HAAGA-HELIA
ammattikorkeakoulu

Määrällisen datan havainnointi visualisoinneilla

Joonas Kujala

Opinnäytetyö

Tietojenkäsittelyn koulutusohjelma

2016

HAAGA-HELIA
ammattikorkeakoulu

Tiivistelmä

10.9.2016

Tekijä(t)
Joonas Kujala

Koulutusohjelma
Tietojenkäsittelyn koulutusohjelma

Opinnäytetyön otsikko
Määrällisen datan havainnointi visualisoinneilla

Sivu- ja liitesivumäärä
30 + 7

Opinnäytetyön otsikko englanniksi
Observing quantitative data with visualizations

Tässä opinnäytetyössä luodaan katsaus datan visualisoinnin suunnitteluvaiheeseen. Lisäksi työn pääasiallisena tavoitteena on tuoda esille datan visualisoinnin tärkeä rooli määrällisen datan tehokkaassa havainnollistamisessa. Produktina opinnäytetyössä valmistetaan visualisointi Väestörekisterikeskuksen tarjoamasta rakennusten sijaintidatasta, jolla esitetään datan visualisoinnin konkreettinen suunnittelu. Tämän suunnitteluprosessin lopputuloksena on visualisointi, jota käytetään rakennusten levinneisyyden havainnollistamiseen eri alueilla Suomea.

Opinnäytetyö on produktityyppinen systeemityö, jossa tietoperusta on hieman tutkimuksellista. Tietoperustassa käydään läpi alan eri ammattilaisten käyttämiä työtapoja, tavoitteena saavuttaa mahdollisimman kiinnostava ja tehokas aloitus projektille. Opinnäytetyön kappaleet koostuvat perustaosuudesta, käytännön osuudesta, sekä lopulta projektiosuudesta. Perusta koskee suurimmilta osin visualisoinnin suunnitteluvaihetta, mutta tuo kevyesti esille myös käytännön asioita tulevaa varten.

Käytännön osuudessa jäi teoria suurimmilta osin taakse ja työ siirtyy katselmoimaan ja analysoimaan käytännön esimerkkejä valittujen visualisointitöiden avulla. Tässä osassa tavoitellaan mielenkiinnon herättämistä tulevaa projektiosuutta varten paljastamalla datan visualisoinnin tehokkuus.

Projektiosuudessa esitellään produktin suunnitteluvaihe ja toteuttaminen, joiden lopputulos on näkyvillä liitteissä. Projektista syntyneessä visualisoinnissa havainnoidaan Suomen rakennusten jakautumista eri alueittain. Lopuksi produktin suunnittelu ja lopputulos tulkitaan ja analysoidaan vaiheittain, sekä viimeiseksi katselmoidaan tapoja jatkokehitykseen.

Lopputuloksen analyysistä paljastuu karkealla tasolla oletetun mukaisesti, että Suomessa leveyspiireittäin laskettuna rakennusten määrät hiipuvat pohjoiseen päin edetessä. Silti joissakin leveyspiireissä olevien yhden tai useamman suuren kaupungin sijainnista voi johtua vaihtelevuutta. Lopputulos tarkemmalla tasolla esitti sitä, että Suomen länsirannikko on enemmän itäpuolta asutettua. Myös rakennusten määrien katselmointi pienalueittain, kuten tiettyjen paikkakuntien yllä paljasti haja-asutusalueiden ja tiheämmin rakennettujen alueiden erot selkeästi.

Asiasanat
Visualisointi, suunnittelu, rakennusdata, havainnointi

HAAGA-HELIA
ammattikorkeakoulu

Tiivistelmä

10.9.2016

Tekijä(t)
Joonas Kujala

Koulutusohjelma
Tietojenkäsittelyn koulutusohjelma

Opinnäytetyön otsikko
Määrällisen datan havainnointi visualisoinneilla

Sivu- ja liitesivumäärä
30 + 7

Opinnäytetyön otsikko englanniksi
Observing quantitative data with visualizations

This thesis examines the design and planning of data visualizations. In addition the primary objective of the study is to express the importance of data visualization in perceiving insights in data efficiently. The product in this thesis includes the creation of a visualization based on building data published by the Finnish Population Register Centre. The final product of this create process is a visualization, which was used to analyze and demonstrate the distribution of buildings in different areas of Finland.

This thesis is mainly a system project, which has a research based theory section. The theory section examined different frameworks used by the fields' professionals. This thesis consists of a theory section, practice and finally the project section. The theory section consists mainly of designing a visualization project, but also touches on some practical conventions for the future sections.

The practical section consists of insights and analyzations of common visualization practices of hand-picked, professionally made visualizations. In this section the main purpose is to raise interest towards the project section by unraveling the effectivity of data visualization.

The project section showed the design and create process of the main visualization, which is shown in the appendix section of the thesis. The final product regards the Finland’s building data published by the Finnish Population Register Centre. The visualization observes the distribution of Finnish buildings within different areas. Finally the design and the end result were analyzed and planning for the advancement was made.

The end result showed clearly and according to assumptions, that mainly the distribution of Finland’s buildings fades down from south to north, if the data is showed by latitude. However on some latitudes, some of the greater cities cause variance in numbers. More detailed end result showed that the west coast is more populated compared to the east. Also the number of buildings detailed on smaller areas, such as cities, showed a clear difference between settlements of varying population densities.

Asiasanat
Visualization, design, building data, insight

Sisällys

1 Johdanto

Visualisointeja on sovellettu esimerkiksi lehtiartikkeleissa muun muassa painottamaan kirjoitetun tekstin todenmukaisuutta. Tyypillisesti myös TV-uutiset vilauttavat keskusteluissaan joko pylväsdiagrammia kertomaan nousevasta trendistä, tai eri väriskaaloilla koristeltuja maita tai maakuntia vertailevana tekijänä. Useat uutissivustot tai aikakausilehdet, kuten yhdysvaltalaiset The Economist ja New York Times pitävät jopa omaa osiota verkkosivuillaan keskittyen pelkästään datan visualisointeihin. Näissä palveluissa journalistit, analyytikot ja graafiset suunnittelijat tarjoavat yksityiskohtaista, syventävää tietoa eri aihealueista. Journalismissa visualisoinnit yleistävät hyvin tutkitun informaation välittämistä selkokielisesti valtavirtamedian kautta. Oikein tehtynä tämä vahvistaa lukijan asioiden perillä olemista käsitellystä asiasta, eikä jätä lukijan tietämystä spekulaation nojaan. Visualisoinneilla annetaan datan puhua, joka edistää puolueettomuutta, sekä yksipuolisten näkökantojen muotoutumista. (Heer & Segel, 2010.)

Visualisointi on raa'an ja statistisen datan analysointia, esillepanoa ja kommunikointia yleisölle, joka ei välttämättä koostu tietotekniikan tai muun informaatioteknologian asiantuntijoista. Visualisoinnin avulla tulkitsemme ja selitämme abstraktia dataa, sekä vahvistamme tärkeitä havaintoja datasta tehokkaasti muistiimme. Visualisointi yhdistää tietojenkäsittelytieteen, statistiikan, graafisen osaamisen, sekä usein myös tarinankerronnan. (Heer & Segel, 2010.)

Valtavirtamedian lisäksi myös työelämässä datan visualisointi on hyvin yleisessä käytössä. Esimerkiksi erilaiset liiketoiminnalliset verkkopalvelut sekä markkinointijärjestelmät ovat suuria työkaluja, jotka tuottavat ja hallitsevat massiivisia määriä monikäyttöistä dataa. Visualisointi on tehokas harmonisointitapa ihmisen ymmärryksen ja tämän monipuolisen, määrällisen datan välillä. Määrällisen datan visualisoinnit ovat nättejä ja laadukkaita tapoja paketoida elintärkeä informaatio ja nopeuttaa oikeiden päätösten tekemistä. (Heer & Segel, 2010.)

1.1 Opinnäytetyön tavoitteet

Opinnäytetyön tarkoituksena on esittää lukijalle visualisointiprojektin aloitusprosessia, sekä miten edetä suunnittelussa toteutukseen asti. Työssä tavoitteena on myös paljastaa datan visualisoinnin tehokkuus esiteltävän informaation kannalta. Suunnittelun lopputuloksena syntyy visualisointi Suomen rakennuksista alueittain, jonka tavoitteena on esittää yleiskuvaa Suomen rakennusten levinneisyydestä.

1.2 Rajaukset

Luvussa 3 käytettyjen visualisointiesimerkkien poiminnoissa on otettu huomioon ainoastaan visualisoinnin ammattilaisten töitä. Nämä käytännön esimerkit ovat valittu tarkasti ja asiakohtaisesti kappaleen aiheeseen liittyen. Tietoperustan luvun 3 valikoitujen visualisointien lisäksi muutama kuvio on luotu itse. Nämä ovat kappaleen 3 kuviot 4, 7 ja 8, sekä tietysti produkti, eli liitteet 4-5.

Opinnäytetyön projektin koodi on kokonaan kirjoitettu MySQL–ohjelmointikielellä ja datan käsitteleminen on tehty ainoastaan tietokannoilla. Lopputuloksena liitteiden 4-5 visualisoinnit ovat tehty kuvankäsittelyohjelmalla, käyttämällä laskelmia kyseisestä datasta. Opinnäytetyössä ei siis ole käytössä minkäänlaista visualisointijärjestelmää, vaan visualisointien suunnittelu ja toteutus on tehty manuaalisella työllä. Tämä myöhemmin aiheutti pieniä ongelmia, joita käydään läpi opinnäytetyön loppuosassa.

Työllä ei ole toimeksiantajaa, vaan työ on itsenäistä, sekä harrastepohjaista. Produktin luomisessa on käytetty ainoastaan ilmaisohjelmia.

1.3 Rakenne

Opinnäytetyö on produktityyppinen systeemityö, jossa tietoperusta on hieman tutkimuksellista. Luvut koostuvat lineaarisesti alkamalla kevyistä aiheista ja abstrakteista käsitteistä, siirtyy luonnollisesti enemmän käytäntöön ja sitä kautta projektityöhön.

Alussa opinnäytetyön perustaosuus koskee suurimmilta osin suunnitteluvaihetta. Ensimmäinen kappale pohjustaa, sekä näyttää mallia miten alan ammattilaiset rakentavat visualisointeja ja mitä he ottavat suunnitteluvaiheessa huomioon. Pohjustaa visualisoinnin eri formaatteja (kuten elävä esitys, painettu teksti tai kokemus) ja formaattien eri vahvuuksia (esimerkiksi tehokkuus, selkokielisyys, kiinnostavuus).

Tietoperustan toisessa luvussa esitetään datavisualisoinnin alan ammattilaisten tekemiä töitä, joilla korostetaan kappaleessa käsiteltäviä aiheita. Luvun tarkoituksena on esimerkillisesti esittää mitä kaikkea datan visualisoinnilla voidaan saavuttaa. Valittuja töitä tulkitaan ja analysoidaan suunnitteluprosessin ymmärtämisen vuoksi, sekä antamalla useita näkökulmia aiheisiin. Teoria jää siis suurimmilta osin taakse. Esimerkkitöillä katsotaan hieman historiaa, kuinka tehokkaita statististen mallien ohella visualisoinnit ovat ja yleensä mitä kaikkea visualisoinneilla voi tehdä.

Projektiosuus käsittelee määrällisen datan visualisoinnin suunnitteluprosessia ja toteutusta. Lopputuloksena on karttavisualisointi Väestörekisterikeskuksen julkaisemasta avoimesta Suomen rakennusten sijaintidatasta. Työosuudessa valmiista karttavisualisoinnista havainnoidaan Suomen rakennusten jakautumista alueittain. Luku käy läpi käytettävän datan laadun, käytettävän tekniikan, visualisoinnin suunnitteluprosessin ja sen toteutusvaiheet yksityiskohtaisesti. Työn toteuttaminen käydään askel kerrallaan läpi, jonka mukana perustellaan päätökset suunnitteluvalintoihin. Lopputuloksesta sen jälkeen nostetaan esille positiiviset asiat, ongelmat lopputuloksissa ja suunnittelussa, sekä lopuksi katselmoidaan jatkokehitystä.

2 Visualisoinnin perustat

Tämä kappale käsittelee datan visualisointia laajemmalla tasolla ja kuvailee visualisointia esitysmuotona, enemmän kuin konkreettisena tuotteena. Kappaleessa käydään läpi huomioon otettavia asioita ja alkuaskelia onnistuneen visualisoinnin valmistamiseen. Kappaleessa on myös mainittuna muutama alan ammattilainen, joiden työt ja työtavat ovat vaikuttaneet suoraan tämän opinnäytetyön tekoon. Kappaleen viimeinen aliotsikko ohjaa lukijaa enemmän käytännön puolelle ja tuo esimerkiksi termistön seuraavaa kappaletta varten tuoreeseen muistiin.

2.1 Tarinankerronta visualisoinneissa

Käytännössä mikä tahansa data on mahdollista visualisoida, mutta mikä tahansa visualisoitu data ei välttämättä ole ymmärrettävä tai hyödyllinen. Varsinkin laajan datan visualisoinnissa tietyt asiat on hyvä kääntää selkokieliseksi katselijalle. Thomas Davenport, Yhdysvaltalainen tietotekniikan professori, kertoo narratiivia seuraavien visualisointien auttavan katsojaa ymmärtämään ja muistamaan sen paremmin. Davenport mainitsee myös sitä, että ideaalitilanteessa hyvä dataan perustuva tarina kertoo tapahtuneesta, sen aiheuttajasta ja sekä tarjoaa jatkosuunnitelmaa rakentaakseen nykytilanteesta parempaa. (Davenport & Kim, 2013.)

