti hallitsevana. Lisäksi molemmissa ryhmissä vakuutukset olivat pääsääntöisesti sidottu elinkustannusindeksiin (11), ja ne olivat olleet voimassa (8) keskiarvon mukaan 6 vuotta ryhmässä 1 ja 5 vuotta ryhmässä 2.

Asiakasroolin (6) perusteella havaittiin ryhmässä 1 olevan paljon yksityisiä henkilöitä, joista miespuolisia oli 60 % ja naisia 20 %. Muut asiakasroolit (20 %) edustivat kuolinpesiä, tuntemattomia henkilöitä tai avio- ja avopareja. Tämä lähes sama jakauma toistui myös ryhmässä 2. Vakuutusten omavastuun (8) keskiarvo oli 140 euroa ryhmässä 1 ja 88 euroa ryhmässä 2. Kummassakaan ryhmässä vakuutuksenottajat eivät ole käyttäneet suoraveloitusta (7) maksutapana ja vakuutuksen alennusprosentit (10) olivat molemmissa ryhmissä yleisesti hyvin pieniä. Vakuutuksen päättymissyy (21) oli melkein kaikissa ryhmän 1 vakuutuksissa jokin muu yhtiö (93,4 %). Tämä näytti olevan yksi yhdistävä tekijä kaikkien tämän ryhmän vakuutusten kesken. Asiakkaan irtisanoma vakuutus kirjataan nollaksi, koska aina ei tiedetä, sitä päättyykö se toiseen yhtiöön siirryttäessä vai sanotaanko se irti tarpeettomana. Lukumäärältään suurimpia vakuutusten saajia olivat If (45 %), Pohjola (21 %) ja Tapiola (16 %). Tämä kuvastaa myös aika pitkälle markkinaosuuden mukaista jakaumaa. Ryhmässä 2 ei näytä olevan tätä ominaisuutta vakuutuksen päättymissyyn muuttujalla.

Korvaukset (14) eivät myöskään olleet merkittävä yhdistävä tekijä kummassakaan ryhmässä, koska vain 5 % (ryhmä 1) ja 2,1 % (ryhmä 2) ryhmien vakuutuksista oli suoritettu korvausta. Tästä havaittiin, että ainakaan korvaustilanteet eivät ole aiheuttaneet tämän ryhmän irtisanomisia. Eniten oli päättynyt sellaisia vakuutuksia, jotka ovat olleet jo aikaisemmin vanhoina asiakkaina tai tulleet voimaan täysin uusina vakuutuksina. Suurimpana menettäjänä vakuutusyhtiöistä oli If, mikä johtuu suuresta markkinaosuudesta. Yleensähän sen on helpompi menettää, jolla on eniten markkinaosuutta.

Vakuutusmäärä (16) oli molempien ryhmien useissa vakuutuksissa nolla euroa. Nämä vakuutukset ovat autovakuutuksia, koska autovakuutusten korvausperusteena on aina ajoneuvon vahinkohetken arvo (päivänarvo) eikä niiden vakuutuksessa mainita korvausmäärää. Nämä autovakuutukset häiritsivät ryhmittelyä vakuutusmäärän osalta. Keskimääräinen vakuutusmaksu (17) oli 303 euroa ryhmässä 1 ja 189 euroa ryhmässä 2. Tästä havaitaan ryhmässä yksi olevan asiakkaita, jotka maksavat suurempia vakuutusmaksuja kuin ryhmässä 2. Vastaavasti viivästyskorko (18) oli molemmissa ryhmissä keskimäärin 0 euroa. Vakuutuksenottajat ovat maksaneet maksunsa eräpäivänään ja molempien ryhmien keskiarvon mukainen alennus vakuutusmaksusta (20) oli 85 euroa.

6.4.2 Ryhmittelyn lopputulos

Lopputuloksena todetaan ensimmäisen ryhmittelyn jälkeen, että ryhmä 1 sisälsi kaikkia vakuutuslajeja, joiden päättymissyy oli asiakkaan siirtyminen jonkin toisen yhtiön asiakkaaksi. Näistä vakuuttajista suurin osa oli miehiä (60 %) ja nämä vakuutukset olivat olleet voimassa keskimäärin 6 vuotta. Korvauksilla ei ollut merkitystä tämän ryhmän vakuutuksen päättymiseen, koska korvausta oli maksettu vain 5 %:lle vakuuttajista. Annetut ylimääräiset alennukset olivat pieniä, joten irtisanomisen perusteena on voinut olla vanhentunut vakuutusturva. Ryhmän 2 osalta oli vaikea löytää mitään syytä vakuutusten päättymisiin. Ainoana yhdistävänä tekijänä oli, että se sisälsi ne asiakkaat, jotka olivat muuttaneet vakuutustaan ja olivat vielä asiakkaita. Lisäksi on myös mahdollista, että ryhmä 2 voi olla vakuutuksien ”roskakori” , jolle ei löydy selitettävää yhtäläisyyttä.

6.5 Autovakuutusaineiston ryhmittely

Seuraavassa ryhmittelyssä tarkasteltiin pienemmän lähdeaineiston ryhmittymistä. Alkuperäisestä lähdeaineistosta poistettiin kaikki muut vakuutuslajit paitsi autovakuutukset, jolloin lähdeaineistoksi jäi 937 päättyneen autovakuutuksen tiedot. Aineistoa pienennettiin, koska aikaisempi ryhmittely ei tuottanut uusia ryhmiä ja nyt haluttiin selkeyttää aineiston sisältöä pelkästään autovakuutuksiin. Nämä autovakuutukset sisälsivät yksityisten ajoneuvojen (henkilö-, paketti- ja kuorma-auto, sekä moottoripyörä ja mopedi) kasko-,

osakasko- ja palovakuutuksia. Tämän lisäksi myös alkuperäistä 21 muuttujan joukkoa pienennettiin niin, että aineistoon jäi 11 muuttujan tiedot taulukon 6.4 mukaisesti.

Pienin keskihajonta löytyi sukupuolesta (4), vakuutuksen voimassaoloajasta (5) ja vakuutuksen päättymissyystä (6) (taulukko 6.10). Aineiston sukupuolijakaumassa (4) miesten osuus oli 77,5 % ja naisten osuus 21,8 %. Tämä sukupuolijakauma oli ihan luonnollinen, koska miehet omistavat ja hoitavat ajoneuvojensa vakuuttamisen itse. Vakuutusten voimassaoloaika (5) oli aineiston keskiarvon mukaan vain 3,2 vuotta. Tämä osaltaan selittyy sillä, että autot eivät ole pysyvää omaisuutta vaan vaihtonopeus ja autojen vanhetessa niiden arvojen laskeminen luovat asiakkaille tarpeen muuttaa vakuutusturvaansa nopeammin kuin kiinteistöjen kohdalla.

Taulukko 6.10: Autovakuutusaineiston muuttujien minimi-, maksimi- ja keskiarvot (Av) sekä keskihajonta-arvot (Sd).

	1	2	3	4	5	6	7	8	9	10	11
Min	51	11	45	1	0	0	0	0	0	0	0
Max	225	83	858	4	24	9	10999	99	1440	1400	99
Av	143,4	55,4	266,9	1,2	3,2	2,8	52,0	10,1	132,7	194,6	1,3
Sd	77,40	14,17	254,72	0,48	2,70	2,46	578,18	18,76	122,36	307,10	8,59

Aluksi ennen ryhmittelyä tutkittiin aineistoa korrelaatioanalyysilla (liite 7). Tällöin havaittiin vakuutuslajin (1) korreloivan lievästi syntymävuoden (2) kesken 0,20 ja hieman enemmän vakuutuksen voimassaolovuosien (5) kanssa –0,53 (taulukko 6.11), kun taas vakuutuslajien lievästä negatiivisesta korreloitumisesta vakuutuksen voimassaoloaikaan ei löytynyt mitään selvää yhteyttä.

Taulukko 6.11: Vakuutuslajin korrelaatio muihin muuttujiin.

	1	2	3	4	5	6	7	8	9	10	11
1	1,00	0,20	---	---	-0,53	---	---	---	---	---	---

Aineiston tarkemmassa tarkastelussa havaittiin, että vakuutuslajeissa oli kahden eri järjestelmän vakuutuslajien numeroita. Vanhaa järjestelmää edustavat lajikoodit 51, 52, 53, 54 ja 67, kun taas uudempaa numerointia edustavat vakuutuslajit 205, 215 ja 225. Nämä olivat samoja vakuutustuotteita eli suppeita ja rajattuja osakaskovakuutuksia sekä laajoja ja laajimpia kaskovakuutuksia. Tämä numeroiden nimeämisen erilaisuus olisi pitänyt korjata jo

tiedon esikäsittelyssä. Kokeeksi muutettiin data-arvot siten, että 51 = 225, 52 = 205, 53 = 215, 54 = 215 ja 67 = 205. Tämän korjauksen jälkeen tarkasteltiin uudelleen skaalaamattoman aineiston korrelaatiotaulun käyttäytymistä, jolloin havaittiin korrelaatioiden muuttuneen (taulukko 6.12). Korrelaatio syntymävuoden (2) ja vakuutuksen voimassaoloajan (5) mukaan oli pienentynyt 0,28, mutta vastaavasti vakuutusmaksu (9) ja alennus euroina (10) olivat saaneet kohtalaisen suuret samankaltaisuuden arvot –0,66 ja –0,56. Tämän johdosta voitiin unohtaa vakuutuksen vakuutuslajin ja sen voimassaolovuosien kohtalainen korreloituminen.

Taulukko 6.12: Vakuutuslajin (muutettu lajinumerot) korrelaatio muihin muuttujiin.

	1	2	3	4	5	6	7	8	9	10	11
1	1,00	---	---	---	0,28	0,26	---	-0,22	-0,66	-0,56	---

Aineistossa olevalla vakuutuksen päättymissyyllä (6) ja entisen yhtiön tunnuksella (8) ei esiinny merkittävää korrelaatiota minkään muuttujan kanssa. Maksettujen korvausten määrän (7) keskiarvona oli 52 euroa, joka oli erittäin pieni. Vuosimaksu (9) korreloi merkittävästi (0,54) vuosimaksun alennus -muuttujan kanssa, joka voi johtua molempien euromääräisestä arvoista. Vuosimaksun keskiarvo oli 132,7 euroa ja alennuksen keskiarvo oli 194,6 euroa. Asiakkaan syntymävuoden (2) keskiarvo oli aineistossa 1955 eli vakuutuksenottajat olivat keskimäärin 46-vuotiaita. Näiden mainittujen havaintojen perusteella ei aineistosta vielä ennen ryhmittelyä pystytty huomaamaan selviä ryhmiä tai muita yhtäläisyyksiä.

6.5.1 Aineiston ryhmittely kahteen ryhmään

Ryhmittelyssä aineiston annettiin vapaasti ryhmittyä, mutta ryhmittyminen tapahtui jälleen ainoastaan kahden ryhmän kesken. Ryhmittelyssä lisättiin iteroinnin määrää aikaisemmasta luotettavamman ryhmittelytuloksen saamiseksi ja lisäksi ensimmäisessä ryhmittelyssä etäisyyden laskennassa käytetty f-kerroinmenetelmä vaihdettiin pienimmän Euclidisisen-etäisyyden menetelmäksi. Näiden molempien ryhmien sisällä havaitaan korreloivan voimakkaimmin koko aineiston muuttujien vuosimaksun (9) 0,64 ja vuosimaksun alennuksen (10) 0,81 keskenään (taulukko 6.13). Aineisto ryhmittyikin lähinnä näiden kahden keskenään korreloivan muuttujan välillä. Tämä on aika yllätyksetön tulos, koska nämä muuttujat olivat aika karkeasti porrastettuja ja riippuvat eromääräisesti toisistaan. Mielestäni nämä korkeat korreloitumisen arvot johtuivat siitä, että kaikissa autovakuutuksissa on melkein aina

alennusta (bonus-, ABS- ja ikäalennukset) ja vuosimaksu on lähes samalla tasolla kaikissa samanlaisissa vakuutuslajeissa (pienet erot automerkkien ja mallien välillä). Näiden kahden muuttujan välillä oli molempien ryhmien osalla paljon samankaltaisuutta, mutta tämä ei kuitenkaan vielä täysin paljasta tämän ryhmäjaon perustetta.

Taulukko 6.13: Muuttujien korrelointi ryhmittelytuloksen kanssa (2 ryhmää/11 muuttujaa).

	1	2	3	4	5	6	7	8	9	10	11
Keskiarvo	---	-0,23	---	---	---	-0,32	---	0,34	0,64	0,81	---

Myös asiakkaan aikaisemmalla vakuutusyhtiöllä (8) oli aineistoon nähden ryhmittelevä vaikutus, joskin vain lievä korrelaatio 0,34. Mikäli aineistosta poistettaisiin vuosimaksu ja alennus vuosimaksusta tai pelkkä alennus, ryhmittelyyn vaikuttaisi ehkä enemmän jokin muu muuttuja. Seuraavana poistettiinkin aineistosta alennus vuosimaksusta (10) ja ryhmiteltiin aineisto uudelleen (taulukko 6.14). Nyt ryhmittelyn kanssa korreloivat voimakkaasti vakuutuslaji (1) –0,99 ja lievemmin asiakkaan sukupuoli (6) 0,54. Tällöin ryhmät jakautuivat vakuutuslajin numeroiden perusteella, jolloin ryhmässä 2 olivat vanhat numerot ja ryhmässä 1 uudet numerot. Tuloksien tulkinta oli hankalaa, mutta voimakasta korrelaatiota ei pitäisi esiintyä, sillä niiden esiintyminen ilmentäisi kyseisen muuttujan hallitsevan aineistoa. Vakuutusmaksun alennus -muuttujan aineistosta poistamisen seurauksesta ei vuosimaksu enää hallinnut aineistoa. Näiden muuttujien vaikutus perustui yhteisvaikutukseen, sillä ne olivat kaksi samaa suuretta mittaavaa muuttujaa, jotka yhdessä hallitsivat ryhmittelyä.

Taulukko 6.14. Muuttujien korrelointi ryhmittelytuloksen kanssa (2 ryhmää/10 muuttujaa).

	1	2	3	4	5	6	7	8	9	11
Keskiarvo	-0,99	-0,21	---	---	0,54	---	---	---	---	---

Nyt tutkittiin lähemmin alkuperäisen aineiston (muuttujia 11) ryhmittelyn lopputulosta, jossa ryhmään 1 kuului 531 ja ryhmään 2 kuului 406 päättyneen vakuutuksen tiedot. Ryhmän 1 asiakkaiden syntymävuoden (2) keskiarvon perusteella vakuutuksenottajat olivat keskimäärin 43-vuotiaita (taulukko 6.15). Vastaavasti ryhmän 2 asiakkaat olivat keskimäärin 49-vuotiaita eli hieman vanhempia kuin ryhmässä 1. Sukupuolijakaumaksi (4) muodostui ryhmässä 1 miesten osalta 75,3 % ja naisten osalta 23,9 %. Vastaavasti ryhmän 2 miespuolisten asiakkaiden osuus oli 80,8 % ja naisten 18,9 % eli miesten osuus oli tässä ryhmässä hieman suurempi kuin ryhmässä 1. Sukupuolijakaumasta oli vähäinen määrä molemmissa ryhmissä

pariskuntia ja perikuntia. Lisäksi ryhmän 1 vakuutusten voimassaoloajan pituus (6) keskiarvona oli vain 2 vuotta, mistä havaittiin ryhmässä olevan lyhytaikaisempia vakuutuksia kuin koko aineistossa (keskiarvo 3,2 vuotta). Vastaavasti ryhmän 2 vakuutusten voimassaoloajan keskiarvo oli 5 vuotta.

Taulukko 6.15: Ryhmittelyn tuloksena syntyneiden ryhmien muuttujien keskiarvot.

	1	2	3	4	5	6	7	8	9	10	11
Ryhmä1	210,7	57,8	255,4	1,3	2,0	2,7	91,7	7,5	136,5	208,2	1,4
Ryhmä2	55,4	52,1	282,0	1,2	4,9	2,8	0	13,4	127,6	176,7	1,2
Koko	143,4	55,4	266,9	1,2	3,2	2,8	52,0	10,1	132,7	194,6	1,3

Ryhmästä 1 havaittiin vakuutuksen lajin (1) korreloivan –0,67 vakuutuksen vuosimaksun (9) ja –0,56 (liite 8). Lisäksi havaitaan vuosimaksun (9) korreloivan –0,53 alennuksen (10) kanssa. Nämä havainnot tukivat aikaisempia havaintoja ryhmän sisällöstä. Ryhmästä 2 havaittiin vakuutuksen lajin (1) 0,61 ja vuosimaksu (9) 0,56 korreloivan vakuutuksen alennuksen (10) kanssa. Vakuutuslajin (1) minimi- ja maksimiarvoista havaittiin, että ryhmän 1 kaikki luvut olivat uudempaa vakuutusnumeroa eli 205–225, joten ryhmä 2 sisältää vanhemmat vakuutuslajin numerot 51–67. Tällöin vakuutuslajin numerointi oli ainakin yksi molempia ryhmiä yhdistävä tekijä. Ryhmän 2 kaikkien vakuutuksien korvaussumma (7) oli nolla, eli tässä ryhmässä ei ole yhdellekään asiakkaalle maksettu korvausta autovahingosta. Tällöin vakuutuslajin numerot ja maksetut korvaukset olivat tätä ryhmää yhdistävä tekijä.

Vakuutuksen päättymissyy (6) ja kuntakoodi (4) oli esitetty lineaarisesti, jolloin näiden keskiarvon tai keskihajonnan perusteella ei voitu tehdä kovinkaan tarkkoja johtopäätöksiä. Tarkemmin eri syiden lukumäärien tarkastelussa havaittiin asiakkaiden itse irtisanoneen vakuutukset 7,9 %:ssa (ryhmä 1) ja 9,1 %:ssa (ryhmä 2) tapauksista. Suoraan toiseen yhtiöön oli siirtynyt molemmissa ryhmissä noin 3,2 % vakuutuksista. Omistussuhteen muuttuminen oli päättänyt vakuutuksen 27,0 %:ssa (ryhmä 1) ja 24,6 %:ssa (ryhmä 2) tapauksista (taulukko 6.16). Uuden vakuutuksen alkamista entisessä yhtiössä ei näissä tapauksissa pystytä osoittamaan varmasti. Tämän ryhmän asiakaspoistuman arvioidaan molemmissa ryhmissä olevan yli 11,1 %, mutta ehkä alle 20 %.

Taulukko 6.16: Vakuutuksen päättymissyy ryhmässä 1 ja 2 sekä koko aineistossa.

Vakuutuksen päättymissyy		Ryhmä 1	Ryhmä 2	Koko aineisto
"1"	uutta vastaan	58,8 %	58,9 %	58,9 %
"2"	maksu suorittamatta	0,2 %	0,5 %	0,3 %
"3"	lunastamaton	0,2 %	0,0 %	0,1 %
"4"	omistussuhde muuttunut	27,0 %	24,6 %	26,0 %
"5"	vakuutuskohde lakannut	0,4 %	0,7 %	0,5 %
"6"	kohde jää vakuuttamatta	0,4 %	0,3 %	0,3 %
"7"	asiakas toiseen yhtiöön	3,2 %	3,2 %	3,2 %
"8"	asiakasirtisanominen	7,9 %	9,1 %	8,5 %
"9"	varaton / tuntematon	1,9 %	2,7 %	2,2 %

6.5.2 Ryhmittelyn lopputulos

Ryhmittelyn yhteenvetona havaittiin ryhmäjaon tapahtuneen lähes pelkästään vakuutuslajin uuden ja vanhan numeron perusteella. Ryhmien sisällä oli havaittavissa mielenkiintoisia kytkentöjä ja mahdollisia yhtäläisyyksiä, mutta niiden luotettavuudelle ei voida laittaa kovinkaan paljon painoa tässä ryhmittelyssä. Laskuntila (10) hallitsi voimakkaasti ryhmien muodostumista, joten se poistettiin ja aineisto ryhmiteltiin uudelleen. Tällöinkin ryhmittelyssä aineisto ryhmittyi vakuutuslajin perusteella edelleen lähes samanlaisesti kahteen ryhmään.

6.5.3 Aineiston ryhmittely kolmeen ryhmään

Kahteen ryhmään ryhmittelyssä aineistosta ei löydetty merkittäviä lopputuloksia eikä siitä pystytty tekemään merkittäviä johtopäätöksiä, jolloin seuraavaksi päätettiin kasvattaa ryhmien määrää. Tarkasteltiin ryhmittelyn tuloksena syntynyttä kolmen ryhmän korrelaatiota koko aineistoon, jolloin taulukosta 6.17 havaitaan ryhmien korreloivan voimakkaammin vakuutuslajin (1) 0,50, vuosimaksun (9) 0,51 ja alennus vuosimaksusta (10) 0,70 muuttujien mukaan. Mielestäni nämä korkeat korreloitumisen arvot johtuivat siitä, että kaikissa autovakuutuksissa on varmaankin aina alennusta vuosimaksusta.

Taulukko 6.17: Muuttujien korrelointi ryhmittelytulosten kanssa (3 ryhmää/11 muuttujaa).

	1	2	3	4	5	6	7	8	9	10	11
Keskiarvo	0,50	---	---	---	-0,39	-0,27	---	0,27	0,51	0,70	---

Taulukosta 6.18 havaitaan ryhmien korrelaation muuttuvan, mikäli aineistosta poistetaan alennus vakuutusmaksusta (10) muuttuja. Lopputuloksena havaitaan syntyvän likipitäen sama korrelaatio kuin aikaisemmin kahteen ryhmään jakautuneessa ryhmittelyssä (taulukko 6.14). Edellisessä ryhmittelyssä on havaittu, että ryhmämäärän lisääminen tai poistaminen ei oleellisesti muuta ryhmittelyn lopputulosta.

Taulukko 6.18: Muuttujien korrelointi ryhmittelyn kanssa (kolme ryhmää/10 muuttujaa).

	1	2	3	4	5	6	7	8	9	11
Keskiarvo	-0,98	-0,21	---	---	0,53	---	-0,22	---	---	---

Tarkastelemme alkuperäisen ryhmittelyn (11 muuttujaa) lopputuloksena syntynyttä ryhmäjakoa, jossa vakuutukset jakautuivat lukumääräisesti kolmen ryhmän kesken melko tasaisesti. Ryhmään 1 ryhmittyi 299 päättyneen vakuutuksen tiedot, joissa havaitaan lievää korrelaatiota eri muuttujien kesken. Ryhmää yhdistävänä tekijänä oli kaikkien vakuutuksien korvausmäärä (7) nolla eli tässä ryhmässä ei ole yhdellekään asiakkaalle maksettu korvausta autovahingosta. Ryhmään 2 kuului 275 vakuutuksen tiedot, eikä siinä havaittu olevan merkittävää korrelaatiota minkään muuttujan kanssa. Ryhmään 3 jäi loput 363 vakuutusta ja tässä ryhmässä havaitaan merkittävää korreloitumista vakuutuslajin (1) ja vakuutuksen voimassaoloajan (5) kanssa –0,51 (liite 9).

Asiakkaiden syntymävuoden (2) keskiarvot ryhmässä 1 ja 3 olivat 1950–52. Vastaavasti ryhmän 2 vakuutuksenottajat ovat keskimäärin 10 vuotta nuorempia kuin ryhmässä 1. Lisäksi vakuutusten voimassaoloajan pituuden (5) keskiarvo on ryhmässä 1 pisin 5,3 vuotta ja ryhmässä 2 lyhin 1,9 vuotta (keskiarvo 3,2 vuotta). Kuntakoodin (3) mukaan ei missään ryhmässä yksikään kunta dominoi, koska kaikkien ryhmien keskiarvot ovat lähekkäin. Lisäksi vakuutuslajin (1) keskiarvosta havaitaan, että ryhmän 1 kaikki luvut edustavat vanhempaa vakuutusnumeroa (51–67) ja vastaavasti ryhmä 2 sisältää enemmän uuden ryhmän vakuutuksia (205–225). Tällöin vakuutuslajin numerot ja maksetut korvaukset ovat ryhmää 1 yhdistävä tekijä.

Taulukko 6.19: Ryhmittelyn tuloksena syntyneiden ryhmien muuttujien keskiarvot.

	1	2	3	4	5	6	7	8	9	10	11	Luku-määrä
Ryhmä 1	54,8	1952,8	298,6	1,2	5,3	3,3	0,0	7,1	82,4	27,6	1,7	299
Ryhmä 2	212,0	1961,4	247,3	1,3	1,9	3,2	81,8	4,7	89,4	35,0	1,9	275
Ryhmä 3	149,2	1950,2	258,3	1,2	2,7	1,6	69,2	20,3	244,3	586,9	0,2	363
Koko aineisto	143,4	1955,4	266,9	1,2	3,2	2,8	52,0	10,1	132,7	194,6	1,3	937

Sukupuolten (4) keskiarvo on kaikissa ryhmissä 1,2–1,3 eli miesten osuus kaikissa ryhmissä on suuri, mutta naiset ovat taulukon 6.20 mukaan lähes tasaisesti jakautuneina (19,0 % ryhmässä 1, 25,6 % ryhmässä 2 ja 19,3 % ryhmässä 3) kaikkien ryhmien kesken. Ryhmään 2 on ryhmittynyt eniten naisia, mutta erot muihin ryhmiin ovat kuitenkin pienet.

Taulukko 6.20: Sukupuolten jakautuminen ryhmien kesken.

Sukupuoli		Ryhmä 1	Ryhmä 2	Ryhmä 3	Koko aineisto
"1"	mies	79,6 %	73,8 %	80,0 %	77,5 %
"2"	nainen	19,0 %	25,6 %	19,3 %	21,7 %
"3"	yritys	0,4 %	0,0 %	0,4 %	0,2 %
"4"	perikunta, kuolinpesä	1,0 %	0,6 %	0,3 %	0,6 %
"5"	pariskunta	0,0 %	0,0 %	0,0 %	0,0 %

Vakuutuksen päättymissyy (6) on esitetty lineaarisesti, jolloin keskiarvon tai keskihajonnan perusteella ei voida tehdä kovinkaan tarkkoja johtopäätöksiä (taulukko 6.21). Tarkemmin eri syiden lukumäärien tarkastelussa havaitaan asiakkaiden itse irtisanoneen 8,4 % tai siirtäneen toiseen yhtiöön 3,2 % vakuutuksista, jolloin nämä muodostavat yhteensä 11,6 %:n asiakaspoistuman (taulukko 6.21). Omistussuhteen muuttuminen on päättänyt vakuutuksen 26,0 %:ssa tapauksista, mutta uuden asiakkaan uuden vakuutuksen alkamista entisessä yhtiössä ei näissä tapauksissa pystytä osoittamaan varmasti.