Jo 2010 ja 2013 alan ammattilaiset ovat ottaneet tosissaan käsittelyyn visualisoinnin tarinallistamisen, sekä vihjaillut sen ottavan jopa trendimäisen ulkomuodon. Tableau Software–tietotekniikkayrityksen analyytikot Robert Kosara ja Jock Mackinlay mainitsevat raportissaan, että tarinankerronnan rakenne-elementtien ottaminen mukaan visualisointiin on ollut seuraava looginen askel visualisointien luojille (Kosara & Mackinlay, 2013). Kertomukset ovat tehokas tapa tarjoilla informaatiota helposti omaksuttavassa ja hyödynnettävässä muodossa. Edward Tufte, Yhdysvaltalainen tietotekniikan professori ja datan visualisoinnin pioneeri, mainitsee erinomaisen visualisoinnin koostuvan monimutkaisten ideoiden kommunikoinnista selvyydellä, tarkkuudella ja tehokkuudella. (Tufte 2001, 51.) Käytännössä tämä ei tarkoita sitä, että yksinkertaiset ideat eivät olisi visualisoinnin arvoisia. Tufte välittää sanomallaan karkeasti konkretisoituna sitä, että jos esitettävänä on jokin asia tai käsite, niin siitä pitää saada helposti selvää, sekä esitettävä statistinen data tulee olla validia. Lukijan ohjaaminen visualisoinnin datan rakennetta pitkin auttaa visualisoinnin luojan näkemyksen ymmärtämisessä ja muistamisessa.

Hans Rosling, Ruotsalainen tohtori ja statistikko, kertoo kymmenen minuutin TED–konferenssissaan, nimeltä Hans Rosling And The Magic Washing Machine, eri talous-

luokkien kasvusta johtuvaan teknologian yleistymiseen ja yleistymisen vaikutuksesta sähkönkulutukseen. Puheessansa hän käyttää tarinankertojan kyvykkyyttänsä tuomaan asiansa esille selkeästi ja luovalla tavalla. Rosling kaappaa yleisön mielenkiinnon samaistuttavalla tarinalla hänen äitinsä ensimmäisestä pyykinpesukoneesta. Sitten Rosling edistää tarinan perspektiiviin maailmanlaajuisella tasolla nykyhetkeen, viitaten sähkölaitteisiin mitä rikkaammat omistaa ja mitä köyhemmät eivät. Kun Rosling on herättänyt tarpeeksi kysymyksiä ja mielenkiintoa, hän etenee esitellen kolme asiankeskeistä datasettiä visualisoituna. Hän tarkoin selittää statistiikan jokaisen arvoryhmän taustalle ja pitää datan rinnastettuna aiempaan tarinaansa konkretian säilyttämiseksi. Lisäksi Rosling tekee esitetyn datan avulla ennusteita talouden kasvusta talousluokkien välillä, sekä ottaa havainnoillaan kantaa kasvusta johtuviin riskeihin ja asettaa tavoitteita tilanteen korjaamiseksi. Lopuksi Rosling vielä palaa äitinsä ensimmäiseen pyykinpesukoneeseen ja mainitsee teknologian yleistymisen lisäävän koulutusta, lopettaen puheensa positiivisella mielellä. Tarina siis kulkee luonnollisesti eteenpäin tarinan lailla, eikä katsoja välttämättä huomaa edes suurta datan määrää, joka kulkee tarinan taustalla. (Rosling, 2011.)

Joissain visualisoinneissa dataa käytetään vain tukemaan päätelmien todenmukaisuutta ja on siten osana esitelmää. Printissä, kuten lehtiartikkeleissa datan näyttö oheistuotteena on yleistä. Esimerkiksi The New York Timesin erikoisartikkeli The Changing American Family, joka sisältää lähes 10 000 sanaa erilaisista Amerikkalaisista ydinperheistä. Artikkelin rinnalla faktoja tukee yhdeksän yksinkertaista diagrammia aiheeseen liittyvästä datasta. Artikkelin kirjoittaja nojaa tietyissä kohdissa datan tarjoamiin faktoihin, eikä pidä dataa artikkelinsa pääasiana. (Angier, 2013.)

2.2 Yleisölle suunnitteleminen

Visualisoinnin suunnitteluvaiheessa on mahdollista ottaa huomioon tekniikoita, joilla pääasioiden kommunikointi olisi luonnollisempaa tietyissä tilanteissa. Robert Kosara ja Jock Mackinlay, Tableau Software–yhtiöstä laativat hyviä esimerkkejä eri tilanteista ja mitä näissä ottaa huomioon, jotta katsojan mielenkiinto pysyisi visualisoinnissa. Pääasiassa Kosaran ja Mackinlayn skenaariot voidaan jakaa kahteen. Ensimmäiseksi autonominen esitys, kuten esimerkiksi mediaesitys, tai artikkeli, joka tehdään kerran ja katsotaan useasti. Yleisöllä ei ole tällöin mahdollisuutta vuorovaikuttaa tai esittää kysymyksiä, vaan ideana on saada pääasiat yleisölle esille uskottavasti, totuudenmukaisesti ja selvästi, jotta esillepantavat faktat olisivat helppo ymmärtää. Toisena skenaariona esitys elävälle yleisölle, joka voi olla joko palaveri, tai julkinen esitys, vaihdellen pienistä yleisöistä suuriin yleisöihin. Tässä tapauksessa on ensimmäiseen skenaarioon verrattuna tärkeämpää kaapata yleisön mielenkiinto, sekä yleisön kokoluokasta riippuen vaihtelevasti vaikeampaa vuoro-

vaikuttaa yleisön kanssa, mutta silti mahdollista ja kannattavaa. Tietyissä tilanteissa data voi olla yleisölle jopa kokonaan avointa, joko verkossa tai elävän yleisön edessä. Verkkopalvelun tapauksessa datalle on esimerkiksi avoin rajapinta, jolla sitä pääsee käsittelemään. Elävän yleisön kuunnellessa visualisoinnin esittäjä voi tulkita tietyn osakohdan hyvinkin tarkasti, johon esittäjä saa välitöntä palautetta yleisöltä. Tässä tilanteessa on myös mahdollista, että jokin seikka tulee luonnostaan esille, jota ei esityksessä alun perin ole ollut edes mukana. (Kosara & Mackinlay, 2013.)

Visualisoinnin ominaisuuksien suunnittelua voi helpottaa yleisölle suunnitellessaan alustavilla kohderyhmäarvioilla. Teoreettiselle yleisölle suunnitteleminen muokkaa visualisointia lopputulokseen, joka miellyttäisi paremmin visualisointityön mahdollista katselijaa. Jonathan Corum, hänen kollegansa Shan Carterin kanssa luovat The New York Times -lehdelle ja kyseisen lehden verkkopalvelulle visualisointeja tietynlaiset tekaistut yleisöt mielessään. Corum kertoo heidän ympärillä pyörivistä kohderyhmistä ja niiden vaikutuksesta hänen visualisointeihinsa, puheessansa datan visualisoinnin Tapestry–nimisessä konferenssissa Yhdysvalloissa. Karkeimmalla tasolla The New York Times -lehden lukijat ovat joko usein selailevia ja lojaaleja tilaajia tai kertaluontaisen kävijäsuhteen tarjoavia vierailijoita. Suunnitteluvaiheen valinnoissa voi myös vaikuttaa tieto käyttäjien jakautumisesta lehden lukijoihin, verkkovierailijoihin tai vaikka mobiililaitteilla selaajiin. Shan Carter, The New York Times -lehden graafinen suunnittelija saa teoreettisiin kohderyhmiin persoonallisuutta mukaan, ajattelemalla työskentelevänsä Bart ja Lisa Simpsonille. Bart edustaa lyhytjänteisyyttä, nopeaa tietoa ja yleiskatsausta haluavaa persoonaa, kun taas Lisa haluaa viettää visualisoinnin kanssa aikaa, nähdä sitä eri näkymästä, tutkia, tarkastella ja ymmärtää sen takana olevaa dataa. Jonathan Corum, The New York Times -lehden tiedealan graafinen suunnittelija, pitää mielessään tieteestä kiinnostuneen nuoren opiskelijan, joka voi vielä lukea printtiä tai selata sitä mobiililaitteella, sekä vaatii kiinnostavan esityksen. Toisena katselijana Corum harkitsee vanhemman ihmisen, joka keskittyy työn kokonaisuuteen ja estetiikkaan. Nämä työtavat huomioivat katselijoiden ikähaarukkaa, mielenkiintoa, keskittymiskykyä ja käsitellyn aihealueen kokemusta, jota lukijalla voisi olla. (Corum, 2013.)

2.3 Tutkimuksellisen data-analyysin visualisointitekniikka

Tutkimuksellinen data-analyysi on suositeltu lähtökohta silloin, kun mielessä ei ole vielä mitään mitä datasta voisi lähteä visualisoimaan. Prosessilla pyritään analysoida raakaa dataa pääasioiden yhteenvedon saamiseksi, usein visuaalisilla metodeilla, käyttäen statistisia malleja. Prosessilla toivotaan tuovan esille kehityskelpoisia hypoteeseja tarkkailun aiheena oleville tapahtumien syille. Tutkiva analyysi koostuu datan rakenteiden ja trendien esilletuomisista ja niiden avainasioiden tutkimisesta. Tavoitteena on synnyttää olettamuk-

sia sille mihin visualisoinnissa käytettävät päätelmät tulevat mahdollisesti pohjautumaan. Kun esiteltävät tulokset ovat vahvistettuja, lopulta tuodaan pääasiat viimeisimmässä, korkealaatuisemmassa tuotteessa selvästi esille. Syntyvä lopputulos, eli selityksellinen visualisointi on ennestään tiedetyn, toteutuneen hypoteesin selittämistä. Tässä vastataan kysymykseen datalla ja perustellaan pääasiat ja kommunikoidaan havainnot katsojalle. (Behrens, 1997; Fayyad, Wierse & Grinstein 2002, 22.)

John Tukey, Yhdysvaltalainen matemaatikko, yleisti kyseisen termin 70-luvulla kirjoittaessaan kirjan aiheesta nimeltään Exploratory Data Analysis. Termi lähti käytäntönä nousuun tietokonerajapintojen yleistyessä, milloin data saatiin helpommin käsiteltyä tietokoneiden avulla. Dataa oli mahdollista soveltaa helpommin eri diagrammeihin ja laskentoihin, joiden avulla dataa pääsi katselmoimaan statistisesti kokeilumielessä eri tavoilla. Myöhemmin tätä voitiin toteuttaa isoimmilla dataseteillä ja useimmilla muuttujilla. John Tukey esittää kirjassaan tutkimuksellisen data-analyysin olevan enemmän ajattelutapa, eikä pelkkä työkalu. (de Smith, 2015.)

Tutkimuksellista data-analyysiä voi ajatella visualisoinnin elinkaarena tai osana isompaa tutkimusta. Kerättyä dataa tutkimalla pyritään saamaan esille päähavainnot tai ongelmat ja niiden ympäröivät yksityiskohdat. Tämä suoritetaan esimerkiksi järjestelemällä tai ryhmittelemällä dataa ja analysoimalla sitä eri statistisissa mallinnuksissa. Päämääränä ovat lisäkysymykset, tai jopa vastaukset kysymyksiin, joita esiintyy tutkiskellessa dataa. Tutkimuksellinen data-analyysi kehottaa tekijää tutkimaan dataa kokeilemalla ja sovittamalla sitä erilaisiin malleihin, josta mahdollisesti koituu lisähavaintoja. On myös mahdollista päätyä tilanteeseen, jossa kysymyksille tarvitaan lisäanalyysiä, joka toistavana prosessina johtaa uuden, syventävän datan keräykseen ja sen tutkiskeluun. Kun tekijän tutkimuksellinen analyysi on ohi, tietää hän käsiteltävästä datasta paljon enemmän. Tästä lähtökohdasta on helpompi jatkaa informatiiviseen osuuteen, jossa esitetään asiat, jotka olivat tutkimuksen aiheena. (Behrens, 1997; de Smith, 2015.)

Tutkimuksellista dataa voi myös tarjota yleisölle palveluna. Esimerkiksi Googlen tarjoama palvelu Public Data, jossa katsoja pääsee tutkimaan eri kolmansien osapuolten tarjoamaa avointa dataa yhdessä paikassa. Google on palvelussaan jo sovittanut avoimen datan helposti tulkittaviin diagrammeihin, joita pääsee tutkimaan aihealueittain. (Google, 2016.)

3 Visualisointiesimerkkien havainnointi ja analyysit

Tässä kappaleessa on koottu valittuja käytännön esimerkkejä visualisoinneista, joita katselmoidaan aiheittain. Kappale muun muassa koskettaa visualisoinnin historiaa, sekä kertoo yksittäisesti siitä, miten menneisyydestä kerätyn tapahtuneen datan lisäksi on mahdollista käyttää myös muun laadun dataa visualisoinneissa. Jokaisessa aiheessa käydään käytännön esimerkkejä läpi eri näkökulmista, hyvin yksityiskohtaisella tasolla. Tämän kappaleen aliotsikoissa on luotu omia diagrammeja esimerkin datan vahvistamiseksi, sekä näyttämään dataa eri perspektiivistä. Kappale tuo edetessään konkretiaa enemmän esille tulevaa kappaletta varten, sekä korostaa visualisoidun datan avulla havainnollistamista. Kappale sisältää tekstin ohella useita kuvia, joiden havainnointi on tärkeä osa kappaletta. Kuviin viitataan kappaleessa useasti, josta johtuen ne ovat selkeyden ja käytettävyyden takia asetettu tiettyyn kohtaan tekstiä.

3.1 Francis Anscomben kvartetti

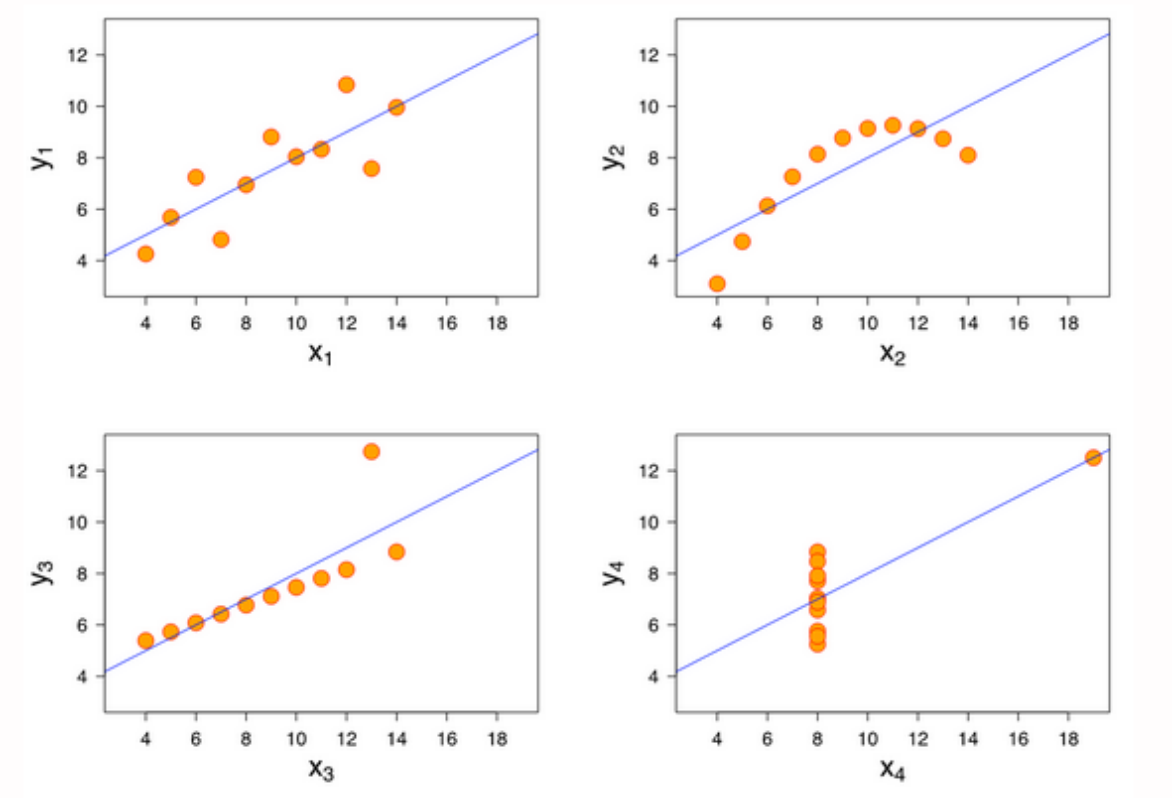
Francis Anscomben vuonna 1973 suunnittelema arvojoukko nimeltään Anscomben kvartetti, on yksi yleisimmistä ja tunnetuimmista esimerkeistä tutkimuksellisen visualisoinnin (engl. exploratory data analysis) hyödyllisyydestä. Kvartetin tavoitteena on esittää arvotaulukkojen erilaisuus visuaalisesti diagrammeilla, vaikka arvoilla on useita statistisia samanlaisuuksia. Francis Anscombe näyttää esimerkillään sitä, että useassa tapauksessa datan visualisointi on hyödyllinen data-analyysin työkalu statistisen tulkinnan lisäksi. (Anscombe, 1973.)

Anscomben kvartetti sisältää neljä arvojoukkoa, joista jokaisesta löytyy yksitoista arvoparia. Matemaattisesti arvojoukoista löytyy paljon yhtenäisyyksiä tutkimalla niitä esimerkiksi keskiarvoilla, variansseilla ja regressiomenetelmillä. Luvut eivät vaihtele suuresti, sekä useampi laskenta lukujen välillä on hyvin samanlaista. Muun muassa x-arvojen keskiarvo jokaisessa joukossa on tasan 9 ja y-arvojen noin 7.50, sekä laskettava korrelaatio x- ja y-arvojen välillä on noin 0.816 jokaisen joukon kohdalla. (Taulukko 1.)

Taulukko 1. Anscomben kvartetti (Giles, 2011.)

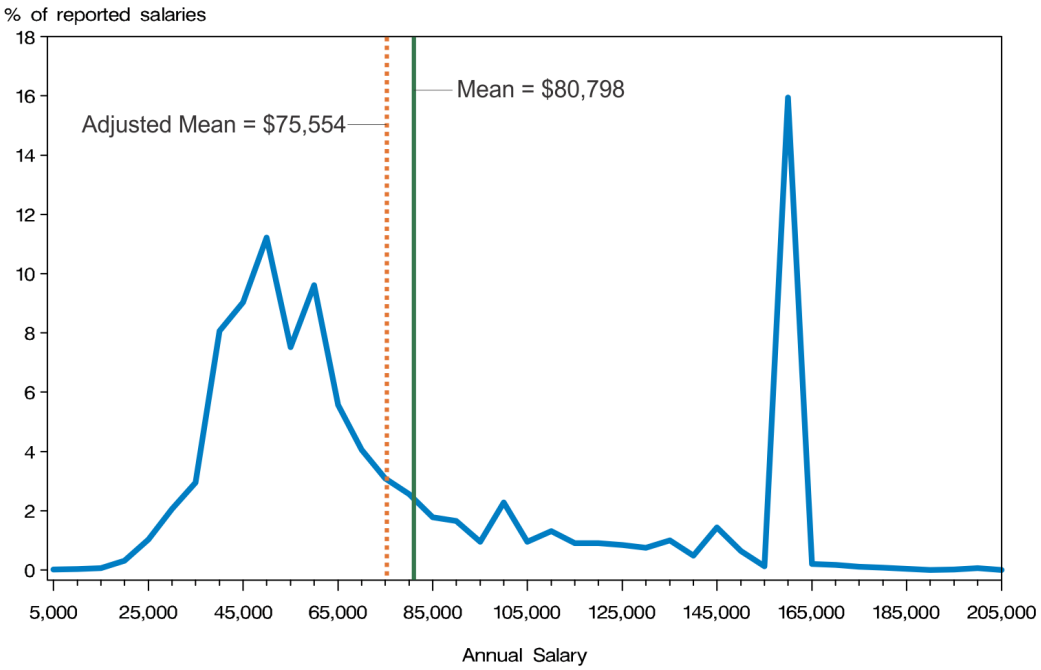
I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,1	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,1	4	5,39	19	12,5
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

Kvartetin arvoja visualisoidessa arvojoukot näyttävätkin hyvin erilaisilta. Jokainen arvojoukko käyttäytyy diagrammissa eri tavalla, vaikka ne esiintyivät samantapaisina taulukossa laskettuna. Diagrammeissa ainoastaan lineaarinen regressiosuora on jokaisessa arvojoukossa lähes identtinen. (Kuvio 1.)



Kuvio 1. Anscomben kvartetti visualisoituna (Parikh, 2014.)

Ravi Parikh, Kalifornialaisen Heap Analytics–yrityksen analyytikko ja bloggaaja, esitti 2014 laatimassaan artikkelissa Anscomben kvartetin esimerkillisen tarkoituksen oikean elämän dataesimerkillä. Parikh oikoi National Association of Law Placementin (NALP) 2013–vuoden raporttia, joka ilmaisi, että vuonna 2012 Yhdysvaltalaiset juristit saivat alkupalkkaa vuositasolla keskiarvoltaan 80 789 dollaria. Dataa visualisoidessa (Kuvio 2.) Parikh huomasi muun muassa, että NALP–yhdistyksen antama keskiarvo on vääristävä, sillä palkansaajat jakautuvat kahteen pääasialliseen osioon.



© NALP 2013
www.nalp.org

Kuvio 2. Yhdysvaltojen juristien vuosipalkan havainnointia (Parikh, 2014.)

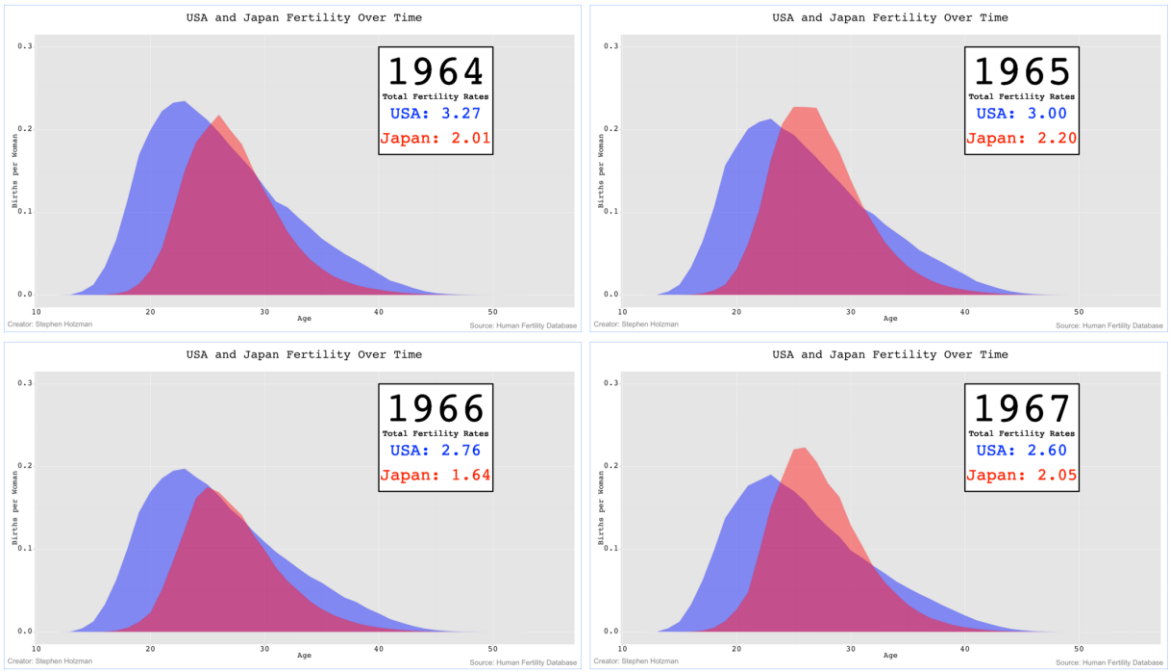
Suurin osa uusista juristeista tienaa alkupalkakseen vuositasolla noin 35 000 – 75 000 dollaria, joka esiintyy kuvio 2:ssa ensimmäisenä piikkinä. Kuvio 2:n mukaan usea uusi juristi tienaa vuositasolla 160 000 dollaria, joka skaalaa keskiarvoa NALP–yhdistyksen raportoimaan 80 798 dollariin. Parikh mainitsee artikkelissaan myös samaa, mitä Francis Anscombe alkuperäisessä vuoden 1973 artikkelissaan: Dataa on hyvä tutkia matemaattisilla laskennoilla, mutta visualisoimalla sitä on mahdollista löytää tärkeitä lisähavaintoja. (Anscombe, 1973; Parikh, 2014.)

3.2 Menneisyys, historiallisen datan havainnointi ja raportointi

Silloin kun ei ole kyse reaaliaikaisesta datan esittämisestä tai ennusteista, niin visualisointi on käytännössä historiatietojen deskriptiivistä raportointia. Raportoinnin prosessilla tehdään raa'asta datasta hyödyllisempää ryhmittelemällä ja järjestämällä se useisiin käsitel-

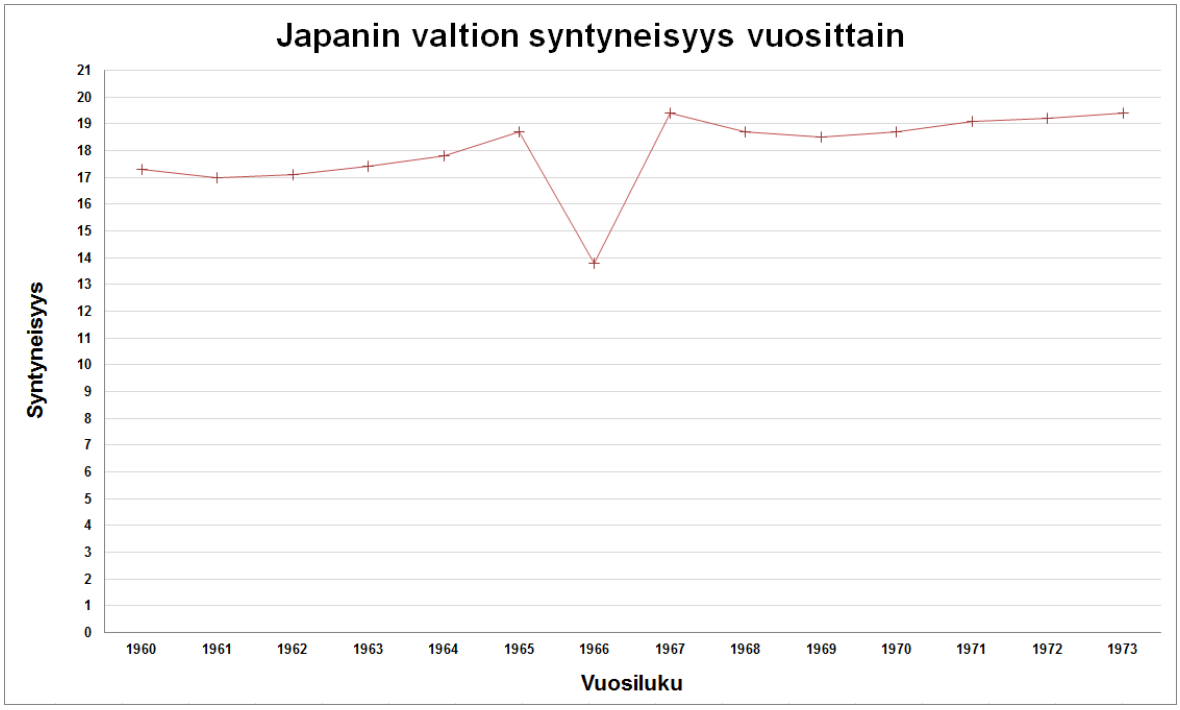
täviin informaatiotiivistelmiin esimerkiksi summaamalla, ryhmittelemällä, optimoimalla, pisteyttämällä, aggregoimalla tai muilla lukuisilla analyyttisillä tai statistisilla tavoilla. Näin datasta saadaan avainluvut esille kerta-analyyseinä projektin tai yrityksen tarpeisiin nähden. Informaatiotiivistelmät kootaan esitettävään muotoon, josta syntyy diagrammi tai visualisointi. Visualisoidusta datasta tehdään havainnointeja ja haetaan vastauksia tapahtuneelle, sekä niistä syntyviä johtopäätöksiä käytetään mahdollisesti tulevaisuudessa päätöksenteossa. Visualisoinnilla siis kerrotaan tietty tapahtunut asia, joka perustuu faktaan, eli aihetta ympäröivään kerättyyn dataan. (Corum, 2013; Dykes, 2010; Fayyad, ym. 2002, 23.)