Taulukko 6.21: Vakuutuksen päättymissyy ryhmässä ja koko aineistossa.

Vakuutuksen päättymissyy		Ryhmä 1	Ryhmä 2	Ryhmä 3	Koko aineisto
"1"	uutta vastaan	49,5 %	49,3 %	83,3 %	58,9 %
"2"	maksu suorittamatta	0,7 %	0,2 %	0,0 %	0,3 %
"3"	lunastamaton	0,0 %	0,00 %	0,3 %	0,1 %
"4"	omistussuhde muuttunut	28,1 %	32,5 %	13,8 %	26,0 %
"5"	vakuutuskohde lakannut	1,0 %	0,6 %	0,0 %	0,5 %
"6"	kohde jää vakuuttamatta	0,3 %	0,6 %	0,0 %	0,3 %
"7"	asiakas toiseen yhtiöön	4,3 %	1,9 %	0,7 %	3,2 %
"8"	asiakaan irtisanominen	12,4 %	10,7 %	4,0 %	8,4 %
"9"	varaton / tuntematon	3,7 %	1,9 %	1,1 %	2,2 %

Ryhmän 3 vakuutuksista yli 50 % on siirtynyt lähes kokonaan kilpailijoilta, koska entinen asiakassuhde on ollut vain alle 44,3 % näiden ryhmien vakuutuksista. Yllätyksenä havaitaan ryhmän 2 vakuutuksissa vanhan asiakassuhteen olleen jopa 84,8 % tapauksista (taulukko 6.22), joten nuoret ovat vaihtaneet ajoneuvoja tiheästi ja ovat myös olleet erittäin asiakasuskollisia.

Taulukko 6.22: Ryhmittäin eriteltynä asiakkaan entinen yhtiö.

Mistä yhtiöstä vakuutus on siirtynyt		Ryhmä 1	Ryhmä 2	Ryhmä 3	Koko aineisto
"0"	Entinen asiakas	67,6 %	84,8 %	44,3 %	71,0 %
"3"	Lähivakuutus	13,0 %	1,1 %	10,2 %	8,9 %
"4"	A-vakuutus	0,3 %	0,8 %	1,5 %	0,3 %
"22"	Tapiola	5,0 %	3,6 %		0,1 %
"30"	Kansa	5,0 %	0,8 %	2,2 %	3,0 %
"36"	Pohjola	3,3 %	3,9 %	8,4 %	4,5 %
"41"	Sampo	2,7 %	3,6 %	6,5 %	4,3 %
"55"	Turva	1,3 %	0,3 %	11,6 %	4,2 %
"65"	Y-Fennia	1,7 %	1,1 %	4,0 %	2,4 %
"99"	Yhdistykset	0,0 %	1,9 %	1,5 %	2,2 %

Uuden yhtiön tunnuksen ollessa nolla voi vakuutus olla irtisanottu tarpeettomana tai ajoneuvo on myyty ja sen tilalle on tullut uusi ajoneuvo. Tämä epätarkka jaottelu haittaa aineiston tarkempaa tarkastelua. Kaikissa ryhmissä havaitaan olevan tasaisesti kilpailijoille siirtyneitä asiakkaita (taulukko 6.23). Lisäksi muut vakuutusyhtiöt eivät ole irtisanoneet vakuutuksia vaan irtisanomisliikenne on ollut vähäisempää kuin uusien asiakkaiden siirtyminen.

Taulukko 6.23: Ryhmittäin eriteltynä asiakkaan uusi yhtiö.

Mistä yhtiöstä vakuutus on siirtynyt		Ryhmä 1	Ryhmä 2	Ryhmä 3	
"0"	Entinen asiakas	96,0 %	94,5 %	99,4 %	96,9 %
"3"	Lähivakuutus	0,0 %	0,0 %	0,3 %	0,1 %
"4"	A-vakuutus	0,0 %	0,0 %	0,3 %	0,1 %
"22"	Tapiola	1,0 %	2,2 %	0,0 %	1,0 %
"30"	Kansa	0,0 %	0,0 %	0,0 %	0,0 %
"36"	Pohjola	1,0 %	0,4 %	0,0 %	0,4 %
"41"	Sampo	1,4 %	1,8 %	0,0 %	1,0 %
"55"	Turva	0,0 %	0,0 %	0,0 %	0,0 %
"65"	Y-Fennia	0,3 %	0,0 %	0,0 %	0,1 %
"99"	Yhdistykset	0,3 %	1,1 %	0,0 %	0,4 %

6.5.4 Ryhmittelyn lopputulos

Ryhmittelyn yhteenvetona havaitaan jaon ryhmien 1 ja 2 välillä tapahtuneen pelkästään vakuutuslajin uuden sekä vanhan numeron perusteella. Ryhmä 2 sisältää määrällisesti enemmän naisia, 25,6 %, kuin muissa ryhmissä. Ryhmien sisällä on havaittavissa muitakin mielenkiintoisia kytkentöjä ja mahdollisia yhtäläisyyksiä, mutta niiden luotettavuudelle ei voida laittaa kovinkaan paljon painoa näiden kahden ryhmän ryhmittelyssä.
Ryhmän 3 sisältö näyttää mielenkiintoiselta ja haastavalta. Ryhmä sisältää sellaisia vakuutuksia, jotka ovat tulleet kilpailijoilta (entisestään asiakkaina vain 44,3 %). Lisäksi vakuutusta on muutettu tai uusi ajoneuvo on tullut vakuutukseen 83,3 %:ssa vakuutuksista ja näiden vakuutusten keskiarvon mukainen vuosimaksu on ollut yli 240 euroa. Tämän ryhmän vakuutuksenottajat ovat keskimäärin 51-vuotiaita eli tähän ryhmään ovat ryhmittyneet vanhimmat asiakkaat. Näiden perusteella voidaan todeta ryhmän sisältävän laajoja kaskovakuutuksia ja muihin ryhmiin ovat jääneet suppeat ja rajoitetut vakuutukset.

6.5.5 Aineiston ryhmittely kymmeneen ryhmään

Seuraavassa ryhmittelyssä tarkastelun kohteena on jo aikaisemmin käsitelty autovakuutusaineisto (937 vakuutusta), mutta ainoastaan aineiston muuttujien lukumäärää pienennettiin yhdestätoista aina viiteen. Ryhmittelyohjelmalle annettiin tavoitteeksi löytää aineistosta 10 ryhmää, joiden kesken vakuutukset ryhmittäytyivät. Poikkeuksena edellisestä suorituksista normalisointi ja skaalaus suoritettiin nyt alusta loppuun Excel-taulukkolaskentaohjelmassa. Lisäksi aikaisemmasta toiminnasta poiketen konvertoinnissa käytettiin professori Fräntin kehittämää ASC2CB-ohjelmaa. Aluksi tarkastelemme lopputuloksena saadun kymmenen ryhmän korrelaatiota koko aineistoon, ja siitä havaitaan ryhmien lievä korrelaatio ainoastaan vakuutuksen voimassaoloajan kanssa (8) –0,31 (taulukosta 6.24).

Taulukko 6.24. Muuttujien korrelaatio ryhmittelyn ryhmien kanssa.

	2	4	5	8	11
Keskiarvo	---	---	---	-0,31	---

Ryhmien korrelaatioiden tarkastelussa havaitaan esiintyvän voimakkaita korrelaatioita ryhmissä 3, 6 ja 8 (liite 10). Ryhmässä 3 asiakkaan syntymävuosi (4) korreloi uuden yhtiön (11) kanssa 0,69 ja samoin tekee myös entinen yhtiö (8) 0,77. Ryhmässä 6 esiintyy korrelaatiota syntymävuoden (2) ja sukupuolen (4) kesken –0,64 sekä myös vakuutuksen voimassaoloajan (5) 0,88 ja entisen yhtiön (8) kesken –0,90. Lisäksi sukupuoli (4) korreloi entisen yhtiön (8) kanssa 0,58 sekä myös vakuutuksen voimassaoloajan (5) –0,78 kanssa. Myös ryhmässä 8 havaitaan sukupuolen (4) korreloivan 0,59 entisen yhtiön (8) kanssa. Näiden lisäksi havaitaan monien ryhmien muuttujissa keskihajonnan olevan nolla, joten tämä muuttuja sisältää ainoastaan samoja arvoja. Tällöin havaitaan tämän muuttujan olevan yksi ryhmien muodostamisen perusteista.

Tarkemmassa tarkastelussa havaitaan ryhmään 2 ryhmittäytyneen nuoria asiakkaita, joiden syntymävuoden keskiarvo on 1972 eli vakuutuksenottajat ovat alle 30-vuotiaita (taulukko 6.25). Lisäksi tässä ryhmässä vakuutukset ovat olleet voimassa lyhimmän ajan eli 2,1 vuotta. Vastaavasti ryhmässä 8 ovat pisimpään voimassa olleet vakuutukset, joiden voimassaoloajan keskiarvo on 11,8 vuotta ja vakuutuksenottajien keski-ikä on 53 vuotta. Vanhimmat vakuutuksenottajat ovat ryhmittyneet ryhmään 4 ja ovat 71-vuotiaita. Koko aineiston vakuutuksenottajien keski-ikä on 46 vuotta ja vakuutusten voimassaoloaika on 3,2 vuotta.

Taulukko 6.25. Ryhmien vakuutuksenottajien syntymävuoden ja vakuutuksen voimassaoloajan keskiarvot.

	Ryhmä 1	Ryhmä 2	Ryhmä 3	Ryhmä 4	Ryhmä 5	Ryhmä 6	Ryhmä 7	Ryhmä 8	Ryhmä 9	Ryhmä 10	Koko aineisto
Vakuutuksenottajan syntymävuosi	1950	1972	1934	1930	1958	1961	1960	1950	1948	1954	1955
Vakuutuksen voimassaolo vuosina	3,4	2,1	3,3	3,1	2,7	2,8	3,0	3,1	11,8	2,8	3,2

Aineistossa havaitaan ryhmien 3 (6 kpl) ja 6 (4 kpl) sisältävän pienen määrän ryhmittyneitä vakuutuksia. Ryhmän 3 kaikissa vakuutuksissa vakuuttajina ovat olleet kuolinpesät ja perikunnat, jonka perustella tämä ryhmä on ryhmittynyt. Aineiston pienuus näissä ryhmissä voi aiheuttaa virheellisen tulkinnan niiden sisällöstä koko aineistoon nähden. Sukupuolen osalta havaitaan ryhmän 2 sisältävän nuoria miehiä 99,5 %. Kuitenkin naisten osuus on kahdessa ryhmässä 6 ja 8 lähes 98 %, joten naisten päättyneet vakuutukset ovat ryhmittäytyneet näihin ryhmiin (taulukko 6.26).

Taulukko 6.26. Ryhmien sukupuolten jakauma.

Sukupuoli	Ryhmä 1	Ryhmä 2	Ryhmä 3	Ryhmä 4	Ryhmä 5	Ryhmä 6	Ryhmä 7	Ryhmä 8	Ryhmä 9	Ryhmä 10	Koko aineisto
Mies	100,0%	99,5%	0%	96,8%	1,3%	50,0%	66,7%	0%	98,0%	100,0%	81,8%
Nainen	0%	0,5%	0%	3,2%	98,0%	50,0%	33,3%	97,6%	2,0%	0%	17,6%
Yritys	0%	0%	0%	0%	0,7%	0%	0%	2,4%	0%	0%	0,2%
Perikunta, kuolinpesä	0%	0%	100,0%	0%	0%	0%	0%	0%	0%	0%	0,4%
Pariskunta	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Lukumäärä	120	182	6	94	152	4	21	41	49	268	937

Entinen asiakas tarkoittaa vanhaa asiakassuhdetta (taulukko 6.27). Ryhmän 1 ja 8 vakuutukset ovat siirtyneet lähes kokonaan kilpailijoilta, koska entinen asiakassuhde on ollut vain alle 1 %:ssa näiden ryhmien vakuutuksista. Yllätyksenä havaitaan, että ryhmän 2 vakuutuksissa vanha asiakassuhde on ollut jopa 95,1 %:ssa tapauksista. Nämä nuoret ovat vaihtaneet ajoneuvoja tiheästi, mutta he ovat myös olleet erittäin asiakasuskollisia.

Taulukko 6.27. Ryhmittäin eriteltynä asiakkaan entinen yhtiö.