Vuoden 2015 heinäkuussa, tietotekniikan bloggaaja nimeltä Stephen Holzman jakoi visualisointianimaation Reddit–nimisessä sosiaalisessa mediassa. Reddit–sivuston Data Is Beautiful–nimisessä aiheosiossa on monipuolinen yleisö, joka koostuu 5,7 miljoonasta lukijasta. Lukijakunta vaihtelee tutkiskelevista tohtoreista, kiinnostuneisiin kouluopiskelijoihin ja hienoa ”datataidetta” etsiviin sisällön selaajiin. Holzmanin visualisointi, esitetty kuviossa 3, koostaa dataa Human Fertility Database–projektin avoimesta datasta, joka sisältää maailmanlaajuista tietoa synnyttäneistä naisista monen vuoden takaa. Animaatiossa näkyy vertailussa Japanin ja Yhdysvaltojen, vuosien 1947–2010 välillä synnyttäneiden äitien ikä x-akselilla ja valtion kokonaishedelmällisyysluku y-akselilla. Kokonaishedelmällisyysluku tarkoittaa vuosittaista naiselle keskimäärin syntyvän lapsimäärän odotetta. (Holzman, 2016; Human fertility database, 2016.)



Kuvio 3. Yhdysvaltojen ja Japanin kokonaishedelmällisyysluvut vuosina 1964–1967 (Holzman, 2016.)

Visualisoinnissa esiintyy muutama selkeä poikkeavuus, joka herättää erityistä mielenkiintoa. Esimerkiksi vuonna 1964 ja 1965, jotka näkyvät kuviossa 3, Japanin hedelmällisyysluku nousee arvosta 2,01 arvoon 2,20 ja laskee vuonna 1966 arvoon 1,64, sekä lopulta nousee takaisin suhteellisen tasaiseksi arvon 2 yläpuolelle. Tämän tyyppinen ponnahdus voi kyseenalaistaa datan laatua ja kaipaa selvästi lisätutkimusta. Yllättävä lasku hedelmällisyysluvussa on tutkittu johtuvan joka kuudeskymmenes vuosi ilmestyvästä Kiinalaisen tulihevosen horoskoopista. Vuotta 1966 pidettiin horoskoopin takia huonon onnen vuotena, joka näkyy huomattavana laskuna Japanin syntyneisyydessä. (Bruckner, Catalano & Subbaraman, 2011.)



Kuvio 4. Japanin valtion syntyneisyys vuosittain

Vertailuna kuvio 4, joka esittää Japanin valtion syntyneisyyden vuosista 1960–1973. Kuviossa 4 data on eri lähteestä, sekä hieman eri näkökannasta. Data on saatu Yhdistyneiden kansakuntien erityisjärjestön Maailmanpankin avoimen datan arkistosta ja diagrammi on tehty Excel-taulukkolaskentaohjelmalla. Y-akselilla esitetty syntyneisyys tarkoittaa kyseisen x-akselilla esiintyvän vuoden aikana syntyneiden lasten keskiarvoa tuhatta asukasta kohden. Datasta näkyy selvästi vuosien 1965 ja 1966 välinen poikkeavuus, jolloin Japanin syntyneisyys tippuu arvosta 18,7 arvoon 13,8 ja nousee vuonna 1967 arvoon 19,4. (The world bank, 2016.)

Holzmanin tekemän animaation avulla on vaikea seurata dataa vuositasoisen rinnakkaisvertailun kannalta, eikä yksityiskohtainen havainnointi juurikaan ole mahdollista. Käytän-

nössä pelkkä tilanteen eteneminen on tässä animaation formaatissa keskeisimpänä, eikä sitäkään ole mahdollista pysähtyä tarkastelemaan. Kehnojen suunnitteluvalintojen jälkeen Holzman päivitti Reddit–yhteisöstä saaneen vuorovaikutuksen ansiosta blogisivustollensa visualisoinnista uuden version, jossa katselija voi edetä aikajanalla vapaasti. Lisäksi kahden ennalta valitun valtion lisäksi visualisointiin voi valita 26 eri valtiota, joista neljä valtiota on mahdollista näyttää samanaikaisesti. (Holzman, 2016.)

3.3 Nykyaika, dynaamisen datan esittäminen

Dynaaminen data, eli tietyistä tapahtumista muokkautuva, reaaliaikainen järjestelmässä kulkeva informaatio, joka esitetään välittömästi tai pienellä viiveellä keräyksen yhteydessä. Dynaamisen datan visualisoinnilla pyritään antamaan reaaliaikaista kuvaa järjestelmän kokonais- tai osaprosessin nykytilanteesta. Yksi yleinen reaaliaikainen visualisointiratkaisu, Performance Dashboard, on eräänlainen mittaristo, joka koostuu esimerkiksi yhtiön avainluvuista, tai järjestelmän suoritusluvuista. Näitä on esimerkiksi sisäänrakennettu useisiin CRM–järjestelmiin, tai räätälöity ratkaisuina analytiikan toimijoiden johdosta projekti- tai järjestelmätasoisesti. (Schacter, 2010.)

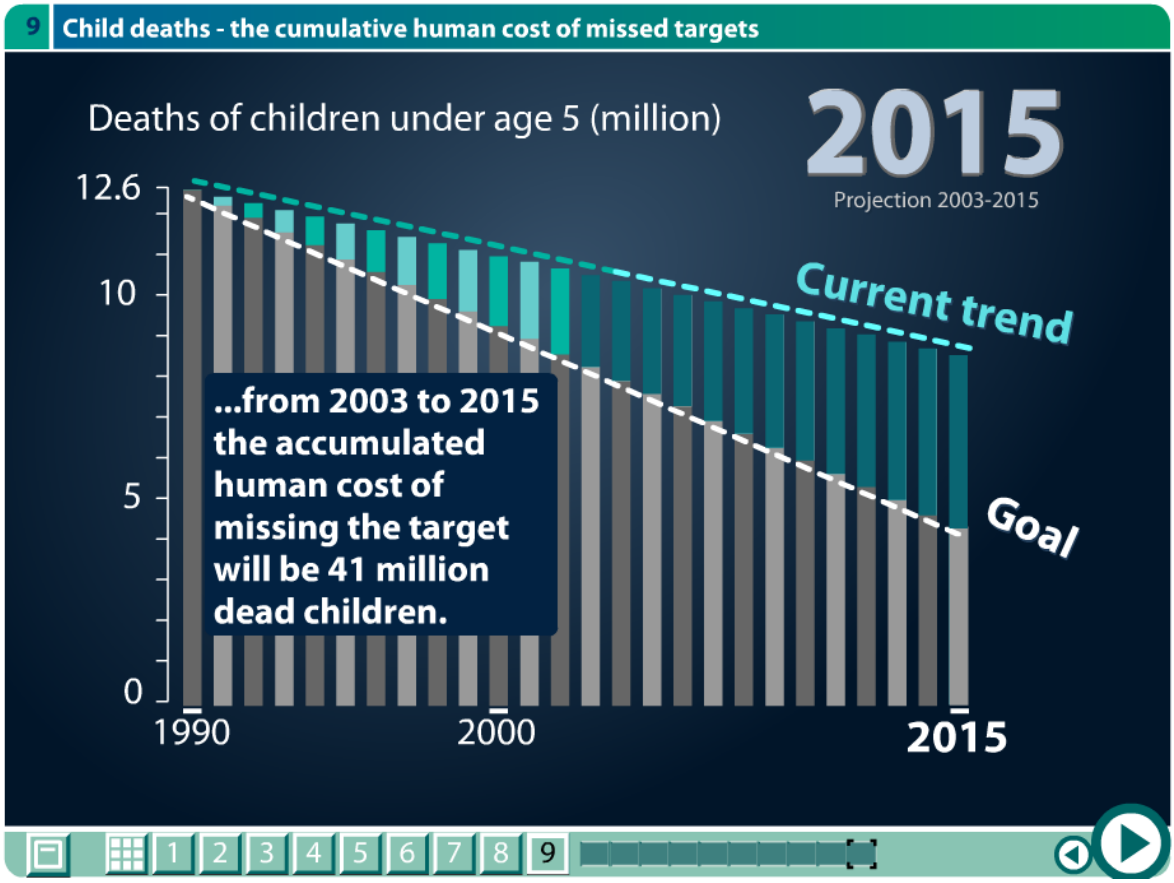
Reaaliaikaisen datan visualisoinnilla on mahdollista nähdä esimerkiksi nykyhetken palvelindatan lähetys ja vastaanotto mittarityyppisellä ratkaisulla, joka voi tuoda esiin mielenkiintoisia havaintoja tietoliikenteestä. Etan Lightstone, käyttökokemuspäällikkö Yhdysvaltalaisesta analytiikkayrityksestä nimeltä New Relic, esittelee vuonna 2013 julkaistussa seminaarissaan, miten reaaliaikaista verkkodataa on mahdollista visualisoida. Kuviossa 5 on mitattu kuvitteellisen sovelluspalvelimen sarjaliikennettä kahden minuutin ajan. Sovelluspalvelimen yhteyskutsuja käsittelevät muuttujat demonstroidaan reaaliajassa eteenpäin ajassa kulkevilla, sinertävillä, läpikuultavilla nauhoilla. (Lightstone, 2013.)

Kuvio 5. Sovelluspalvelimen yhteyskutsujen vastausajat (Lightstone, 2013.)

Kuvion 5 yhteyskutsut käsittelevät ruokareseptiapplikaatiota, jota käyttäjät käyttävät joko listaamalla kaikki reseptit, näyttämällä yhden niistä tai lisäämällä uuden. Visualisoinnissa nähdään myös päällekkäisiä käsiteltäviä yhteyskutsuja. Esimerkiksi kuvion 5 yläosassa on näkyvillä läpikuultavien nauhojen päällekkäisyydestä koituva tummempi osa, kun viisitoista sekuntia pitkä palvelinkutsu menee kaksikymmentä sekuntia pitkän kutsun päälle. Nauhojen läpikuultavuuden suunnitteluvalinta tuo esille erinomaisesti nämä päällekkäisyydet, jonka ansiosta suuret tietoliikenneruuhkat herkästi korostuvat tummempina osina nopeaa havainnointia varten. (Lightstone, 2013.)

3.4 Tulevaisuus, estimointi ja kehityssuunnat

Ruotsalaisen Gapminder–säätiön 2005 vuonna julkaisema visualisointi Human Development Trends on sisällöltään hyvin laaja visualisointi. Tässä moniosaisessa diasarjan kaltaisessa esityksessä käytetään tarkasti suunniteltua rakennetta ja useita datan visualisoinnin käytäntöjä. Pääasiassa Gapminder–säätiön visualisointi kuvastaa globaalin kehittymisen trendejä raha-ansiossa ja terveydessä. Visualisoinnin viimeisessä diagrammissa (Kuvio 6) Gapminder estimoi Yhdistyneiden Kansakuntien esittämää globaalin tason kehityssuuntaa lasten kuolematapauksien estämiseksi. (Gapminder, 2008.)



Kuvio 6. Osa Gapminder–säätiön tekemästä, globaalia kehitystä esittävässä visualisoinnista (Gapminder, 2008.)

Esityksessä käytetään lukuja vuodesta 1990 vuoteen 2003 pohjana estimoinnille. Esityksessä puhutaan tosin vain suhdanteesta, eli luultavasti näytettävä data perustuu faktojen sijasta pyöristettyihin, karkeisiin arvioihin. Diagrammissa näkyvien palkkien mukaan kymmenessä vuodessa päästiin 12.6 miljoonasta 11 miljoonaan. Yhdistyneiden Kansakuntien tavoitteen mukaan vuonna 2015 luku tulisi olla 4 miljoonaa, jonka visualisoinnissa on asetettu vierekkäin nykyisen tason rinnalle. Diagrammissa painotetaan, että pelkästään vuonna 2003 YK:n asettaman tavoitteen ja oikean tilanteen välissä on kaksi miljoonaa. Lisäksi vuosien 2003–2015 YK:n tavoitteen kehityssuunnan, sekä todellisen tason estimoitu kumulatiivinen erotus on 41 miljoonaa. (Gapminder, 2008.)

3.5 Saman datasetin eri visualisointimenettelyt

Visualisointia on aina mahdollista parannella ja kehittää ulkomuodoltaan tai käytettävyydeltään. Joissain tapauksissa datasta on vaikea tehdä havainnointeja, jolloin voi olla mahdollista näyttää sitä eri muodossa. Usein tietyn tarkoituksen mukaista havainnointia varten visualisoinnin suunnittelu vaatii ajattelemista eri näkökannoilta, jotta yleisö saa tarkemman vastauksen tiettyyn kysymykseen.