Mistä yhtiöstä vakuutus on siirtynyt		Ryhmä 1	Ryhmä 2	Ryhmä 3	Ryhmä 4	Ryhmä 5	Ryhmä 6	Ryhmä 7	Ryhmä 8	Ryhmä 9	Ryhmä 10	Koko aineisto
"0"	Entinen asiakas	0,6 %	95,1 %	33,2 %	63,8 %	83,7 %	75 %	61,9 %	0 %	83,7 %	44,3 %	70,0 %
"3"	Lähivakuutus	0,0 %	2,8 %	16,7 %	17,0 %	9,2 %	25 %	0 %	0 %	1,1 %	10,2 %	58,9 %
"4"	A-vakuutus	0,0 %	0 %	0 %	5,3 %	0,6 %	0 %	0 %	0 %	0,8 %	1,5 %	0,3 %
"22"	Tapiola	0,8 %	1,6 %	16,7 %	7,5 %	6,5 %	0 %	9,5 %	0 %	3,5 %	9,8 %	0,1 %
"30"	Kansa	13,3 %	0 %	0 %	4,3 %	0 %	0 %	4,8 %	10,0 %	0,8 %	2,2 %	26,0 %
"36"	Pohjola	22,5 %	0,5%	16,7 %	2,1 %	0 %	0 %	9,5 %	29,8 %	3,9 %	8,4 %	0,5 %
"41"	Sampo	24,2 %	0 %	0 %	0 %	0 %	0 %	14,3 %	19,9 %	3,5 %	6,5 %	0,3 %
"55"	Turva	22,1 %	0 %	16,7 %	0 %	0 %	0 %	0 %	22,8 %	0,3 %	11,6 %	3,2 %
"65"	Y-Fennia	11,7 %	0 %	0 %	0 %	0 %	0 %	0 %	14,8 %	0,7 %	4,0 %	8,4 %
"99"	Yhdistykset	2,5 %	0 %	0 %	0 %	0 %	0 %	0 %	2,7 %	1,7 %	1,5 %	2,2 %
Lukumäärä		120	182	6	94	152	4	21	41	49	268	937

Uuden yhtiön tunnuksen ollessa nolla voi vakuutus olla irtisanottu tarpeettomana tai ajoneuvo on myyty ja sen tilalle on tullut uusi ajoneuvo. Tämän osalta alussa tehty epätarkka jaottelu haittaa nyt aineiston tarkempaa tarkastelua. Ryhmissä 3, 8 ja 7 havaitaan asiakkaiden siirtyneen kilpailijoille (taulukko 6.28). Lisäksi muut vakuutusyhtiöt eivät ole irtisanoneet vakuutuksia vaan irtisanomisliikenne on ollut vähäisempää kuin uusien asiakkaiden siirtyminen asiakkaiksi. Muissa ryhmissä vakuutus on jatkunut entisessä yhtiössä muutoksen jälkeenkin (100 %).

Taulukko 6.28. Ryhmittäin eriteltynä asiakkaan uusi yhtiö.

Mihin yhtiöön vakuutus on siirtynyt		Ryhmä 1	Ryhmä 2	Ryhmä 3	Ryhmä 4	Ryhmä 5	Ryhmä 6	Ryhmä 7	Ryhmä 8	Ryhmä 9	Ryhmä 10	Koko aineisto
"0"	Entinen asiakas	100 %	100 %	83,3 %	97,9 %	100 %	0 %	0 %	100 %	100 %	100 %	96,9 %
"3"	Lähivakuutus	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0,1 %
"4"	A-vakuutus	0 %	0 %	0 %	0 %	0 %	0 %	0,1 %	0 %	0 %	0 %	0,1 %
"22"	Tapiola	0 %	0 %	0 %	2,1 %	0 %	0 %	28,6 %	0 %	0 %	0 %	0,9 %
"30"	Kansa	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0,0 %
"36"	Pohjola	0 %	0 %	0 %	0 %	0 %	0 %	23,8 %	0 %	0 %	0 %	0,5 %
"41"	Sampo	0 %	0 %	16,7 %	0 %	0 %	0 %	41,9 %	0 %	0 %	0 %	1,1 %
"55"	Turva	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0,0 %
"65"	Y-Fennia	0 %	0 %	0 %	0 %	0 %	0 %	4,8 %	0 %	0 %	0 %	0,1 %
"99"	Yhdistykset	0 %	0 %	0 %	0 %	0 %	100 %	0 %	0 %	0 %	0 %	0,4 %
LUKUMÄÄRÄ		120	182	6	94	152	4	21	41	49	268	937

6.5.6 Ryhmittelyn lopputulos

Hyvinä piirteinä havaitaan monen ryhmän sisältävän miespuolisia keski-ikäisiä vakuutuksenottajina, joiden vakuutukset ovat olleet voimassa keskimäärin 3 vuotta. Nämä asiakkaat ovat siirtyneet kilpailijoilta, mutta kun vakuutusta on muutettu tai ajoneuvoa uusittu nämä ovat jääneet silti asiakkaiksi. Yllätyksellistä on ryhmän 2 nuorten miespuolisten vakuutuksenottajien vakuutuksien lyhyt voimassaoloaika, mutta he ovat olleet myös uskollisia vakuutusyhtiölle uusiessaan vakuutuksiaan (edelliset vakuutukset samassa yhtiössä 95,1 %:ssa ja uusi vakuutus on jatkunut samassa yhtiössä 100 %:ssa tapaiksissa). Tämän perustella kirjoittajan aikaisempi epäily nuorten herkkyydestä vaihtaa yhtiötä ei välttämättä pidä paikkaansa tässä aineistossa. Ryhmien 3 ja 6 lukumäärällisesti pienen aineiston kriittinen tulkinta voi johtaa virhearvioihin. Ryhmässä 5 ja 8 vakuutuksenottajat ovat pääosin naisia (n. 98 %), joiden keski-ikä on 43–51 vuotta ja vakuutusten voimassaoloaika noin 3 vuotta. Todennäköisesti ryhmään 8 on ryhmittäytynyt kilpailijoilta siirtyneet naisasiakkaat, jotka ovat uusineet vakuutustaan ja jääneet nykyiseen yhtiöön.

Huolestuttavana piirteenä havaitaan ryhmän 7 sisältävän eniten asiakasmenetyksiä kilpailijoille. Tämän ryhmän päättyneistä vakuutuksista ei yksikään ole jatkunut uutena entisessä yhtiössä ja lisäksi näistä vain 61,9 % on ollut aikaisemmin asiakkaana. Tämä ryhmä sisältää sellaisia asiakkaita, jotka ovat tulleet kilpailijoilta asiakkaiksi ja palanneet takaisin kilpailijoille. Asiakasmenetyksen syyhyn ei voida muuttujien vähyyden vuoksi antaa mitään muuta järkevää vastausta kuin ehkä vakuutuksissa oli liian korkeat vakuutusmaksut.

Ryhmittelyn huonona puolena havaitaan, että aineiston tulkintaa varten pitäisi muuttujatietoja olla käytettävissä enemmän kuin viisi, jotta pystyttäisiin selvittämään tarkemmin asiakkaiden piirteitä. Nyt monessa tapauksessa ryhmät ovat muodostuneet yhden muuttujan ympärille, jolloin muiden muuttujien mahdollinen vaikutus ryhmäjakautumiseen on sekava.

6.6 Kaskovakuutusaineiston järjestyminen

Edellisestä ryhmittelystä havaittiin, että aineiston ja muuttujien lukumäärän pieneneminen aiheuttaa ryhmittelyn tulkinnassa vaikeuksia. Tällöin ryhmittyminen tapahtuu herkästi yhden muuttujan ympärille ja ryhmien sisällöstä on vaikea tehdä paikkansa pitäviä päätelmiä. Tämän johdosta katson, että alustavan suunnitelman mukaista kaskovakuutusten ryhmittelyä ei ole enää tarpeellista suorittaa, koska en usko sen perustella enää löytyvän uusia asiakkaita yhdistäviä piirteitä.

7 Yhteenveto

Tämän tutkielman tarkoituksena oli löytää ryhmittelyn avulla vakuutusyhtiön asiakaspoistumaa ennalta ehkäiseviä asiakkaiden ja vakuutusten ominaisuuksia. Tehtävä oli erittäin mielenkiintoinen ja haastava, mutta myös osaltaan vaikea.

Tiedonlouhinnalle perustan muodostaa louhittavan aineiston luonti, jossa juuri oikeiden muuttujien valinta korostuu. Yksikään muuttuja ei saa olla aineistossa liian hallitsevana, koska tällöin ryhmittely tapahtuu tämän muuttujan ympärillä. Lisäksi tiedon pitää olla oikeaa ja virheetöntä, koska tiedon oikeellisuuden tarkistaminen ja muokkaaminen kuluttavat aikaa turhan paljon. Korjaukset tietoihin on tehtävä erittäin huolellisesti, koska mahdollisesti tästä aiheutuu virheitä itse ryhmittelyyn. Ryhmien sisältämien vakuutusten sisäisten yhteyksien löytäminen on vaikeata, sillä tiedon tulkitseminen virheellisesti voi aiheuttaa väärän toimintatavan, jonka seurauksesta yritykselle voi aiheutua haittaa. Ryhmittelyjen tuottaman tiedon analysoijan pitäisi tietää jonkin verran myös analysoitavan yrityksen toiminnasta. Vastaavasti tämä saattaa aiheuttaa analysointiin ennakkoluuloja ja rasitteita jo aikaisemmista havainnoista ja ympäristön käyttäytymisestä. Kuitenkin alasta tietämätön voi ennakkoluulottomasti huomata jonkin ryhmien sisältöä yhdistävän tekijän, mutta virhehavainto voi myös osoittautua erittäin kalliiksi yritykselle.

Koko aineiston ryhmittelyn osalta ei aineistoja pystytty ryhmittelemään riittävän tarkasti, jotta siitä olisi löydetty lopputuloksena jotakin uutta ja mielenkiintoista tietoa asiakaspoistuman tueksi. Tämä ryhmittely vahvisti siten aikaisempaa uskomusta siitä, että asiakaspoistuman syynä on usein unohdettu asiakas. Vanhemmat vakuutuksenottajat (yli 60-vuotiaat) ovat myös tämän tutkimuksen mukaan asiakasuskollisimpia, mikä johtuu usein pitkäaikaisesta hyvin hoidetusta asiakassuhteesta (luottamus) tai positiivisesta korvauskokemuksesta. Näiden vakuuttajien osalta ei ole syytä huoleen, mutta keski-ikäiset vakuutuksenottajat ovat tämän tutkimuksen mukaan ns. riskiryhmää. Yllätyksenä aikaisemmasta olettamastani poiketen nuoret vakuutuksenottajat olivat tämän aineiston mukaan erittäin asiakasuskollisia ja hyviä asiakkaita, joihin panostaminen on järkevää ajatellen myöhempää asiakasuskollisuutta. Sukupuolten välistä eroa ei pystytty näyttämään toteen irtisanomistilanteissa, koska naisten osuus aineistossa vakuuttajina oli vähäinen. Tämä johtuu näkemykseni mukaan vanhasta perinteestä, jolloin mies huolehtii vakuutuksista ja ne ovat aina miehen nimissä.

Toisen ryhmittelyn osalta tehtiin myös saman suuntaiset havainnot kuin koko aineiston ryhmittelyssä. Vastaavasti kolmannen ryhmittelyn aineiston muuttujien lukumäärä oli liian pieni ja muutamat muuttujat muuttuivat ryhmittelyssä liian hallitseviksi, jolloin riippuvuuksien löytäminen oli suorastaan mahdotonta ajatellen sitä asiakkuudenhallinnan näkökulmasta.

Tutkimus vahvisti näkemystäni, että asiakkaaseen pitäisi saada yhteydenotto vähintään 5–6 vuoden välein. Tämä on tärkeää, koska asiakkaan elämässä on käännekohtia, joihin vakuutusyhtiöllä ei välttämättä ole mahdollisuutta heti reagoida, mikäli asiakkaalta ei tule yhteydenottoa. Tässä vaiheessa vakuutusyhtiöllä on ainoastaan toivo siitä, että pitkäaikainen asiakassuhde hyvin hoidettuna pelastaa tilanteen ja asiakas ottaa itse yhteyttä ensimmäiseksi nykyiseen yhtiöönsä. Mutta pelkästään tähän uskomukseen ei kenenkään ole varaa tuudittautua.

Kuitenkin näidenkin ryhmittelyjen ja näkemykseni mukaan vakuutuksen hinta on yksi tärkeimmistä kynnyksistä vakuutusyhtiön vaihdossa. Mikäli asiakkaan vakuutus on tuotteena vanhanmallinen, johon vakuutusturvien muutokset ovat tekemättä sekä ajan tuomat kilpailualennukset puuttuvat, voi asiakassuhde olla herkemmin vaarassa. Nuoret saattavat sanoa vakuutuksen herkemmin irti, koska tähän käyttäytymiseen voi vaikuttaa käytettävän rahan niukkuus tai ennakkoluulottomuus valittaessa vakuutusyhtiötä, jolloin hinta ja vakuutusten houkutteleva markkinointi nousevat tärkeimmiksi kriteereiksi.

Lisäksi vain muutamille vakuutuksenottajille oli suoritettu korvausta, jota ne asiakkaat, joille oli maksettu korvausta ovat olleet tyytyväisiä korvaustilanteiden hoitamiseen. Kuitenkin on mahdollista, että kielteiset korvauspäätökset ovat aiheuttaneet asiakkaiden irtisanomisen. Lisäksi irtisanomisiin voi myös vaikuttaa asioiden huono hoitaminen tai epävarma asiakaspalvelu. Mikäli haluttaisiin saada tietoa korvausten tai asiakaspalvelun vaikutuksesta irtisanomisiin, olisi mielenkiintoista lähettää kyselylomake näille poislähteneille asiakkaille. Kyselyssä olisi maksimissaan 10 kysymystä irtisanomisen perusteista. Vastausprosentti näihin kyselyihin saattaisi kuitenkin olla erittäin pieni, jolloin paras toteutustapana olisi ehkä puhelinkeskustelu entisen asiakkaan kanssa, jonka toteuttaisi vaikka call-center. Lisäksi tämä menettely antaisi myös vielä mahdollisuuden sopia tapaaminen myyntihenkilöstölle mahdollisen asiakassuhteen pelastamiseksi. Tällöin olisi mahdollista saada tarkempaa tietoa muista irtisanomisen syistä kuin vain niistä, jotka saadaan selville vakuutustiedoissa. Tämän

kyselyn vastaukset yhdistettäisiin vakuutusaineistoon, jolloin louhinnasta voitaisiin saada varmasti tarkempaa ja uskottavampaa tietoa kuin pelkästään vakuutusaineiston tutkistelussa. Asiakkaiden kyselyn pitäisi olla jatkuvaa, jonka perusteella aineistoa päivitettäisiin jatkuvasti. Tämä pitäisi myös yhdistää yhdeksi asiakkuudenhallinnan osa-alueeksi.