Metrokartat voivat olla yksi esimerkki siitä, miten käytettävän informaation käyttötarkoitus vaikuttaa esitystapaan. Metron kulku on hyvin suoraviivainen. Metrojunan kyytiläinen valitsee joko poistuvansa metrosta junan pysähdyttyä, tai olla kyydissä vielä esimerkiksi kahden pysäkin ajan. Kyytiläinen laskee matkustaessaan pysäkkejä, eikä tarkista esimerkiksi geologisesti korrektista kartasta olinpaikkaansa. Oikean kartan sijaan skemaattinen kuviointi metron pysäkeistä on katsottu tähän ehkä käytännöllisyydeltään parhaiten sopivammaksi. Metronjunan kyydissä on tärkeintä havainnoida matkan päämäärän, eli pysäkin sijainti verrattuna nykyiseen sijaintiisi. Joissakin metrokartoissa on pysäkkien väleihin merkitty välimatkan keskimääräinen aika. Tällä tavalla matkustamiseen käytettävä aika olisi hahmotettavissa helpommin.

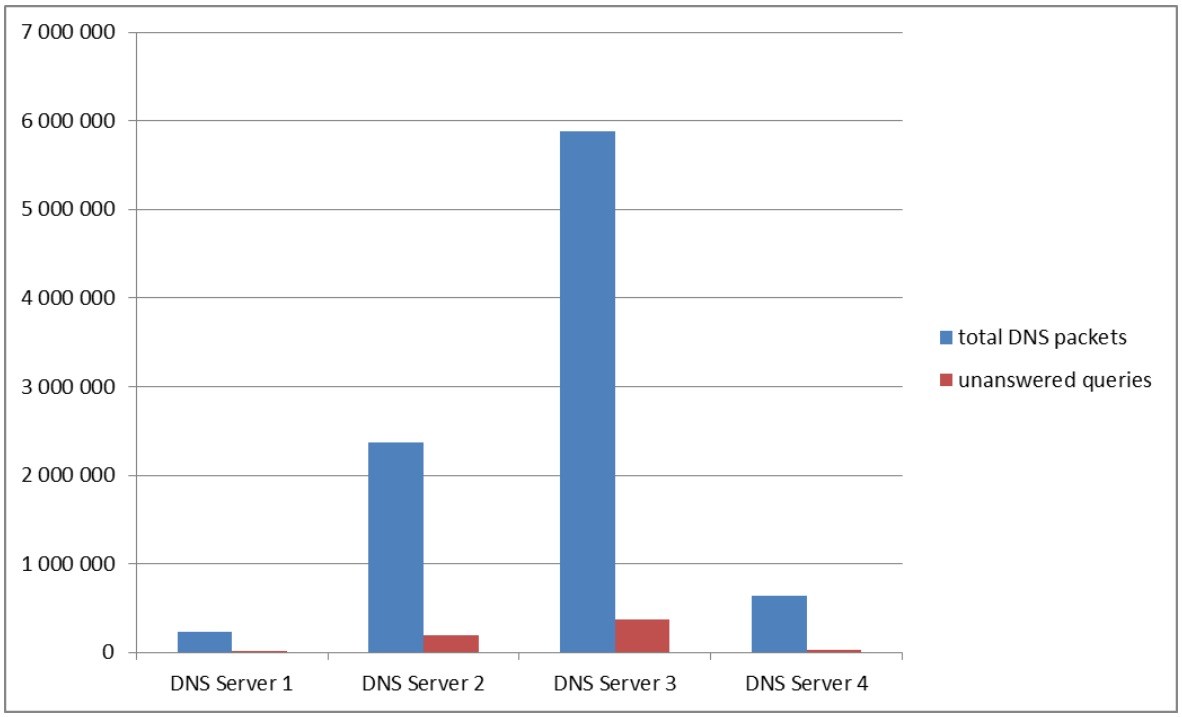
Metron ratojen visualisoiminen geografisella kartalla voi toki palvella myös tiettyä käyttötarkoitusta. Geografinen metrokartta voi paljastaa esimerkiksi reitin levinneisyyden ja toisaalta alueet, joihin metrolla on vaikea päästä. Geografisella metrokartalla puolestaan tiettyyn sijaintiin kulkiessa on helpompi havainnoida oikeiden metrojunien ja metropysäkkien yhdistelmä, jos alue on hieman tuntemattomampi. Kartan tapauksessa keskustan pysäkit voivat olla vaikeasti luettavissa pysäkkien tiheyden vuoksi. Tiettyjen pysäkkien väliset matkat ovat myös helposti verrattavissa. Geografisella metrokartalla voi samaistua siihen minkälaisen matkan metrot kulkevat kartalla.

Joskus visualisoitava data näyttää huonolta tai on hankala saada selkeään muotoon. Jos tavallinen diagrammi ei esitä dataa tarkasti ja selkeästi, niin voi olla että on olemassa toinen diagrammi joka tekee sen. Esimerkkinä taulukko 2, jossa arvotaulukkoina on neljä eri kuvitteellista DNS–palvelinta, joilta on kerätty vastaanotetut DNS–paketit ja vastaamattomat DNS–kyselyt. Datan otokset ovat yhdeltä kuukaudelta ja kyseiset neljä eri palvelinta ovat samasta toimipisteestä. Analysoinnin kohteena on palvelimille tulevien DNS–kyselyihin vastaamiskyky, sekä kehitystarpeena on kuormituksen normalisoiminen tasaiseksi neljälle palvelimelle. Haasteena on datan visualisoiminen siten, että diagrammissa pysyisi datan eheys ja että palvelimen väliset suhteet tai niiden eroavaisuudet korostuisivat.

Taulukko 2. Neljän kuvitteellisen DNS–palvelimen tiedonsiirtodatat.

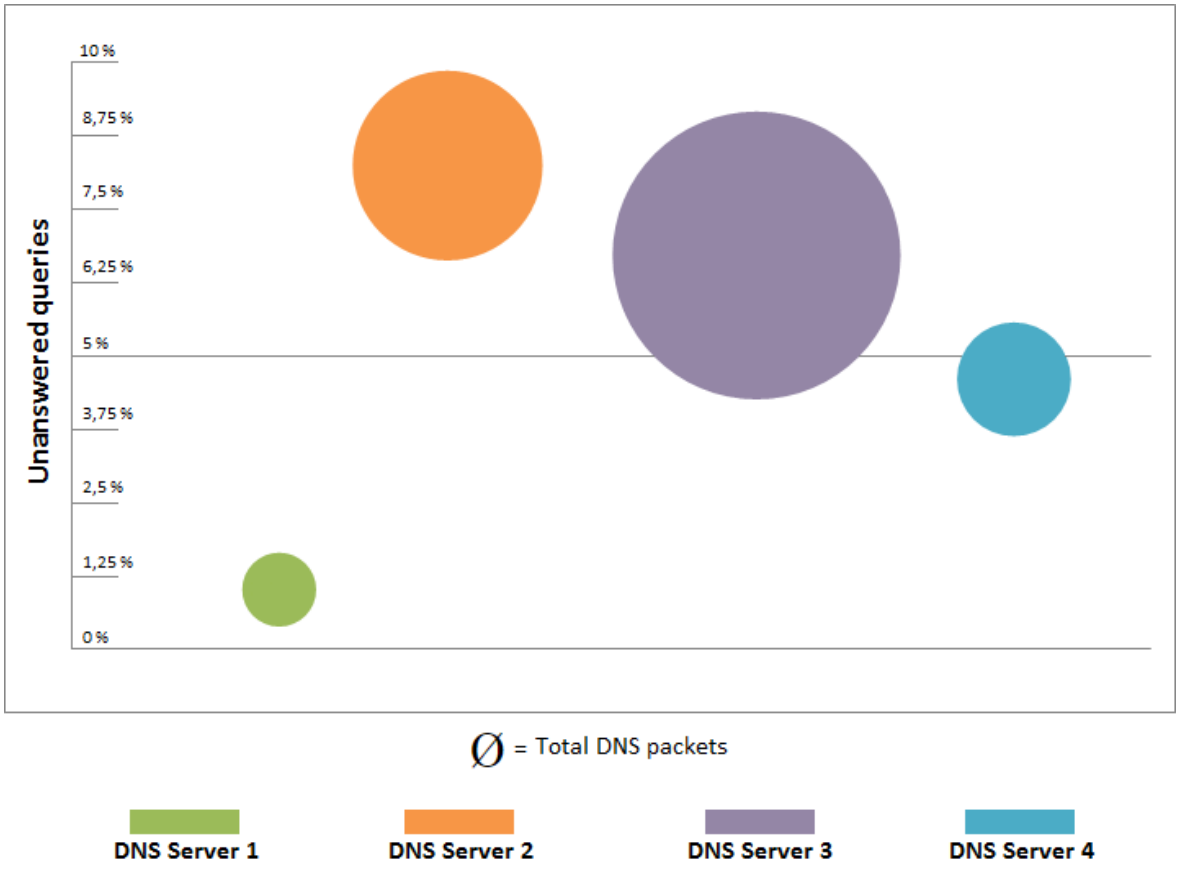
	DNS Server 1	DNS Server 2	DNS Server 3	DNS Server 4
total DNS packets	228 883	236 5875	5 885 995	642 427
unanswered queries	2 475	195 537	374 859	29 987
percentage	1,0813385	8,264891425	6,368659844	4,6677677

Data on kuviossa 7 esitelty kaikille tutussa pylväskaaviossa. Dataa esittäessä tässä muodossa huomataan, että parametrien vaihtelevaisuus suhteessa toisiinsa on todella korkea. DNS–pakettien suuri määrä skaalaa diagrammin niin korkealle, ettei pieniarvoiset DNS–kyselyt tule esille käytännössä ollenkaan. Molemmat parametrit ovat tässä tapauksessa avainarvoja, jotka halutaan ehdottomasti sisällyttää visualisointiin. Jos tätä dataa haluttaisiin näyttää pylväsdiagrammissa, joutuisi DNS–paketit ja vastaamattomat kyselyt jakaa kahteen erilliseen diagrammiin jotta pylväiden keskeinen skaalaus ei ilmentyisi näköhaittana. Kahteen erilliseen diagrammiin jakaminen vaikeuttaisi palvelinten vertailua kokonaisuutena.



Kuvio 7. Taulukossa 2 esitettyjen DNS–palvelimien visualisointi pylväsdiagrammilla

Kuviossa 8 saman datan esittämiseen on käytetty epävirallista ilmapallokilpailudiagrammia (engl. Balloon Race). Jokainen ilmapallo tai kupla esittää eri DNS–palvelinta omalla värikoodillaan, kuplan halkaisija osoittaa vastaanotettujen DNS–pakettien totaalimäärää ja pystysuuntainen sijainti diagrammissa merkitsee vastaamattomien kyselyiden suhdetta pakettien totaalimäärään verrattuna. Diagrammi näyttää hillitymmin palvelinten suhteellisen vaihtelevaisuuden ja on lukujen kannalta tarpeeksi tarkka nopeaan havainnointiin.



Kuvio 8. Taulukossa 2 näkyvien DNS–palvelimien data ilmapallokilpailudiagrammissa

Usein alueellisesti tai suuruudella ilmaistut arvot liioittelevat dataa tai ovat virheellisesti suhteutettuja. Myös näissä tapauksissa ihmisen näköaisti voi vääristää näkyviä objekteja ali- tai yliarvioivasti. Kuitenkin nopeissa vaikutelmissa ne ovat usein tehokkaita. Kuvio Y:n avulla voidaan havainnollistaa, että DNS–palvelimet 2 ja 3 ovat ruuhkaantuneimpia, DNS–palvelin 1:llä on kapasiteettiä vapauttaa muilta palvelimilta kaistaa, sekä DNS–palvelin 4 näyttää olevan ihannetilanteessa, jota halutaan muilla palvelimilla tavoitella.

4 Visualisointi Suomen rakennusdatasta

Projektiosuudessa on visualisoituna Väestörekisterikeskuksen julkaisema avoin rakennusdata, jonka avulla lopulta tehdään havaintoja Suomen rakennusten jakautumisesta. Projektiosuus käy läpi visualisointiprojektin toteutusvaiheita suunnittelusta toteutukseen. Visualisoinnin tarkoituksena on toimia esimerkkinä, kuten edellisten kappaleiden visualisoinnitkin, mutta enemmän käytännönläheisemmin. Tässä esimerkissä pelkän analysoinnin lisäksi on luotu katsaus visualisoinnin suunnitteluprosessiin ja toteutukseen. Alueellisen datan visualisointia ei ole käyty entisissä esimerkeissä läpi ollenkaan, jonka ansiosta se sopii hyvin projektiosuuteen.

4.1 Projektin tavoitteet

Tavoitteena on saada yleiskuva siitä, miten koko Suomen rakennukset jakautuvat alueellisesti. Lisäksi on tarkoitus pienentää mittakaavaa ja havainnoida kahden erilaisen kunnan rakennusten jakautuminen. Ajatuksena on saada tehtyä datasta tietokantaa käyttämällä massa, jonka perustalta voi lähteä suunnittelemaan ja luomaan visualisointia.

Visualisointiprojekti tulisi olla ulkomuodoltaan selkeä ja yhtenäinen, noudattaen tiettyä tyyliä. Visualisoinnin tulisi myös olla mahdollisimman yksinkertainen, sekä selvästi tuoda esille päähavainnot rakennusten sijaintidatasta. Lisäksi visualisointi tulisi olla suhteellisen nopea toteuttaa. On odotettavissa, että projektilta jää varaa parantamiseen, sekä kehitysehdotuksia varmasti tulee projektin edetessä eteenpäin.