Tulevia tiedon louhintoja silmällä pitäen pitäisi kiinnittää vielä enemmän huomiota aineiston luomiseen ja muuttujien valintaan. Muuttujat pitäisi valita huolella, jotta ryhmittelyn lopputuloksena syntyisi luotettavaa tietoa. Lisäksi ryhmiteltävien muuttujien lukumäärän pitäisi olla aina suhteessa koko aineiston lukumäärään. Mikäli muuttujien lukumäärä on pieni, kuten viimeisessä ryhmittelyssä (viisi muuttujaa), ryhmien tulkinta on vaikeaa, koska ryhmittely tapahtuu jonkin yleisen muuttujan perusteella, joka on aineistossa hyvin yleinen. Näitä saman kriteerin omaavia vakuutuksenottajia on varmasti muissakin ryhmissä ja niiden löytäminen sieltä on todennäköisesti vaikeaa. Lisäksi tapahtumaketjua pitäisi automatisoida, jolloin käsin aiheutetut virhemahdollisuudet voitaisiin eliminoida kokonaan pois.

Kokonaisuutena tutkimuksesta löydettiin mielenkiintoisia ryhmiä, joiden sisältöjen piirteet tukivat osaltaan jo aikaisempia olettamia. Mutta uuden ja mullistavan vakuutustiedon löytyminen asiakkaiden käyttäytymisestä jäi vain haaveksi tulevia louhintoja varten.

VIITELUETTELO:

Ahola J. & Rinta-Runsala E. : *Data Mining Case Studies in Customer Profiling.*
VTT Information technology. Research Report TTE1-2001-29. Version 1.0 .
Internet WWW-sivu, URL: http://www.vtt.fi/datamining/publications/dm_case_studies.pdf
(12.12.2003)

Antikainen R. : *Kassakuitti kertoo kaiken.*
Sanomalehti. Sanoma, Helsingin Sanomat 13.4.1999.

Billington J.: *Asiakashallinta, näin pidämme parhaat asiakkaamme*.
Talouslehti. Talentum, FAKTA, Harvard Management Update. 1/1999

Bounsaythip C. & Rinta-Runsala E. : Overview of Data Mining for Customer Behavior *Modelling.* VTT Information Technology. Research Report TTE1-2001-18. Version 1.
Internet WWW-sivu, URL : http://www.vtt.fi/datamining/publications/dm_case_studies.pdf
(23.03.2003)

Berry M. & Linoff G. : *Mastering Data Mining: The art and science of customer relation-ship management.*. New York. Wiley Computer Publishing.2000.

Codd E.F., Codd S.B. & Salley C.T. : *Providing OLAP to User-Analysts:An IT Mandate.*
E.F.Codd Accosiates. Internet WWW-sivu, URL :
http://dev.hyperion.com/download_files/resource_library/white_papers/providing_olap_to_user_analysts.pdf (23.03.2004)

Dyché J. : *The CRM Handbook : A Business Guide to Customer Relationship Management.*
Addison-Wesley Information Technology Series. Pearson Education, 2002.

Erkkilä M. : *CRM nousi toiselle aallolle.*
Ammattilehti. Talentum, Tietoviikko 11.10.2001.

Fu L. : *Neural Networks in Computer Inteligence.*
McGraw-Hill, 1994.

Fränti P. & Kivijärvi J. : *Randomized local search algorithm for the clustering problem.*
University of Joensuu. Department of computer science. Report A-1999-5. 24.4.2000.

Han J. & Kamber M. : *Data Mining, Concepts and Techniques.*
Morgan Kaufmann Publishers, Academic Press, 2001.

Hovi A. : *Data Warehousing*,
Tietovarastotekniikka. Suomen Atk-kustannus, 1997

Hursch J. & Hursch C. : *SQL : Structured Query Language.*
Mcraw-Hill, 1991.

Hyvönen E., Karanta I. & Syrjänen M. : *Tekoälyn ensyklopedia*.
Suomen Tekoälyseura ry. Oy Gaudeamus Ab 1993.

Inmon W. H. : *Building the Data Warehouse.*
New York: John Wiley & Sons, 1996.

Inmon W. H. : *Tech Topic, What is a data warehouse?*
Prism Volume 1 No.1. Internet WWW-sivu,
URL: http://www.cait.wustl.edu/cait/papers/prism/vol_no1/index.htm (10.9.2002)

Jain A. K., Murty M. N. & Flynn P, J. : *Data Clustering: A Review*
ACM Computing Surveys, Vol. 31, No. 3, September 1999. Internet WWW-sivu,
URL: http://www.cs.rutgers.edu/~mlittman/courses/lightai03/jain99data.pdf (30.8.2003)

Karanta I. : *Tiedonlouhinta ja sen käyttömahdollisuuksia vakuutusalalla.*
VTT, Tietotekniikka, 6.3.2002.
Internet WWW-sivu, URL: http://www.actuary.fi/ac_kk2_2002/ (30.5.2003)

Kohonen T. : *Self Organizing Maps*.
Springer Series in Information Sciences. Second Edition. Springer, Berlin 1997.

Koikkalainen P. : *Neurolaskennan mahdollisuudet*.
TEKES- julkaisu 43/94. Helsinki 1994.

Kuvaja A. : *VisualBasic, tietokantaohjelmointi.*
Suomen ATK-kustannus Oy. Jyväskylä 1995.

Kärkkäinen J. : *Kanta-asiakastietoja ei osata vielä hyödyntää.*
Talouslehti. Talentum, Kauppalehti Extra 12.01.2001

Laine H. : Tietokantojen perusteet.
Opetusmoniste D404, Helsingin yliopisto, Tietojenkäsittelytieteen laitos 11.1.2000

Luomala J., Heikkinen J., Virkajärvi K., Heikkilä J., Karjalainen A., Kivimäki A., Käkölä T., Uusitalo O. & Lähdevaara H. : *Digitaalinen verkostotalous*.
Tietotekniikan mahdollisuudet liiketoiminnan kehittämisessä.
TEKES**,** Teknologiakatsaus 110/2001. Helsinki 2001.

Myllyniemi P. & Sarjakoski T.: *Itseorganisoituvat kartat alueellisessa analyysissä.*
Geodeettinen laitos, Tiedote 13, 1996.

Newell F. : *Loyalty.com*.
New York: Mcgraw-Hill 2000.

Parr Rud O. : *Data Mining Cookbook.*
John Wiley & Sons, INC. United States of America 2001.

PWHC, PriceWaterHouseCoopers : *The CRM Handbook: from Group to multiindividual.*
PriceWaterHouseCoopers, July 1999.

Rantanen J., Sainio A., Laiho M., Renkonen E. & Silpiö K. : *Relaatiotietokannat.*
Valtion painatuskeskus. Helsinki 1989.

Roiger J. R. & Geatz W. M. : *Data Mining, a tutorial-based primer.*
Person Education, Inc. United States of America 2001.

Rojas R. : *Neural Networks. A Systematic Introduction.*
Springer-Verlag 1996.

Toivanen O. : *Asiakkuudenhallinta on säveltämistä.*
SysOpen, Timo Peltosen haastattelu
Ammattilehti. Talentum, Tietoviikko 19.10.2000.

Törmä M. : *Neuraaliverkot ja niiden käyttö kuvien analysoinnissa.*
Ammattilehti. Maankäyttö 1/1997.

Ullman J. D. : *Principles of DATABASE SYSTEMS*
Second Edition, Computer Science Press, Inc 1982.

Liite 1: Koko aineiston tietuekuvaukset (1/3)

Nro	Nimi	Tyyppi		Pituus	Kommentti
1	**VLLAJI**	kok.luku	i	3	Vakuutuslajin numero
2	**VLKOHDET**	kok.luku	i	1	Kohdetyyppi mikä tulostetaan vakuutuskirjalle "1" = palokohde "2" = yrityskohde "3" = vastuukohde "4" = autokohde "5" = venekohde "6" = yt-kohde
3	**ASYNTVV**	kok.luku	I	2	Asiakkaan syntymävuosi
4	**AKUNTA**	kok.luku	i	3	Kuntakoodi (kolme ensimmäistä merkkiä on KELAn kuntakoodi) "0" = ulkomailla
5	**ASUKUP**	kok.luku	i	1	Sukupuoli "1" = mies "2" = nainen "3" = yritys "4" = perikunta, kuolinpesä "5" = pariskunta (molemmat nimet)
6	**AROOLI**	kok.luku	i	1	Asiakasrooli "1" = yksityinen "2" = Yritys "3" = kuolinpesä "4" = päämies ryhmälle, muttei muuten asiakas "5" = aviopari "6" = avopari "7" = rekisteröity yhdistys "8" = asiakas ei halua antaa henkilötunnusta "0" = tuntematon
7	**ASUORAV**	kok.luku	i	1	Suoraveloitusaineisto "0" = ei suoraveloitusta "1" = suoraveloitusasiakas
8	**VAKIKA**	kok.luku	i	2	Vakuutuksen voimassa olo aika vuosina
9	**VKOMAV**	kok.luku	i	6	Omavastuu euroina
10	**VKALPR**	des.luku	f	5	Alennus %

Nro	Nimi	Tyyppi		Pituus	Kommentti
11	**VKINKO**	kok.luku	i	1	Indeksityyppi "0" = ei sidottu indeksiin "1" = elinkustannusindeksi "2" = rakennuskustannusindeksi "3" = kone indeksi "4" = kuluttajahintaindeksi
12	**VKPSYY**	merkkijono	c	1	Vakuutuksen päättymissyy "1" = uutta vastaan "2" = maksu suorittamatta "3" = lunastamaton "4" = omistussuhde muuttunut "5" = vakuutuskohde lakannut "6" = kohde jää vakuuttamatta "7" = asiakas toiseen yhtiöön "8" =asiakasirtisanominen "9" = varaton / tuntematon "10" = määräaikainen
13	**VKMIES**	merkkijono	i	9	Asiamiestunnus
14	**VKKORYHT**	des.luku	d	9	Vakuutuksesta maksettujen korvausten yhteissumma
15	**VKSYHT**	merkkijono	i	2	Mistä yhtiöstä vakuutus on siirtynyt "0" = Entinen asiakas "3" = Lähivakuutus "4" = A-vakuutus "13" = Hämeen vakuutus "22" = Tapiola "30" =Kansa "32" = Varma "36" = Pohjola "41" = Sampo "50" = Svensk- Finland "52" = Teollisuusvakuutus "56" = Turva "65 = Y-vakuutus / Fennia "66" = Ålands "99" =Yhdistykset
16	**LAVKMK**	des.luku	d	14	Vakuutusmäärä euroa
17	**LAVUMK**	des.luku	d	10	Vuosimaksu euroa (alennukset huomioitu, sisältä lisäturvamaksun)
18	**LAVIKO**	des.luku	f	7	Viivästyskorko euroa

Nro	Nimi	Tyyppi		Pituus	Kommentti
19	**LATILA**	merkkijono	c	2	Laskun tila "1" = maksettu "3" = avoin "4" = muistutettu 1 kerran "5" = ulosmittauksessa "6" = siirretty luottotappioksi "7" = muistutettu 2 kertaa "8" = kokonaan palauttamatta "9" = käsitelty toisella laskulla "10" = ulosotto tulostettu "11" = mitätöity "12" = kokonaan palautettu "13" = karhunesto
20	**LAALMK**	des.luku	d	7	Alennus markkoina vuosimaksusta
21	**VKSYHT2**	kok.luku	i	2	Yhtiökoodi päättyneille vakuutuksille "0" = Entinen asiakas "3" = Lähivakuutus "4" = A-vakuutus "13" = Hämeen vakuutus "22" = Tapiola "30" =Kansa "32" = Varma "36" = Pohjola "41" = Sampo "50" = Svensk- Finland "52" = Teollisuusvakuutus "56" = Turva "65 = Y-vakuutus / Fennia "66" = Ålands "99" =Yhdistykset

Liite 2: Autovakuutusaineiston tietuekuvaukset (1/2)

Nro	Nimi	Tyyppi		Pituus	Kommentti
1	**VLLAJI**	kok.luku	i	3	Vakuutuslajin numero
2	**ASYNTVV**	kok.luku	I	2	Asiakkaan syntymävuosi
3	**AKUNTA**	kok.luku	i	3	Kuntakoodi (kolme ensimmäistä merkkiä on KELAn kuntakoodi) "0" = ulkomailla
4	**ASUKUP**	kok.luku	i	1	Sukupuoli "1" = mies "2" = nainen "3" = yritys "4" = perikunta, kuolinpesä "5" = pariskunta (molemmat nimet)
5	**VAKIKA**	kok.luku	i	2	Vakuutuksen voimassa olo aika vuosina
6	**VKPSYY**	merkkijono	c	1	Vauutuksen päättymissyy "1" = uutta vastaan "2" = maksu suorittamatta "3" = lunastamaton "4" = omistussuhde muuttunut "5" = vakuutuskohde lakannut "6" = kohde jää vakuuttamatta "7" = asiakas toiseen yhtiöön "8" =asiakasirtisanominen "9" = varaton / tuntematon "M" = määräaikainen
7	**VKKORYHT**	des.luku	d	9	Vakuutuksesta maksettujen korvausten yhteissumma
8	**VKSYHT**	merkkijono	i	2	Mistä yhtiöstä vakuutus on siirtynyt "3" = Lähivakuutus "4" = A-vakuutus "13" = Hämeen vakuutus "22" = Tapiola "30" =Kansa "32" = Varma "36" = Pohjola "41" = Sampo "50" = Svensk- Finland "52" = Tseollisuusvakuut "56" = Turva "65 = Y-vakuutus / Fennia "66" = Ålands "99" =Yhdistykset