4.2 Tekniikat ja työtavat

Tietokantaympäristönä toimii MySQL 5.6, InnoDB–tietokantamoottorilla ja tietokannan hallintajärjestelmänä toimii HeidiSQL 9.3. Visualisoinnissa käytettävän datan koordinaatit, välimatkat ja muu alueellinen vahvistaminen, sekä pisteiden välimatkojen mittaus on suoritettu Maanmittauslaitoksen Karttapaikka–sovelluksella. Visualisoinnin luonnissa on käytetty kuvankäsittelyohjelmaa GIMP 2.8.

Apuna ei ole käytetty minkäänlaista visualisointijärjestelmää. Järjestelmän avulla lopputuloksena olisi lyhyt prosessi, sekä käytettävillä olevien teknologioiden myötä erilainen visualisointi. Ilman järjestelmää projektin suunnittelun eri osuudet tulevat selkeämmin esille ja luominen on vapaampaa.

4.3 Käytettävän datan laatu

Visualisointiprojektissa käytettävä data perustuu pääasiassa Väestörekisterikeskuksen 22.4.2016 julkaistuun avoimeen rakennusdataan, joka on ladattavissa ilmaiseksi avoindata.fi–palvelussa. Rakennusdatan rakennusten yksilöllistämiseen on käytetty pysyvää rakennustunnusta, jota seuraa rakennuksen muut tiedot. Muuna tietona on rakennuksen sijaintikunta, maakunta, sekä rakennuksen kokonainen sijaintiosoite, jota voi olla yksi tai useampi. Useampi osoite löytyy niiltä, joilla on joko kulmaosoite tai kortteliosoite. Jokaisella rakennuksella on sijaintitiedon lisäksi karttakoordinaatti, eli pohjoinen ja itäinen ETRS-TM35FIN–karttaprojektion arvo. Jokaisella rakennuksella on datassa myös tieto rakennuksen pääasiallisesta käyttötarkoituksesta, joka koostuu kahdesta arvosta: 1, eli asuin- tai toimitilarakennus ja 2, eli tuotanto- tai muu rakennus. Rakennusmassassa on mukana kaikki Suomen rakennukset, joilla on koordinaattitieto, sekä omistavat osoitetietonaan vähintään postinumeron. Liitteessä 1 on esimerkki rakennusdatasta, jossa on esillä kolme eri rakennusta.

Datan laatua on tarkasteltu ennen datan käyttöönottamista projektin luonnissa. Tarkastelemalla yksittäisten kuntien sijaintidataa, esille tuli muutama huolenaihe. Esimerkiksi Kuopion pohjoisin rakennus on Nilsiästä hieman yli 20 km pohjoiseen ja noin 10 km Rautavaaran kunnasta etelään. Tämä sijoittaa kyseisen rakennuksen noin kuusikymmentä kilometriä Kuopion keskustasta koilliseen. Kuopion läntisin rakennus sijaitsee datan mukaan Suomen länsirannikolla sijaitsevan Närpiön kunnasta 150 km Pohjanlahdelle, joka viittaa hyvin vahvasti koordinaattidatan kirjausvirheeseen. Tämänlaista vaihtelevaisuutta voi esiintyä datassa ympäri Suomea. Vaihtelevaisuuksien syyt voivat johtua useista eri asioista. Esimerkiksi kirjaamisvirheitä, järjestelmäsekaannuksia tai inhimillisiä virheitä on voinut esiintyä kuntien rakennusvalvontaviranomaisten luovuttaessa tietoa Väestörekisterikeskukselle. Vaihtelevaisuudet esimerkiksi rakennusten kunta- tai postinumerotiedossa voivat johtua mahdollisista kuntaliitoksista tai kuntarajojen muuttumisesta.

Rakennusdatan sijaintitiedoissa käytetään Suomen ETRS-TM35FIN–tasokoordinaatistoa. Tasokoordinaatistojen karttaprojektiot kuvaavat kolmiulotteista aluetta kaksiulotteisella pinnalla, josta johtuen välimatkat voivat vääristyä. Lisäksi Suomen tasokoordinaatisto poikkeaa standardista niin, että se luo maan länsiosassa huomattavia mittakaavavirheitä. Virheistä huolimatta visualisoinnissa käytettävät alueet ovat suuntaa antavia, mutta mittaamalla varmistettuja. (JHS-suositukset, 2016.)

4.4 Sijaintidatan pohjustaminen ja suunnitteluvalinnat

Väestörekisterikeskuksen julkaisema avoin rakennusdata on karttakoordinaattien, sekä osoitetiedon perusteella pääasiallisesti sijaintitietoa, eli alueellista dataa. Alueellista dataa on käytännössä hankalampaa visualisoida muuten kuin kartan avulla. Koordinaatteja ei ole kuitenkaan tavoitteena projektoida koko Suomen tasolla yksittäisinä pisteinä kartalle, koska kyseisen sijaintidatan massa koostuu noin 3,5 miljoonan rakennuksen sijainnista. Näin laajan tason data voi olla huomattavasti helpommin katselmoitavissa, kun se on ryhmiteltyinä laajempien fyysisten alueiden rajojen sisäpuolelle. Laajan datan ryhmittely parantaa myös esityskelpoisuutta, koska määrällinen data esiintyy raa'assa muodossaan abstraktina. (Few, 2009.)

Yleisesti sijaintidatassa karttakoordinaatisto on assosioitu joidenkin arvojen kanssa, jotka riippuvat käytettävillä olevista tietueista. Tässä tapauksessa esimerkiksi karttakoordinaatilla voi raakadatassa katselmoida rakennuksen yksittäistä sijaintia maakunnassa, kunnassa tai postinumeroalueella. Datassa rakennuksilla on näihin alueisiin yhteys, joihin data on myös mahdollista ryhmittää. Mutta maakunnat ovat projektin visualisoinnin käyttötarkoitukseen hieman liian sopimattomat epämääräisten alueiden kokojen takia. Kunnat eivät myöskään sovellu, koska datan laatua tarkasteltaessa tietyn kunnan rakennukset voivat sijainnillisesti hajaantua hyvinkin laajasti. Postinumeroalueet olisivat sijainnillisesti erittäin tarkkoja, mutta tällä tavoin visualisointi koostuisi todella tiiviistä matriisista, sillä datan mukaan postinumeroalueita on melkein kuusituhatta. (Few, 2009.)

Suunnitelmana on luoda sijaintidatan karttakoordinaattitiedoilla omat fyysiset alueet, jotka ovat riippumattomia raakadatan arvoista. Tarkoitus on tehdä havainnointeja näiden räätälöityjen alueiden avulla Suomen rakennusdatasta.
Alueiden avulla yksittäisten pisteiden sijaan käsitellään alueellisesti ryhmitettyjä datakeskittymiä, jolloin havainnointi, vertailu ja käsittely on kevyempää. Karttadatan visualisoitavia alueita voidaan esittää joko koolla tai väreillä. Suuruuksilla ei tässä tapauksessa ole käyttöä, sillä se vaikuttaisi suoraan aluejakoihin, eikä siten ole järkevää tai edes mahdollista. Jaottelu tietyllä väripaletilla voi tarjota alueista hyviä havaintoja varsinkin, jos alueet ovat keskenään samankokoisia. (Few, 2009.)

4.5 Suunnittelu

Tavoitteena on jaotella koko Suomen rakennukset viiteentoista poikittaiseen, korkeudeltaan samanlaiseen liuskaan, joiden sisällä yksittäiset rakennukset ryhmiteltäisiin. Tällöin käytännössä tiettyyn liuskaan kohdistuvat yksittäiset rakennukset saataisiin yhteenlaskettua fyysisen alueen sisälle. Käsiteltävissä olisi viisitoista kappaletta alueellisia massoja,

3,5 miljoonan kohteen sijaan. Nämä viisitoista datamassaa olisivat myös valmiina visualisoitavaksi, jolloin jokaisen liuskan avulla koko Suomen rakennukset olisivat mahdollista havainnollistaa leveyspiireittäin. Havainnollistamisen vuoksi liuskat väritettäisiin väripaletin avulla merkitsemään eri volyymejä. Liuskat koostuisivat nyt kymmenistä, tai sadoista tuhansista yksittäisistä rakennuksista ja rakennusten sijaintitiedoista. Datamassoja käsiteltäisiin lisää niin, että jokaisesta liuskasta arvioitaisiin keskipiste, josta liuskat jaettaisiin kahtia läntiseen ja itäiseen puoliskoon. Tämä kohdentaisi matriisia siten, että nähtävillä olisi rakennusten jakautuminen liuskoittain molemmilla puolella Suomea. Liuskojen länsi- ja itäpuolet väritettäisiin myös massan suuruutta merkitsevällä väriskaalalla, jotta datasta syntyisi toinenkin visualisointi.

Näiden aluejakojen perusteella on mahdollista kartoittaa miten rakennukset jakautuvat ympäri Suomea viidentoista ja kolmenkymmenen fyysisen alueen sisällä. Aluejakoa voi tästä eteenpäin tiivistää eksponentiaalisesti, joka myös kasvattaisi työtä, sekä myös objekteja kartalla. Visualisoinnin muuttujien kasvattamisen sijaan, tarkoituksena on kohdentaa tiiviimpi alue kahteen kaupunkiin. Jotta lopputuloksena olisi mahdollisimman erilaiset tulokset, alueina olisi yksi kaupunki tiheämmästä ympäristöstä ja toinen kaupunki mahdolliselta haja-asutusalueelta. Tiheyttä voidaan arvioida edellisellä kolmenkymmenen ruudun aluejaolla. Näiden kahden kaupungin päälle sijoitettaisiin ruudukko koostuen kahdestakymmenestäviidestä kappaleesta viiden kilometrin levyistä ja korkuista ruutua. Ruudukon sisällä ryhmiteltyjen rakennusten lukumäärät merkittäisiin samanlaisella väriskaalamenettelyllä ja teemoilla, kuten edelliset kaksi visualisointia.

Visualisoinneissa on annettu jokin väripaletin väri edustamaan tiettyä lukumäärää määräasteikolla. Visualisoinnissa tavoitteena on asettaa värit kertomaan massojen volyymeistä siten, että ne olisivat keskenään vertailukelvollisia. Esimerkiksi sateenkaaren värit eivät tässä toimisi, koska eri värit kertoisivat enemmän arvojen yksittäisyydestä. Väriskaalan avulla on mahdollista esittää vaihtelevaisuutta, sekä värin sakeudella on mahdollista hahmottaa järjestystä. Visualisoinneissa käytettävä väripaletti tulee koostumaan kahdesta hajautuvasta väristä, punaisesta ja sinisestä, joiden keskiasteikolla värit ovat haaleat. Tarkoituksena on tuoda esille alueen tiheys esille molemmista ääripäistä, eikä korostaa keskitason arvoja. Kirkas punainen väri esittäisi tiheää aluetta, jonka väri haalistuu arvon laskiessa siniseen, eli hajanaisempaan alueeseen. Keskitasossa näkyisi silti värillään se, että kallistuuko alue keskitason tiheän vai keskitason hajanaisen puoleen.

4.6 Projektin eteneminen

Visualisoinnissa käytettävä rakennusdata oli projektin alussa 3 475 736 riviä, sekä sisälsi samat tiedot kuten esimerkissä liitteenä 1. Liuskoja tehdessä osoitemassassa tärkeimmät tietueet ovat rakennustunnus, pohjoiskoordinaatti ja itäkoordinaatti, joten muut tietueet ovat tarpeettomia. Lisäksi jokaisesta yksittäisestä rakennuksesta riittää ensisijaisen sijaintiosoitteen karttakoordinaatisto, joten moniosoitteisista rakennuksista jää käsiteltävään dataan vain yksi koordinaattitieto. Karsimisen jälkeen käsiteltävään dataan jäi 3 420 942 riviä.

Seuraavaksi koko massa on jaettu viiteentoista alueeseen. Ensimmäisen alueen pohjoisin raja on valittu Keravan kaupungista hieman pohjoiseen, sijaiten pohjoiskoordinaatissa 6697869. Alue laskee sisällensä kaikki Suomen rakennukset datassa, jotka ovat kyseisen pohjoiskoordinaatin eteläpuolella. Toinen aluerajaus on tehty edellisen alueen pohjoiskoordinaatista laskemalla 77 kilometriä pohjoissuuntaan. Käytännössä pohjoiskoordinaattiin 6697869 on lisätty 77000, joka muuttaa pohjoiskoordinaatin arvoon 6774869. Näiden kahden koordinaattiarvon kilometrimittaus on vahvistettu Maanmittauslaitoksen Karttapaikka–sovelluksella. Tähän toiseen alueeseen kuuluu datasta rakennukset, joiden pohjoiskoordinaatin arvo on pienempi kuin 6774869 ja suurempi kuin 6697869. Kolmannen alueen pohjoisraja on toisen alueen pohjoiskoordinaatista 77000 yksikköä pohjoissuuntaan. Tällä tavalla on jatkettu, kunnes koko Suomen pituinen matka on jaoteltuna datassa. Alueen rakennuksen itäinen koordinaatti voi olla mikä tahansa, joka tekee kaikista viidestätoista alueesta liuskamaisia. Aluejako näkyy liitteessä 4.