Nro	Nimi	Tyyppi		Pituus	Kommentti
9	**LAVUMK**	des.luku	d	10	Vuosimaksu euroa (alennukset huomioitu, sisältä lisäturvamaksun)
10	**LAALMK**	des.luku	d	7	Alennus markkoina vuosimaksusta
11	**VKSYHT2**	kok.luku	i	2	Yhtiökoodi päättyneille vakuutuksille Katso VKPSYY

Liite 3: Koko aineiston korrelaatiokertoimet taulukossa ennen ryhmittelyä

	1	2	3	4	5	6	7	8	9		11	12	13	14	15	16	17	18	19	20	21
1	1,00	0,49	0,33	---	---	-0,22	---	-0,37	-0,26	---	-0,35	---		---	---	-0,24	---	---	---	0,27	---
2	0,49	1,00	0,37	---	---	-0,24	---	-0,23	---	-0,23	-0,68	---	---	---	---	-0,23	---	---	---	0,32	---
3	0,33	0,37	1,00	---	-0,58	0,65	---	-0,35	---	---	-0,24	---	---	---	---	-0,21	---	---	---	---	---
4	---	---	---	1,00	---	---	---	---		---	---	---	---	---	---	---	---	---	---	---	---
	---	---	-0,58	---	1,00	0,40	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
6	-0,22	-0,24	-0,65	---	0,40	1,00	---	0,33	---	---	---		---	---	---	---	---	---	---	---	---
7	---		---	---	---	---	1,00	---	---	---	---	---	---	---		---	---	---	---	---	---
8	-0,37	-0,23	-0,35	---		0,33	---	1,00	---	---	---	---	-0,28	---	---	---	---		---	---	---
9	-0,26	---	---	---	---	---	---		1,00	---	---	---	---	---	---	---	0,25	---	---	---	
10	---	-0,23	---	---	---	---	---	---	---	1,00		---	---	---	---	0,27	0,25	---	---	---	---
11	0,35	-0,68	-0,24	---	---	---	---	---	---	0,23	1,00	---	---		---	0,21	0,21	---	0,20	---	---
12	---	---	---		---	---	---	---	---	---	---	1,00	---	---	---	---	---	---	0,65	---	0,35
13	---	---	---	---	---	---	---	-0,28	---	---	---	---		---	---	---	---	---	---		---
14	---	---	---		---	---	---	---	---	---		---	---	1,00	---	---	---		---	---	---
15	---	---	---	---	---	---	---	---	---	---	---	---	---	---	1,00	---	---	---	---	0,23	---
16	-0,24	-0,23	-0,21	---	---	---	---	---	---	0,27	0,21	---	---	---	---	1,00	0,36	---	---	---	---
17		---	---	---	---	---	---		0,25	0,25	0,21	---	---	---		0,36	1,00	---	---	0,30	---
18	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	1,00	---	---	---
19	---	---	---	---	---	---		---	---	---	0,20	0,65	---	---	---	---	---	---	1,00	---	-0,21
20	0,27	0,32	---	---	---	---	---	---	---	---	---	---	---	---	0,23	---	0,30	---	---	1,00	---
21	---	---	---	---	---	---	---	---	---	---	---	0,35	---	---	---	---	---	---	-0,21	---	1,00

Liite 4: Ryhmittely I, keskipisteet CBSHOW-ohjelmalla tulostettuna ASCII-muotoon (loppu-cb.txt)

CBSHOW V.0.17 9.11.00

Input file:	loppu.cb
Version:	VQ CODEBOOK 2.0
BlockSizeX:	21
BlockSizeY:	1
Booksize:	2
Total frequency:	3396
Bytes per element:	3
MinValue:	0
MaxValue:	1000000
Preprocessing:	1
Generation method:	RLS v0.2.8

26572 25543 29146 27960 29433 33012 21147 29082 28539 29204 33951 28996 28306 28607 32765 27517 27653 27749 23142 26529 122372 (141)

45404 57634 41517 28041 22350 21454 29090 22359 22368 22268 9425 29448 28061 28753 23960 22435 24056 28251 27965 35406 26345 (3255)

Liite 5. Ryhmittely I, ryhmittelyohjelman tuottama tuloste (tulos.lop) (1/4)

CBSWDRLS Version 0.2.3 8.1.2003

Clustering evaluation function = F-test
Distance function = Euclidean squared
Number of initial solutions = 1
Centroid swapping method = Random swap
Min number of clusters = 2
Max number of clusters = 256
Number of GLA-iterations = 2
Halting criterion = Fixed
Minimum number of iterations = 100
M = 256, best = 0.0115732, current = 0.0115732, iter = 100.
M = 255, best = 0.0111281, current = 0.0111281, iter = 100.
M = 254, best = 0.0106430, current = 0.0106430, iter = 100.
M = 253, best = 0.0106430, current = 0.0108059, iter = 100.
M = 252, best = 0.0105969, current = 0.0105969, iter = 100.
M = 251, best = 0.0101046, current = 0.0101046, iter = 100.
M = 250, best = 0.0100169, current = 0.0100169, iter = 100.
M = 249, best = 0.00989180, current = 0.00989180, iter = 100.
M = 248, best = 0.00818445, current = 0.00818445, iter = 100.
M = 247, best = 0.00810053, current = 0.00810053, iter = 100.
M = 246, best = 0.00804201, current = 0.00804201, iter = 100.
M = 245, best = 0.00789031, current = 0.00789031, iter = 100.
M = 244, best = 0.00789031, current = 0.00789349, iter = 100.
M = 243, best = 0.00785890, current = 0.00785890, iter = 100.
M = 242, best = 0.00775871, current = 0.00775871, iter = 100.
M = 241, best = 0.00772562, current = 0.00772562, iter = 100.
M = 240, best = 0.00772562, current = 0.00772997, iter = 100.
M = 239, best = 0.00767821, current = 0.00767821, iter = 100.
M = 238, best = 0.00767165, current = 0.00767165, iter = 100.
M = 237, best = 0.00765381, current = 0.00765381, iter = 100.
M = 236, best = 0.00762602, current = 0.00762602, iter = 100.
M = 235, best = 0.00762061, current = 0.00762061, iter = 100.
M = 234, best = 0.00759321, current = 0.00759321, iter = 100.
M = 233, best = 0.00759321, current = 0.00759925, iter = 100.
M = 232, best = 0.00759321, current = 0.00765032, iter = 100.
M = 231, best = 0.00759321, current = 0.00900127, iter = 100.
M = 230, best = 0.00759321, current = 0.00896009, iter = 100.
M = 229, best = 0.00759321, current = 0.00896100, iter = 100.
M = 228, best = 0.00759321, current = 0.00887116, iter = 100.
M = 227, best = 0.00759321, current = 0.00881922, iter = 100.
M = 226, best = 0.00759321, current = 0.00877352, iter = 100.
M = 225, best = 0.00759321, current = 0.00881880, iter = 100.
M = 224, best = 0.00759321, current = 0.00880160, iter = 100.
M = 223, best = 0.00759321, current = 0.00880404, iter = 100.
M = 222, best = 0.00759321, current = 0.00875134, iter = 100.
M = 221, best = 0.00759321, current = 0.00870280, iter = 100.
M = 220, best = 0.00759321, current = 0.00860635, iter = 100.
M = 219, best = 0.00759321, current = 0.00857655, iter = 100.
M = 218, best = 0.00759321, current = 0.00857491, iter = 100.
M = 217, best = 0.00759321, current = 0.00856742, iter = 100.
M = 216, best = 0.00759321, current = 0.00858263, iter = 100.
M = 215, best = 0.00735488, current = 0.00735488, iter = 100.
M = 214, best = 0.00734752, current = 0.00734752, iter = 100.
M = 213, best = 0.00731658, current = 0.00731658, iter = 100.
M = 212, best = 0.00728648, current = 0.00728648, iter = 100.
M = 211, best = 0.00728632, current = 0.00728632, iter = 100.
M = 210, best = 0.00721619, current = 0.00721619, iter = 100.
M = 209, best = 0.00719479, current = 0.00719479, iter = 100.
M = 208, best = 0.00718947, current = 0.00718947, iter = 100.
M = 207, best = 0.00715711, current = 0.00715711, iter = 100.
M = 206, best = 0.00715711, current = 0.00723873, iter = 100.
M = 205, best = 0.00715711, current = 0.00723598, iter = 100.
M = 204, best = 0.00715711, current = 0.00721095, iter = 100.
M = 203, best = 0.00715711, current = 0.00721746, iter = 100.

M = 202, best = 0.00715711, current = 0.00721083, iter = 100.
M = 201, best = 0.00715711, current = 0.00720676, iter = 100.
M = 200, best = 0.00715711, current = 0.00716633, iter = 100.
M = 199, best = 0.00713819, current = 0.00713819, iter = 100.
M = 198, best = 0.00713819, current = 0.00717235, iter = 100.
M = 197, best = 0.00713819, current = 0.00718236, iter = 100.
M = 196, best = 0.00713819, current = 0.00715646, iter = 100.
M = 195, best = 0.00713217, current = 0.00713217, iter = 100.
M = 194, best = 0.00711262, current = 0.00711262, iter = 100.
M = 193, best = 0.00709366, current = 0.00709366, iter = 100.
M = 192, best = 0.00704374, current = 0.00704374, iter = 100.
M = 191, best = 0.00704374, current = 0.00704530, iter = 100.
M = 190, best = 0.00702165, current = 0.00702165, iter = 100.
M = 189, best = 0.00700457, current = 0.00700457, iter = 100.
M = 188, best = 0.00697773, current = 0.00697773, iter = 100.
M = 187, best = 0.00697773, current = 0.00698428, iter = 100.
M = 186, best = 0.00697589, current = 0.00697589, iter = 100.
M = 185, best = 0.00696754, current = 0.00696754, iter = 100.
M = 184, best = 0.00695786, current = 0.00695786, iter = 100.
M = 183, best = 0.00695786, current = 0.00697012, iter = 100.
M = 182, best = 0.00693589, current = 0.00693589, iter = 100.
M = 181, best = 0.00690229, current = 0.00690229, iter = 100.
M = 180, best = 0.00686422, current = 0.00686422, iter = 100.
M = 179, best = 0.00684899, current = 0.00684899, iter = 100.
M = 178, best = 0.00681438, current = 0.00681438, iter = 100.
M = 177, best = 0.00681438, current = 0.00687827, iter = 100.
M = 176, best = 0.00681438, current = 0.00685701, iter = 100.
M = 175, best = 0.00680957, current = 0.00680957, iter = 100.
M = 174, best = 0.00679899, current = 0.00679899, iter = 100.
M = 173, best = 0.00677522, current = 0.00677522, iter = 100.
M = 172, best = 0.00673973, current = 0.00673973, iter = 100.
M = 171, best = 0.00672309, current = 0.00672309, iter = 100.
M = 170, best = 0.00672309, current = 0.00675407, iter = 100.
M = 169, best = 0.00670982, current = 0.00670982, iter = 100.
M = 168, best = 0.00669352, current = 0.00669352, iter = 100.
M = 167, best = 0.00661960, current = 0.00661960, iter = 100.
M = 166, best = 0.00660666, current = 0.00660666, iter = 100.
M = 165, best = 0.00659913, current = 0.00659913, iter = 100.
M = 164, best = 0.00659913, current = 0.00660709, iter = 100.
M = 163, best = 0.00659449, current = 0.00659449, iter = 100.
M = 162, best = 0.00657583, current = 0.00657583, iter = 100.
M = 161, best = 0.00657527, current = 0.00657527, iter = 100.
M = 160, best = 0.00655246, current = 0.00655246, iter = 100.
M = 159, best = 0.00653869, current = 0.00653869, iter = 100.
M = 158, best = 0.00653869, current = 0.00657953, iter = 100.
M = 157, best = 0.00653869, current = 0.00654435, iter = 100.
M = 156, best = 0.00653869, current = 0.00655974, iter = 100.
M = 155, best = 0.00653869, current = 0.00654027, iter = 100.
M = 154, best = 0.00649878, current = 0.00649878, iter = 100.
M = 153, best = 0.00649874, current = 0.00649874, iter = 100.
M = 152, best = 0.00643015, current = 0.00643015, iter = 100.
M = 151, best = 0.00641851, current = 0.00641851, iter = 100.
M = 150, best = 0.00637166, current = 0.00637166, iter = 100.
M = 149, best = 0.00637166, current = 0.00637388, iter = 100.
M = 148, best = 0.00632028, current = 0.00632028, iter = 100.
M = 147, best = 0.00631926, current = 0.00631926, iter = 100.
M = 146, best = 0.00630626, current = 0.00630626, iter = 100.
M = 145, best = 0.00626988, current = 0.00626988, iter = 100.
M = 144, best = 0.00626947, current = 0.00626947, iter = 100.
M = 143, best = 0.00624310, current = 0.00624310, iter = 100.
M = 142, best = 0.00622583, current = 0.00622583, iter = 100.
M = 141, best = 0.00621458, current = 0.00621458, iter = 100.
M = 140, best = 0.00618550, current = 0.00618550, iter = 100.
M = 139, best = 0.00615457, current = 0.00615457, iter = 100.
M = 138, best = 0.00615457, current = 0.00626085, iter = 100.
M = 137, best = 0.00615457, current = 0.00623453, iter = 100.
M = 136, best = 0.00615457, current = 0.00620264, iter = 100.

M = 135, best = 0.00615457, current = 0.00618625, iter = 100.
M = 134, best = 0.00615457, current = 0.00621359, iter = 100.
M = 133, best = 0.00615457, current = 0.00617679, iter = 100.
M = 132, best = 0.00608189, current = 0.00608189, iter = 100.
M = 131, best = 0.00608189, current = 0.00610358, iter = 100.
M = 130, best = 0.00604122, current = 0.00604122, iter = 100.
M = 129, best = 0.00602435, current = 0.00602435, iter = 100.
M = 128, best = 0.00602435, current = 0.00604190, iter = 100.
M = 127, best = 0.00602435, current = 0.00602952, iter = 100.
M = 126, best = 0.00602435, current = 0.00602454, iter = 100.
M = 125, best = 0.00600958, current = 0.00600958, iter = 100.
M = 124, best = 0.00600958, current = 0.00601701, iter = 100.
M = 123, best = 0.00595803, current = 0.00595803, iter = 100.
M = 122, best = 0.00593869, current = 0.00593869, iter = 100.
M = 121, best = 0.00590340, current = 0.00590340, iter = 100.
M = 120, best = 0.00590340, current = 0.00593610, iter = 100.
M = 119, best = 0.00590340, current = 0.00591644, iter = 100.
M = 118, best = 0.00589279, current = 0.00589279, iter = 100.
M = 117, best = 0.00583493, current = 0.00583493, iter = 100.
M = 116, best = 0.00583493, current = 0.00584792, iter = 100.
M = 115, best = 0.00581453, current = 0.00581453, iter = 100.
M = 114, best = 0.00579055, current = 0.00579055, iter = 100.
M = 113, best = 0.00577693, current = 0.00577693, iter = 100.
M = 112, best = 0.00577693, current = 0.00578476, iter = 100.
M = 111, best = 0.00569719, current = 0.00569719, iter = 100.
M = 110, best = 0.00569587, current = 0.00569587, iter = 100.
M = 109, best = 0.00568875, current = 0.00568875, iter = 100.
M = 108, best = 0.00567500, current = 0.00567500, iter = 100.
M = 107, best = 0.00564882, current = 0.00564882, iter = 100.
M = 106, best = 0.00556349, current = 0.00556349, iter = 100.
M = 105, best = 0.00556349, current = 0.00559538, iter = 100.
M = 104, best = 0.00556349, current = 0.00563378, iter = 100.
M = 103, best = 0.00556349, current = 0.00561665, iter = 100.
M = 102, best = 0.00556349, current = 0.00561070, iter = 100.
M = 101, best = 0.00555513, current = 0.00555513, iter = 100.
M = 100, best = 0.00555513, current = 0.00556243, iter = 100.
M = 99, best = 0.00551397, current = 0.00551397, iter = 100.
M = 98, best = 0.00548565, current = 0.00548565, iter = 100.
M = 97, best = 0.00548287, current = 0.00548287, iter = 100.
M = 96, best = 0.00548287, current = 0.00552868, iter = 100.
M = 95, best = 0.00548287, current = 0.00549357, iter = 100.
M = 94, best = 0.00548287, current = 0.00549900, iter = 100.
M = 93, best = 0.00541668, current = 0.00541668, iter = 100.
M = 92, best = 0.00539082, current = 0.00539082, iter = 100.
M = 91, best = 0.00539082, current = 0.00553270, iter = 100.
M = 90, best = 0.00538791, current = 0.00538791, iter = 100.
M = 89, best = 0.00533758, current = 0.00533758, iter = 100.
M = 88, best = 0.00527177, current = 0.00527177, iter = 100.
M = 87, best = 0.00521923, current = 0.00521923, iter = 100.
M = 86, best = 0.00519849, current = 0.00519849, iter = 100.
M = 85, best = 0.00519849, current = 0.00538105, iter = 100.
M = 84, best = 0.00519849, current = 0.00537740, iter = 100.
M = 83, best = 0.00519849, current = 0.00538419, iter = 100.
M = 82, best = 0.00519849, current = 0.00524518, iter = 100.
M = 81, best = 0.00519310, current = 0.00519310, iter = 100.
M = 80, best = 0.00512236, current = 0.00512236, iter = 100.
M = 79, best = 0.00508684, current = 0.00508684, iter = 100.
M = 78, best = 0.00507666, current = 0.00507666, iter = 100.
M = 77, best = 0.00507000, current = 0.00507000, iter = 100.
M = 76, best = 0.00500627, current = 0.00500627, iter = 100.
M = 75, best = 0.00494359, current = 0.00494359, iter = 100.
M = 74, best = 0.00492536, current = 0.00492536, iter = 100.
M = 73, best = 0.00490536, current = 0.00490536, iter = 100.
M = 72, best = 0.00487465, current = 0.00487465, iter = 100.
M = 71, best = 0.00486509, current = 0.00486509, iter = 100.
M = 70, best = 0.00481604, current = 0.00481604, iter = 100.
M = 69, best = 0.00479052, current = 0.00479052, iter = 100.

M = 68, best = 0.00477889, current = 0.00477889, iter = 100.
M = 67, best = 0.00476271, current = 0.00476271, iter = 100.
M = 66, best = 0.00474124, current = 0.00474124, iter = 100.
M = 65, best = 0.00473285, current = 0.00473285, iter = 100.
M = 64, best = 0.00469370, current = 0.00469370, iter = 100.
M = 63, best = 0.00464779, current = 0.00464779, iter = 100.
M = 62, best = 0.00462075, current = 0.00462075, iter = 100.
M = 61, best = 0.00455007, current = 0.00455007, iter = 100.
M = 60, best = 0.00453429, current = 0.00453429, iter = 100.
M = 59, best = 0.00450148, current = 0.00450148, iter = 100.
M = 58, best = 0.00446293, current = 0.00446293, iter = 100.
M = 57, best = 0.00442404, current = 0.00442404, iter = 100.
M = 56, best = 0.00440986, current = 0.00440986, iter = 100.
M = 55, best = 0.00436776, current = 0.00436776, iter = 100.
M = 54, best = 0.00434986, current = 0.00434986, iter = 100.
M = 53, best = 0.00434986, current = 0.00438615, iter = 100.
M = 52, best = 0.00434986, current = 0.00442641, iter = 100.
M = 51, best = 0.00434986, current = 0.00436497, iter = 100.
M = 50, best = 0.00433360, current = 0.00433360, iter = 100.
M = 49, best = 0.00430814, current = 0.00430814, iter = 100.
M = 48, best = 0.00430814, current = 0.00445091, iter = 100.
M = 47, best = 0.00430814, current = 0.00433736, iter = 100.
M = 46, best = 0.00430814, current = 0.00437751, iter = 100.
M = 45, best = 0.00430814, current = 0.00434520, iter = 100.
M = 44, best = 0.00421355, current = 0.00421355, iter = 100.
M = 43, best = 0.00416561, current = 0.00416561, iter = 100.
M = 42, best = 0.00403862, current = 0.00403862, iter = 100.
M = 41, best = 0.00396121, current = 0.00396121, iter = 100.
M = 40, best = 0.00390958, current = 0.00390958, iter = 100.
M = 39, best = 0.00389336, current = 0.00389336, iter = 100.
M = 38, best = 0.00388030, current = 0.00388030, iter = 100.
M = 37, best = 0.00387589, current = 0.00387589, iter = 100.
M = 36, best = 0.00377736, current = 0.00377736, iter = 100.
M = 35, best = 0.00377736, current = 0.00378324, iter = 100.
M = 34, best = 0.00373555, current = 0.00373555, iter = 100.
M = 33, best = 0.00373555, current = 0.00373655, iter = 100.
M = 32, best = 0.00371387, current = 0.00371387, iter = 100.
M = 31, best = 0.00364304, current = 0.00364304, iter = 100.
M = 30, best = 0.00361002, current = 0.00361002, iter = 100.
M = 29, best = 0.00356895, current = 0.00356895, iter = 100.
M = 28, best = 0.00352963, current = 0.00352963, iter = 100.
M = 27, best = 0.00352963, current = 0.00354179, iter = 100.
M = 26, best = 0.00347238, current = 0.00347238, iter = 100.
M = 25, best = 0.00342005, current = 0.00342005, iter = 100.
M = 24, best = 0.00336539, current = 0.00336539, iter = 100.
M = 23, best = 0.00333633, current = 0.00333633, iter = 100.
M = 22, best = 0.00328031, current = 0.00328031, iter = 100.
M = 21, best = 0.00323674, current = 0.00323674, iter = 100.
M = 20, best = 0.00321847, current = 0.00321847, iter = 100.
M = 19, best = 0.00321847, current = 0.00327451, iter = 100.
M = 18, best = 0.00321847, current = 0.00327926, iter = 100.
M = 17, best = 0.00318192, current = 0.00318192, iter = 100.
M = 16, best = 0.00318192, current = 0.00318537, iter = 100.
M = 15, best = 0.00317089, current = 0.00317089, iter = 100.
M = 14, best = 0.00317089, current = 0.00320201, iter = 100.
M = 13, best = 0.00317089, current = 0.00323343, iter = 100.
M = 12, best = 0.00317089, current = 0.00322469, iter = 100.
M = 11, best = 0.00317089, current = 0.00323068, iter = 100.
M = 10, best = 0.00317089, current = 0.00325623, iter = 100.
M = 9, best = 0.00317089, current = 0.00337183, iter = 100.
M = 8, best = 0.00277976, current = 0.00277976, iter = 100.
M = 7, best = 0.00262953, current = 0.00262953, iter = 100.
M = 6, best = 0.00246414, current = 0.00246414, iter = 100.
M = 5, best = 0.00246414, current = 0.00263310, iter = 100.
M = 4, best = 0.00221876, current = 0.00221876, iter = 100.
M = 3, best = 0.00181998, current = 0.00181998, iter = 100.
M = 2, best = 0.00114106, current = 0.00114106, iter = 100.

Liite 6: Ryhmittely I, kahden ryhmän muuttujien korrelaatiot (21 muuttujaa)

Ryhmä 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	1,00	0,42	---	---	---	-0,27	---	-0,31	---	---	---	---	---	---	---	---	---	---	---	0,21	---
2	0,42	1,00	0,30	---	---	---	---	-0,22	---	---	-0,51	---	---	---	---	---	---	---	---	---	---
3	---	0,30	1,00	---	-0,49	-0,68	---	-0,49	-0,20	---	---	0,28	---	---	---	-0,24	-0,25	---	---	---	0,24
4	---	---	---	1,00	---	---	---	0,24	---	---	---	---	---	---	---	---	---	---	---	---	0,27
5	---	---	-0,49	---	1,00	---	---	---	---	---	---	---	---	---	---	---	0,21	---	---	---	---
6	-0,27	---	-0,68	---	---	1,00	---	0,44	---	---	---	---	---	---	---	---	---	---	---	---	---
7	---	---	---	---	---	---	1,00	---	---	---	0,65	-0,32	---	---	---	---	---	0,20	---	---	0,26
8	-0,31	-0,22	-0,49	0,24	---	0,44	---	1,00	---	---	---	---	-0,31	---	---	---	---	---	---	---	---
9	---	---	-0,20	---	---	---	---	---	1,00	0,31	---	-0,58	---	---	---	0,21	0,72	---	---	0,23	-0,28
10	---	---	---	---	---	---	---	---	0,31	1,00	---	-0,51	---	---	---	0,70	0,37	---	---	0,86	-0,32
11	---	-0,51	---	---	---	---	0,65	---	---	---	1,00	-0,34	---	---	---	---	---	0,21	---	---	---
12	---	---	0,28	---	---	---	-0,32	---	-0,58	-0,51	-0,34	1,00	---	---	---	-0,68	-0,64	---	-0,26	-0,40	0,38
13	---	---	---	---	---	---	---	-0,31	---	---	---	---	1,00	---	---	---	---	---	---	---	---
14	---	---	---	---	---	---	---	---	---	---	---	---	---	1,00	---	---	---	---	---	---	---
15	---	---	---	---	---	---	---	---	---	---	---	---	---	---	1,00	---	---	---	---	---	---
16	---	---	-0,24	---	---	---	---	---	0,21	0,70	---	-0,68	---	---	---	1,00	0,42	---	0,24	0,61	-0,34
17	---	---	-0,25	---	0,21	---	---	---	0,72	0,37	---	-0,64	---	---	---	0,42	1,00	---	---	0,31	-0,39
18	---	---	---	---	---	---	0,20	---	---	---	0,21	---	---	---	---	---	---	1,00	---	---	---
19	---	---	---	---	---	---	---	---	---	---	---	-0,26	---	---	---	0,24	---	---	1,00	---	---
20	0,21	---	---	---	---	---	---	---	0,23	0,86	---	-0,40	---	---	---	0,61	0,31	---	---	1,00	---
21	---	---	0,24	0,27	---	---	0,26	---	-0,28	-0,32	---	0,38	---	---	---	-0,34	-0,39	---	---	---	1,00