Jokainen liuska sitten jaetaan liitteen 5 mukaisesti kahteen osaan. Jaottelu on suoritettu lajittelemalla tietokannassa liuskan rakennukset itäisellä koordinaatilla laskevasti, että saadaan liuskan läntisimmät rakennukset, joiden viidentoista ensimmäisen rakennuksen itäisistä koordinaateista on laskettu keskiarvo. Tämän jälkeen sama on tehty lajittelemalla itäisellä koordinaatilla nousevasti, jotta saadaan viisitoista itäisintä koordinaattia. Näistä kahdesta koordinaattiotosten keskiarvoista on laskettu vielä keskiarvo, jotta päästään liuskan arvioituun keskipisteeseen. Päädyin 15 kpl otokseen, koska keskiarvoa laskiessa se on tarpeeksi sopeutuva esimerkiksi Kouvolan läntisimmän rakennuksen tapaukseen, joka tuli esille datan laaduntarkastuksessa. Laskiessa viidentoista pisteen keskiarvot keskenään, tasoittaa yhden virheellisesti kaukaisen koordinaatin muiden neljäntoista Suomen rajalla oleviin, jolloin saadaan hyvä läntinen tai itäinen arvio. Tästä työvaiheesta on osittainen lähdekoodi liitteessä 2.

Koko Suomen rakennusdatan katselmoinnin jälkeen projekti etenee kaupunkiympäristöön kohdistuvaan rakennusdataan. Liitteen 5 visualisoinnin mukaan on mahdollista hahmotella kahden tutkittavien kaupunkiympäristöjen ehdokkaita. Ehdokkaiksi valittiin Kemijärvi ja Kotka. Kemijärven kaupunki siksi, koska suurempiin kaupunkeihin verrattuna siellä mahdollisesti tulisi esille haja-asutusta. Kotkan kaupunki on toisena ehdokkaana, koska alue on hyvin asuttua, sekä kaupungin vesiraja-alueella olisi mahdollisesti näkyvillä selvä erotus datassa. Datassa rakennusten jakamisessa ruutuihin on tekijänä pelkän pohjoiskoordinaatin lisäksi myös itäinen koordinaatti. Ruutujen pinta-ala on noin viisi kilometriä kertaa viisi kilometriä. Ruutuihin on ryhmitelty rakennukset, jotka sijaitsevat kahden pohjoiskoordinaatin välissä ja kahden itäkoordinaatin välissä. Ruutujen koordinaatistot ovat laskettu ja mitattu samalla tavalla, kuten liitteen 4 visualisoinnin kanssa. Ruuturyhmittelystä on osittainen lähdekoodi liitteessä 3.

5 Pohdinta

Tässä opinnäytetyön viimeisessä kappaleessa katselmoidaan toteutuneen projektin lopputulosta. Kappale koostuu lopputuloksen tulkinnasta, arvioinnista, projektin kompastuskohdista, sekä lopulta jatkokehityksen miettimiselle.

5.1 Produktin tulkinta ja analyysit

Liitteen 4 liuskoja visualisoiva kartta, näyttää oleellisen todella selvästi. Oletettava lopputulos näkyy, kun rakennusten määrä hiipuu pohjoiseen päin edetessä. Tasainen lasku kuitenkin katkeaa Suomen leveimmässä kohdassa, etelästä katsottuna viidennellä liuskalla. Neljännen liuskan kohdalla sattuu olemaan suuria kaupunkeja, kuten Jyväskylä, Keuruu, Varkaus ja suurin osa Savonlinnaa. Viidennellä liuskalla on silti enemmän rakennuksia, koska liuskalle sijoittuu esimerkiksi Kuopion, Joensuun ja Seinäjoen ympäröimät kuntaalueet, sekä suuri osa Vaasasta.

Liite 5, joka visualisoi keskeltä jaetut liuskat, on hieman tukkoinen. Liitteessä 4 ja liitteessä 5 olevien visualisointien välillä voi hyvin huomata miten havainnointiin vaikuttaa visualisoinnissa olevien objektien määrä. Mahdollisesti jos liuskoissa olevien rakennusten lukumäärät eivät olisi esitettynä liuskojen alakulmassa, visualisoinnista välittyisi helpommin olennainen tieto. Ulkonäöstä huolimatta visualisoinnista tulee hyvin selville, miten esimerkiksi Keski-Suomen länsirannikko on paljon itäpuolta enemmän asutettua. Myöskin esimerkiksi pohjoisesta laskettuna neljännen liuskan vaihtelevaisuus korostuu huomattavasti. Liitteen 5 visualisoinnin avulla saa aivan erilaisen, sekä paljon tarkemman kuvan verrattuna liitteen 4 visualisointiin. Silti datasta havainnointi on suhteellisen helppoa, eikä toteutukset vienyt suurta aikaa, joten lopputulos oli erittäin positiivinen.

Kemijärven ruudukko, liitteessä 6 ei ole asetettuna keskustan alueen päälle, vaan hieman keskustasta koilliseen. Ruudukossa on silti mukana iso osa keskustaa, mutta suurin osa ruudukosta sijaitsee haja-asutusalueen päällä, joka tulee selvästi esille ruudukon koillisosassa. Ruudukosta on mahdollista havainnoida asutuksen jakautuminen suurempien vesialueiden ympärille, ruudukon länsi- ja lounaisosassa. Myös kaakkoisosassa ruudukkoa on näkyvissä Joutsijärven kylä, jonka osa on sattunut ruudukolle.

Kotkan kaupungin päällä olevassa ruudukossa, liitteessä 7, on esillä rakennusten levinneisyyttä, varsinkin E18–valtatien varrella. Levinneisyys on selvästi esillä myös keskustaa ympäröivän veden äärellä niemissä ja saaristoissa. Edellisessä vaiheessa odotettu vesirajan vaihtuvaisuus näkyy ruudukon kaakkoisosassa todella vahvasti. Kaakkoisosassa

Kuutsalon saaristo näkyy olevan myös suositusti asuttua ja mahdollisesti pitää sisällään enimmäkseen vapaa-ajan asutuksia. Lopuksi jos vertailee liitteen 7 ja liitteen 5 visualisointeja, niin voi huomata, että Kotkan keskusta pitää sisällään enemmän rakennuksia kuin suurin osa Pohjois-Lapista.

5.2 Työn arviointi

Kaiken kaikkiaan visualisointiprojektin lopputulokset näyttävät positiivisilta ja antavat hyvän kuvan siitä miten Suomen rakennukset jakautuvat eri alueilla. Lopputulosten avulla voi tehdä selviä havainnointeja avoimesta rakennusdatasta. Visualisoinnit myös herättävät mielenkiintoa jatkokehityksen kannalta. Eri alueita voi katselmoida samalla tavalla tai tarkemmin, sekä suorittaa työ esimerkiksi paremmalla teknologialla.

Työssä kirjoitettu koodi ulottui jopa kahteentuhanteen riviin tai melkein viiteenkymmeneenyhdeksääntuhanteen kirjaimeen. Pituudesta huolimatta koodi ei missään vaiheessa ollut sekavaa, sekä sisältää hyvän määrän selostavia kommentteja. Lauseet olivat silti toistuvia, joten esimerkiksi muuttujia, sekä silmukoita käyttämällä olisi päässyt paljon vähemmällä.

5.3 Haasteet ja ongelmat

Pelkkää tietokantaa ja kuvankäsittelyohjelmaa käyttäessä eri visualisointielementtien käyttö rajoittuu. Ilman visualisointijärjestelmää työaskeleet ovat hitaampia ja vaativat enemmän manuaalista työtä ja altistavat virheille. Esimerkiksi visualisointien ruudukkoja tehdessä kartta ja ruudukko meni monesti toisistaan ohi, jolloin joko ruudukon alla oli osittainen kartta tai kartan päällä oli osittainen ruudukko. Kartan skaalaaminen ruudukon kanssa otti muutaman yrityskerran ja vei pari lisätuntia aikaa.

Rakennusdatassa oli ryhmiteltynä kymmeniä eri rakennusten pääasiallisia käyttötarkoituksia kahteen eri kategoriaan. Kategoriat ovat nimeltään asuin- tai toimitilarakennukset ja tuotanto- tai muut rakennukset. Rakennusdatan kategoria 1 sisältää Tilastokeskuksen rakennusluokitukset A-H, kategoria 2 sisältää rakennusluokitukset J-N. Käytännössä datassa on ryhmitelty 108 kappaletta rakennusten luokitusta kahteen eri osaan, joka on käytännön kannalta liian karkea jako. Tilastokeskuksen rakennusluokitusten mukaan kategoria 1 siis sisältää muun muassa omakotitalot, asuinkerrostalot, mökit, vankilat, sairaalat ja elokuvateatterit. Tämän avulla ei ole mahdollista näyttää esimerkiksi asuinrakennusten levinneisyyttä koko Suomen osalta. Kategoria 2 sisältää esimerkiksi kouluja, tutkimuslaitoksia, voimalaitoksia, varastoja, navettoja, kasvihuoneita, talousrakennuksia ja saunarakennuk-

sia. Laajemmista rakennusten käyttötarkoituksista saisi dataan mukaan tietoja, jolla sitä olisi mahdollista ryhmitellä. (Tilastokeskus, 1994.)

Yhdeksän värin skaala koitui harmilliseksi arvoasteikkojen kannalta. Esimerkiksi tuplasti tiheämpi, kahdeksantoista kappaleen väriskaala olisi tarkentanut visualisoinnin lopputulosta. Myös arvojen asettaminen oikeaan kohtaan skaalaa oli hankalaa. Kuten liitteissä 4-7 on näkyvillä, kaikissa visualisoinneissa on käytössä oma arvoasteikkonsa. Yhtä staattista arvoasteikkoa ei ollut mahdollista käyttää, koska värit niin sanotusti loppuivat kesken.

5.4 Jatkokehitys

Rakennusdataa on mahdollista jatkokäsitellä siten, että koko Suomen data olisi jaoteltuna pinta-alaltaan viisi kilometriä kertaa viisi kilometriä ruudukkomatriisiin. Ruudukkoa voitaisiin värittää alueiden rakennusten tiheyden mukaisesti, jolloin lopputuloksena olisi korkealaatuisempi ja laajempi visualisointi. Myös nelikulmaisen ruudun sijaan alueen muotona voisi käyttää kahdeksankulmaista aluetta joka mahdollisesti tarkentaisi massaa.

Rakennusdatan ETRS–koordinaatit voisi kääntää maailmanlaajuiseen WGS84–karttaprojektioon, jolloin niitä voisi käsitellä helpommin visualisointijärjestelmien kanssa. Rakennusdatan käsittelemisen tietokannoilla voisi jättää kokonaan välistä pois ja käsitellä dataa visualisointijärjestelmillä. Dataa voisi ryhmitellä järjestelmien paremmilla teknologioilla siten, että rakennusmassojen volyymien jakautuminen näkyisi yksityiskohtaisimmin kartalla. Visualisointijärjestelmien avulla datan käsittely ja visualisointi olisi myös huomattavasti nopeampaa ja vaivattomampaa.

Lähteet

Angier, N. 2013. The changing american family. Luettavissa: http://www.nytimes.com/2013/11/26/health/families.html. Luettu: 9.3.2016.

Anscombe, F. 1973. Graphs in statistical analysis. Luettavissa: http://www.sjsu.edu/faculty/gerstman/StatPrimer/anscombe1973.pdf. Luettu: 25.2.2015.

Behrens, J. 1997. Principles and procedures of exploratory data analysis. Luettavissa: http://cll.stanford.edu/~willb/course/behrens97pm.pdf. Luettu: 5.4.2016.

Bruckner, T., Catalano, R. & Subbaraman, M. 2011. Transient cultural influences on infant mortality: Fire-horse daughters in Japan. Luettavissa: http://onlinelibrary.wiley.com/doi/10.1002/ajhb.21174/pdf. Luettu: 11.4.2016.

Corum, J. 2013. Storytelling with data. Luettavissa: http://style.org/tapestry. Luettu: 18.4.2016.

Davenport, T. & Kim, J. 2013. Keeping up with the quants: Your guide to understanding and using analytics. Harvard business review press. Boston, Massachusetts. Luettu: 5.4.2016.

De Smith, M. 2015. Statistical analysis handbook. Luettavissa: http://www.statsref.com/HTML/index.html?exploratory_data_analysis.html. Luettu: 5.4.2016.

Dykes, B. 2010. Reporting vs. analysis: What's the difference? Luettavissa: https://blogs.adobe.com/digitalmarketing/analytics/reporting-vs-analysis-whats-the-difference. Luettu: 21.3.2016.

Fayyad, U., Grinstein, G. & Wierse, A. 2002. Information visualization in data mining and knowledge discovery. Morgan Kaufmann. Burlington, Massachusetts. Luettu: 21.3.2016.

Few, S. 2009. Introduction to geographical data visualization. Luettavissa: https://www.perceptualedge.com/articles/visual_business_intelligence/geographical_data_visualization.pdf. Luettu: 1.5.2016.

Gapminder 2008. Human development trends 2005. Luettavissa: http://www.gapminder.org/downloads/human-development-trends-2005. Luettu: 14.9.2015.

Giles, D. 2011. Anscombe's Quartet. Luettavissa: web.uvic.ca/~dgiles/blog/anscombe.xls. Luettu: 25.2.2015.

Google 2016. Googlen Public Data–palvelun vastuuvapauslauseke. Luettavissa: https://www.google.com/publicdata/disclaimer. Luettu: 1.5.2016.

Heer, J. & Segel, E. 2010. Narrative visualization: Telling stories with data. Luettavissa: http://vis.stanford.edu/files/2010-Narrative-InfoVis.pdf. Luettu: 25.2.2015.

Holzman, S. 2016. On GIFs and responsive design in dataviz. Luettavissa: http://chartsoncharts.com/opinions/no-more-gifs. Luettu: 11.4.2016.

Human fertility database 2016. Human fertility database–projektin seloste. Luettavissa: http://humanfertility.org/cgi-bin/about.php. Luettu: 11.4.2016.

JHS-suositukset 2016. JHS 197 EUREF-FIN -koordinaattijärjestelmät, niihin liittyvät muunnokset ja karttalehtijako. Luettavissa: http://docs.jhs-suositukset.fi/jhs-suositukset/JHS197/JHS197.html. Luettu: 1.5.2016.