Ryhmä 2.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	1,00	0,49	0,34	---	---	-0,22	---	-0,40	-0,30	---	-0,36	---	---	---	---	-0,30	---	---	---	0,27	---
2	0,49	1,00	0,37	---	---	-0,24	---	-0,24	---	-0,24	-0,69	---	---	---	0,20	-0,26	---	---	---	0,34	---
3	0,34	0,37	1,00	---		-0,64	---	-0,37	---	---	-0,24	---	---	---	---	-0,22	---	---	---	---	---
4	---	---	---	1,00	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
5	---	---	-0,59	---	1,00	0,41	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
6	-0,22	-0,24	-0,64	---	0,41	1,00	---	0,35	---	---	---	---	---	---	---	---	---	---	---	---	---
7	---	---	---	---	---	---	1,00	---	---	---	---	---	---	---	---	---	---	---	---	---	---
8	-0,40	-0,24	-0,37	---	---	0,35	---	1,00	---	---	---	---	-0,29	---	---	---	---	---	---	---	---
9	-0,30	---	---	---	---	---	---	---	1,00	---	---	---	---	---	---	---	---	---	---	---	---
10	---	-0,24	---	---	---	---	---	---	---	1,00	0,25	---	---	---	---	---	0,22	---	---	---	---
11	-0,36	-0,69	-0,24	---	---	---	---	---	---	0,25	1,00	---	---	---	---	0,25	0,22	---	0,22	---	---
12	---	---	---	---	---	---	---	---	---	---	---	1,00	---	---	---	---	---	---	0,65	---	---
13	---	---	---	---	---	---	---	-0,29	---	---	---	---	1,00	---	---	---	---	---	---	---	---
14	---	---	---	---	---	---	---	---	---	---	---	---	---	1,00	---	---	---	---	---	---	---
15	---	0,20	---	---	---	---	---	---	---	---	---	---	---	---	1,00	---	---	---	---	0,24	---
16	-0,30	-0,26	-0,22	---	---	---	---	---	---	---	0,25	---	---	---	---	1,00	0,34	---	---	---	---
17	---	---	---	---	---	---	---	---	---	0,22	0,22	---	---	---	---	0,34	1,00	---	---	0,30	---
18	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	1,00	---	---	---
19	---	---	---	---	---	---	---	---	---	---	0,22	0,65	---	---	---	---	---	---	1,00	---	---
20	0,27	0,34	---	---	---	---	---	---	---	---	---	---	---	---	0,24	---	0,30	---	---	1,00	---
21	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	1,00

Liite 7: Autovakuutusaineiston korrelaatiokertoimet taulukossa ennen ryhmittelyä

	1	2	3	4	5	6	7	8	9	10	11
1	1,00	0,20-	---	---	0,53	---	---	---	---	---	
2	0,20	1,00	---	---	-0,21	---	---	---		---	---
3	---	---	1,00	---	---	---	---	---	---	---	---
4	---	---	---	1,00	---	---	---	---	---	---	---
5	0,53	-0,21	---	---	1,00	---	---	---	---	---	---
6	---	---	---	---	---	1,00	---	---	-0,21	-0,23	0,26
7	---	---	---	---	---	---	1,00	---	---	---	---
8	---	---	---	---	---	---	---	1,00	---	0,24	---
	---	---	---	---	---	-0,21	---	---	1,00	0,54	---
10	---	---	---	---	---	-0,23	---	0,24	0,54	1,00	---
11	---	---	---	---	---	0,26	---	---	---	---	1,00

Liite 8: Ryhmittely II, kahden ryhmän muuttujien korrelaatiot (11 muuttujaa)

Ryhmä 1.

	1	2	3	4	5	6	7	8	9	10	11
1		---	---	---	---	0,23	---	---	-0,67	-0,56	---
2	---	1,00	---	---	---	---	---	---	---	-0,29	---
3	---	---	1,00	---	---	---	---	---	---	---	---
4	---	---	---	1,00	---	---	---	---	---	---	---
5	---	---	---	---	1,00	---	---	---	---	---	---
6	0,23	---	---	---	---	1,00	---	---	---	-0,21	0,27
7	---	---	---	---	---	---	1,00	---	---	---	---
8	---	---	---	---	---	---	---	1,00	---	---	---
9	-0,67	---	---	---	---	---	---	---	1,00	0,53	---
10	-0,56	-0,29	---	---	---	-0,21	---	---	0,53	1,00	---
11	---	---	---	---	---	0,27	---	---	---	---	1,00

Ryhmä 2.

	1	2	3	4	5	6	7	8	9	10	11
1		---	---	---	---	-0,23		0,33	0,36	0,61	---
2	---	1,00	---	---	---	---	#####	---	---	---	---
3	---	---	1,00	---	---	---	#####	---	---	---	---
4	---	---	---	1,00	---	---	#####	---	---	---	---
5	---	---	---	---	1,00	---	#####	---	-0,26	---	---
6	-0,23	---	---	---	---	1,00	#####	-0,23	-0,25	-0,27	0,26
7	####	####	#####	####	####	####		####	#####	#####	####
8	0,33	---	---	---	---	-0,23	#####	1,00	0,29	0,46	---
9	0,36	---	---	---	-0,26	-0,25	#####	0,29	1,00	0,56	---
10	0,61	---	---	---	---	-0,27	#####	0,46	0,56	1,00	---
11	---	---	---	---	---	0,26	#####	---	---	---	1,00

Liite 9: Ryhmittely III, kolmen ryhmän muuttujien korrelaatiot (11 muuttujaa)

Ryhmä 1.

	1	2	3	4	5	6	7	8	9	10	11
	1,00	---	---	---	---	---	######	---	---	---	---
	---	**1,00**	---	---	-0,23	---	######	---	---	---	---
3	---	---	**1,00**	---	---	---	######	---	---	---	---
4	---	---	---	**1,00**	---	---	######	---	---	---	---
5	---	-0,23	---	---	**1,00**	---		---	-0,24	---	---
6	---	---	---	---	---	**1,00**	######	---	---	---	0,25
7	######	######	######	######	######	######	######	######	######		######
	---	---	---	---	---	---	######	**1,00**	---	---	---
9	---	---	---	---	-0,24	---	######	---	**1,00**	0,32	---
10	---	---	---	---	---	---	######	---	0,32	**1,00**	---
11	---	---	---	---	---	0,25	######	---	---	---	**1,00**

Ryhmä 2.

	1	2	3	4	5	6	7			10	11
1	**1,00**	---	---	---					-0,38	---	---
2	---	**1,00**	---	---					---	---	---
3	---	---	**1,00**	---					---	---	---
4	---	---	---	**1,00**					---	---	---
5	---	---	---	---	**1,00**	---	---				
6	---	---	---	---	---	**1,00**	---	---	---	---	0,26
7	---	---	---	---	---	---	**1,00**	---	---	---	---
8	---	---	---	---	---	---	---	**1,00**	---	---	---
9	-0,38	---	---	---	---	---	---	---	**1,00**	0,26	---
10	---	---	---	---	---	---	---	---	0,26	**1,00**	0,26
11	---	---	---	---	---	0,26	---	---	---	0,26	**1,00**

Ryhmä 3.

	1	2	3	4	5	6	7	8	9	10	11
1	**1,00**	---	---	---	-0,51	---	---	-0,35	---	---	---
2	---	**1,00**	---	---	---	---	---	---	---	---	---
3	---	---	**1,00**	---	---	---	---	---	---	---	---
4	---	---	---	**1,00**	---	---	0,21	---	---	---	---
5	-0,51	---	---	---	**1,00**	---	---	---	---	---	---
6	---	---	---	---	---	**1,00**	0,25	---	---	---	0,29
7	---	---	---	0,21	---	0,25	**1,00**	---	---	---	---
8	-0,35	---	---	---	---	---	---	**1,00**	---	---	---
9	---	---	---	---	---	---	---	---	**1,00**	---	---
10	---	---	---	---	---	---	---	---	---	**1,00**	---
11	---	---	---	---	---	0,29	---	---	---	---	**1,00**

Liite 10: Ryhmittely IV, kahden ryhmän muuttujien korrelaatiot (5 muuttujaa) (1/2)

Ryhmä 1.

	2	4	5	8	11
2	*1,00*	#####	---	---	#####
4	#####	####		#####	#####
5	---	#####	*1,00*	-0,20	#####
8	---	#####	-0,20	*1,00*	#####
11	#####	#####	#####	#####	####

LKM=120	2	4	5	8	11
Minimi	26	1	1	22	0
Maksimi	77	1	7	99	0
Keskiarvo	50,4	1,0	3,4	46,2	0,0
Keskihajonta	12,24	0,00	1,66	14,17	0,00

Ryhmä 2.

	2	4	5	8	11
2	*1,00*	---	---	---	#####
4	---	*1,00*	---	---	#####
5	---	---	*1,00*	---	#####
8	---	---	---	*1,00*	#####
11	#####	#####	#####	#####	####

LKM=182	2	4	5	8	11
Minimi	62	1		0	0
Maksimi	83	2		36	0
Keskiarvo	72,1	1,0	2,1	0,7	0,0
Keskihajonta	5,71	0,07	1,22	3,88	

Ryhmä 3.

	2	4	5	8	11
2	*1,00*	#####	---	0,25	0,69
4	#####	####	#####	#####	#####
5	---	#####	*1,00*	-0,30	---
8	0,25	#####	-0,30	*1,00*	0,77
11	0,69	#####	---	0,77	*1,00*

LKM=6	2	4	5	8	11
Minimi	24	4	1	0	0
Maksimi	45	4	7	55	41
Keskiarvo	34,2	4,0	3,3	19,3	6,8
Keskihajonta	7,68	0,00	2,16	22,68	16,74

Ryhmä 4.

	2	4	5	8	11
2	*1,00*	---	---	---	---
4	---	*1,00*	---	---	---
5	---	---	*1,00*	0,24	---
8	---	---	0,24	*1,00*	---
11	---	---	---	---	*1,00*

LKM=94	2	4	5	8	11
Minimi	11	1	1	0	0
Maksimi	43	2	8	36	22
Keskiarvo	30,3	1,0	3,1	4,4	0,5
Keskihajonta	7,61	0,18	1,88	9,26	3,19

Ryhmä 5.

	2	4	5	8	11
2	*1,00*	---	-0,25	---	#####
4	---	*1,00*	---	---	#####
5	-0,25	---	*1,00*	---	#####
8	---	---	---	*1,00*	#####
11	#####	#####	#####	#####	####

LKM=152	2	4	5	8	11
Minimi	33	1	1	0	0
Maksimi	82	3	8	22	0
Keskiarvo	59,8	2,0	2,7	1,8	0,0
Keskihajonta	11,04	0,14	1,74	5,45	0,00

Ryhmä 6.

	2	4	5	8	11
2	*1,00*	-0,64	0,88	-0,90	#####
4	-0,64	*1,00*	---	0,58	#####
5	0,88	---	*1,00*	-0,78	#####
8	-0,90	0,58	-0,78	*1,00*	#####
11	#####	#####	#####	#####	####

LKM=4	2	4	5	8	11
Minimi	55	1	1	0	99
Maksimi	67	2	4	3	99
Keskiarvo	61,8	1,5	2,8	0,8	99,0
Keskihajonta	4,99	0,58	1,50	1,50	0,00

Ryhmä 7.

	2	4	5	8	11
2	*1,00*	---	---	---	---
4	---	*1,00*	---	---	---
5	---	---	*1,00*	0,30	---
8	---	---	0,30	*1,00*	0,32
11	---	---	---	0,32	*1,00*

LKM=21	2	4	5	8	11
Minimi	45	1	1	0	22
Maksimi	77	2	7	41	65
Keskiarvo	60,2	1,3	3,0	12,8	35,5
Keskihajonta	9,56	0,48	1,69	17,40	10,58

Ryhmä 8.

	2	4	5	8	11
2	*1,00*	---	---	---	#####
4	---	*1,00*	---	0,59	#####
5	---	---	*1,00*	-0,22	#####
8	---	0,59	-0,22	*1,00*	#####
11	#####	#####	#####	#####	####

LKM=41	2	4	5	8	11
Minimi	30	2	1	30	0
Maksimi	67	3	7	99	0
Keskiarvo	54,7	2,0	3,1	46,8	0,0
Keskihajonta	9,69	0,16	1,64	14,28	0,00

Ryhmä 9.

	2	4	5	8	11
2	*1,00*	---	-0,31	---	#####
4	---	*1,00*	---	---	#####
5	-0,31	---	*1,00*	---	#####
8	---	---	---	*1,00*	#####
11	#####	#####	#####	#####	#####

LKM=49	2	4	5	8	11
Minimi	24	1	8	0	0
Maksimi	67	2	24	22	0
Keskiarvo	48,7	1,0	11,8	0,6	0,0
Keskihajonta	10,09	0,14	3,56	3,18	0,00

Ryhmä 10.

	2	4	5	8	11
2	*1,00*	#####	---	---	---
4	#####	####	#####	#####	#####
5	---	#####	*1,00*	---	---
8	---	#####	---	*1,00*	---
11	---	#####	---	---	*1,00*

LKM=268	2	4	5	8	11
Minimi	43	1	0	0	0
Maksimi	66	1	7	22	13
Keskiarvo	54,1	1,0	2,8	2,7	0,1
Keskihajonta	5,54	0,00	1,63	6,68	0,83