Kosara, R. & Mackinlay, J. 2013. Storytelling: The next step for visualization. Luettavissa: http://kosara.net/papers/2013/Kosara_Computer_2013.pdf. Luettu: 14.9.2015.

Lightstone, E. 2013. Data Visualization Design by Etan Lightstone at FutureStack13. Katseltavissa: http://youtu.be/VG6fFjIFxbo?t=8m48s. Katsottu: 25.2.2015.

Parikh, R. 2014. Anscombe’s quartet, and why summary statistics don’t tell the whole story. Luettavissa: http://data.heapanalytics.com/anscombes-quartet-and-why-summary-statistics-dont-tell-the-whole-story. Luettu: 25.2.2015.

Rosling, H. 2011. Hans Rosling and the magic washing machine. Luettavissa: https://www.ted.com/talks/hans_rosling_and_the_magic_washing_machine/transcript. Luettu: 14.9.2015.

Schacter, M. 2010. The art of the performance dashboard. Luettavissa: http://media.wix.com/ugd/dadb01_67024a8b0ac5243ee0d4e8f668ca1e4a.pdf. Luettu 11.4. 2016.

The world bank 2016. Maailmanpankin avoimen syntyneisyysdatan seloste. Luettavissa: http://data.worldbank.org/indicator/SP.DYN.CBRT.IN. Luettu: 11.4.2016.

Tilastokeskus 1994. Rakennusluokitus 1994. Luettavissa: http://www.stat.fi/meta/luokitukset/rakennus/001-1994/koko_luokitus.html. Luettu 1.5.2016.

Tufte, E. 2001. The visual display of quantitative information. Graphics press. Cheshire, Connecticut. Luettu: 21.3.2016.

Liitteet

Liite 1. Esimerkki avoimesta rakennusten sijaintidatasta

rakennustunnus	sijaintikunta	maakunta	kayttotarkoitus	pkoord	ikoord	osoitenumero	lahiosoite	lahiosoite_sv	katunumero	postinumero
103142355R	091	01	1	6673563	386926	1	Lintulahdenkuja	Fågelviksgränden	4	00530
1032542752	091	01	1	6674923	385846	1	Liukulaakerinkuja	Glidlagergränden	4	00520
1032542752	091	01	1	6674923	385846	2	Konepajanraitti	Maskinverkstadsstråket		00520
1032762863	091	01	1	6675676	385476	1	Veturimiehenkatu	Lokmannagatan	6	00520
1032762863	091	01	1	6675676	385476	2	Rautatieläisenkatu	Järnvägsmannavägen	7	00520
1032762863	091	01	1	6675676	385476	3	Ratapihantie	Bangårdsvägen	13	00520

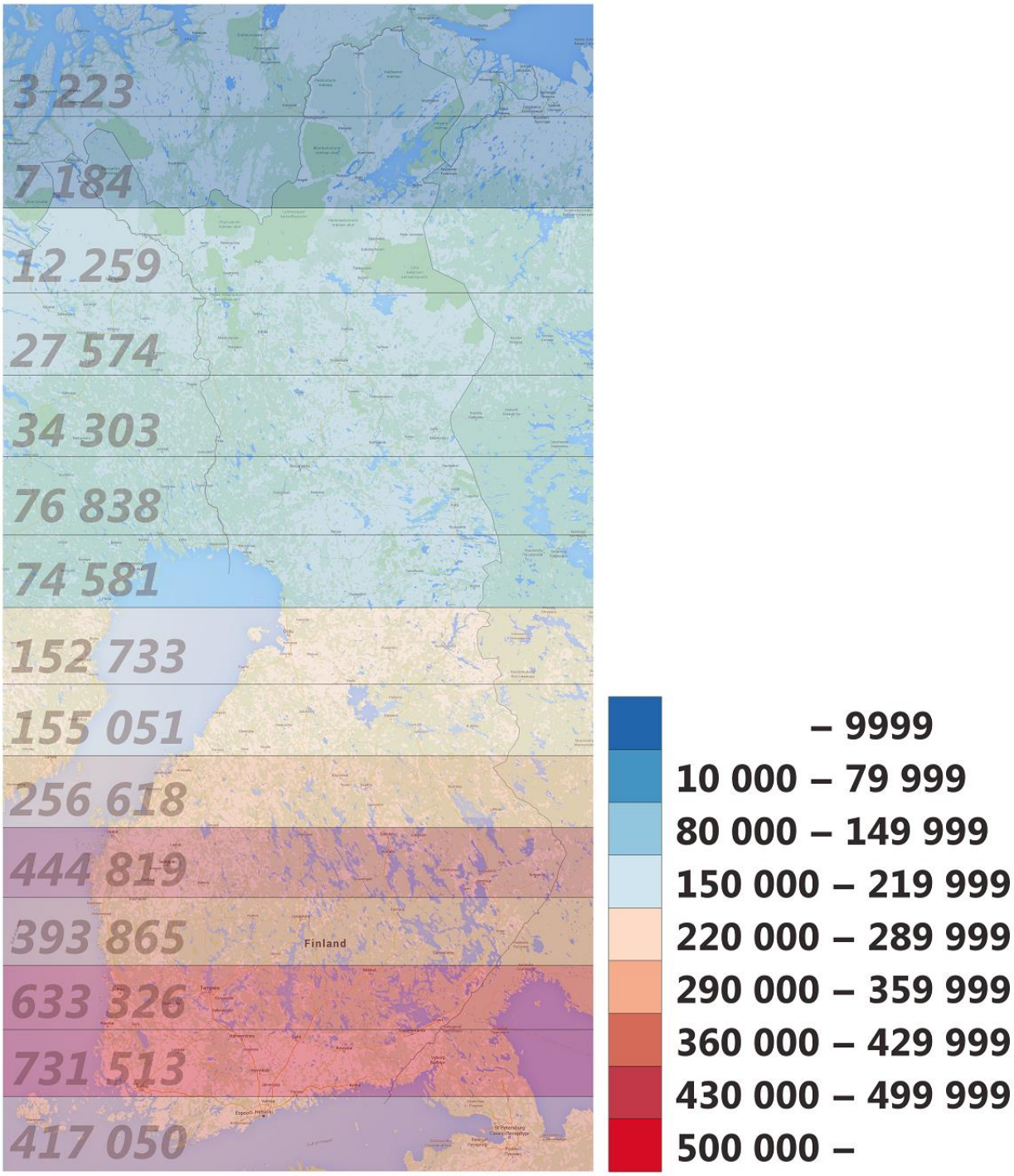
Liite 2. Liuskojen keskipisteiden arvioiminen

```
-- luodaan itäiselle koordinaatin sarakkeelle indeksointi lajittelua varten liuskalle (liuska numero 1)
create index iidx on `77km_liuska01` (ikoord);

-- luodaan taulu laskettaville keskiarvoille
create table `liuskojenkeskiarvot` (
		`pkoord` INT(11) NULL DEFAULT NULL,
		`ikoord` INT(11) NULL DEFAULT NULL,
		`liuska` VARCHAR(13) NOT NULL DEFAULT '' COLLATE 'utf8_general_ci',
		`laji` VARCHAR(3) NOT NULL DEFAULT '' COLLATE 'utf8_general_ci');

-- sijoitetaan 15 kpl läntisimpiä ja itäisimpiä koordinaattipareja tauluun
-- liuska-sarakkeeseen liuskan nimi ja laji-sarakkeeseen arvot läntinen tai itäinen
-- jokaisen liuskan kohdalla dataa syntyy 30 riviä
insert into liuskojenkeskiarvot
select pkoord, ikoord, '77km_liuska01' as liuska, 'lan' as laji
from `77km_liuska01`
order by ikoord desc
limit 15;

insert into liuskojenkeskiarvot
select pkoord, ikoord, '77km_liuska01' as liuska, 'ita' as laji
from `77km_liuska01`
order by ikoord asc
limit 15;

/*
** ... toistetaan sama muiden liuskojen osalta ...
*/


-- luodaan taulu läntisten ja itäisten pisteiden keskiarvojen laskemiseen
CREATE TABLE `liuskat_lansijaita` (
		`pkoord` INT(11) NULL DEFAULT NULL,
		`ikoord` INT(11) NULL DEFAULT NULL,
		`liuska` VARCHAR(13) NOT NULL DEFAULT '' COLLATE 'utf8_general_ci',
		`laji` VARCHAR(3) NOT NULL DEFAULT '' COLLATE 'utf8_general_ci');

-- lasketaan koordinaattien pyöristetyt keskiarvot per 15 koordinaattia
-- data pienenee jokaisen liuskan kohdalla kahteen riviin, itäisimpään ja läntisimpään koordinaattiin
insert into liuskat_lansijaita
select round(avg(pkoord)) as pkoord, round(avg(ikoord)) as ikoord, liuska, laji
from liuskojenkeskiarvot
group by liuska,laji;

-- luodaan taulu keskipisteen arvioinnille
CREATE TABLE `liuskat_keskipiste` (
		`pkoord` INT(11) NULL DEFAULT NULL,
		`ikoord` INT(11) NULL DEFAULT NULL,
		`liuska` VARCHAR(13) NOT NULL DEFAULT '' COLLATE 'utf8_general_ci');

-- lasketaan pyöristetyt keskiarvot läntisimmän ja itäisimmän pisteen kesken liuskoittain
-- periaatteessa itäisellä koordinaatilla ainoastaan on väliä, mutta pidetään pkoordinaatti mukana,
jotta pisteen voi tarkastaa
-- data pienenee jokaisen liuskan kohdalla yhteen riviin, arvioituun keskipisteeseen
create table liuskat_keskipiste
select round(avg(pkoord)) as pkoord, round(avg(ikoord)) as ikoord, liuska
from liuskat_lansijaita
group by liuska;
```

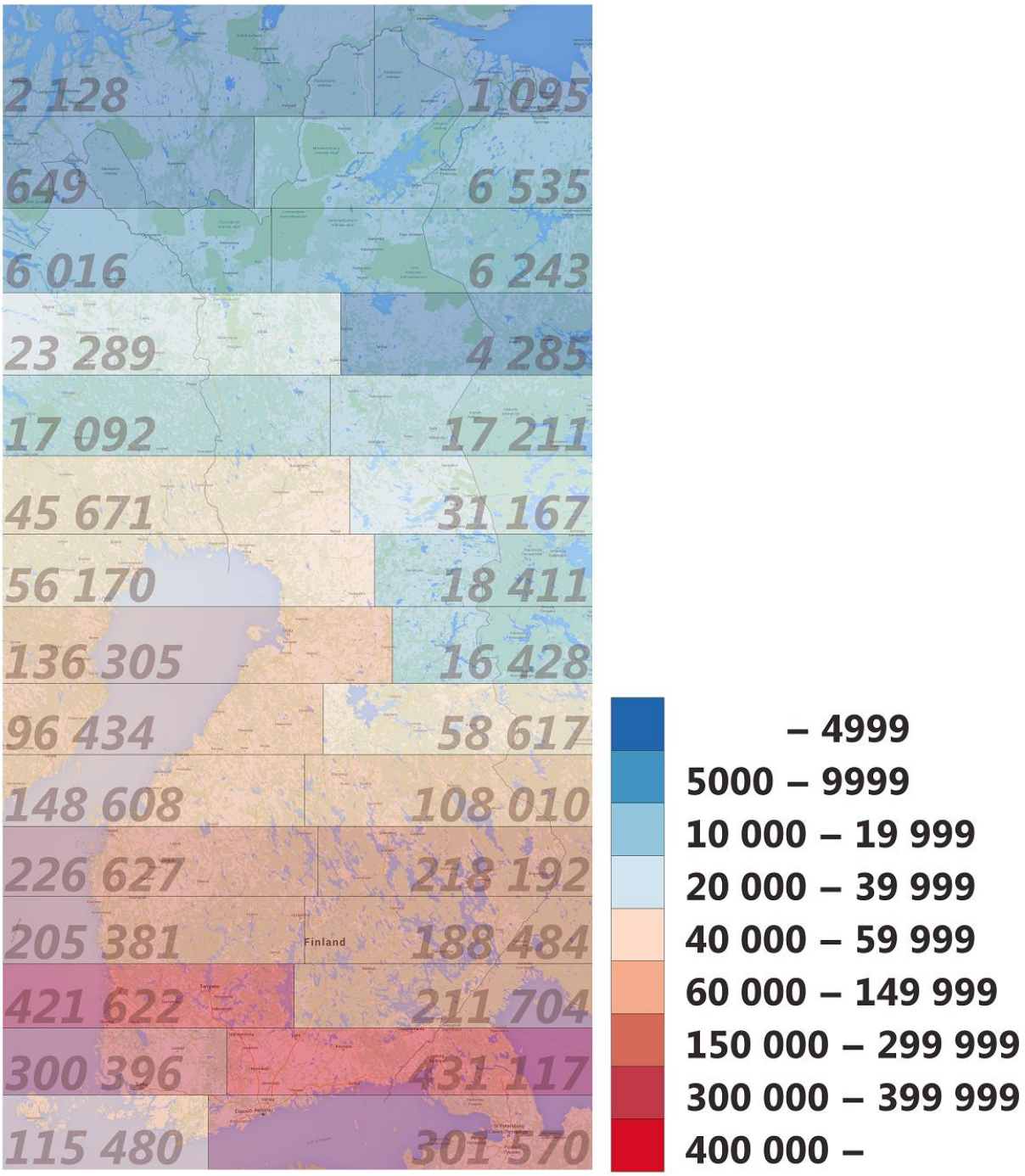
Liite 3. Kemijärven ruutumatriisin rakennusten laskeminen

```
-- Kemijärvi
-- poimitaan edellisten liuskojen avulla käsiteltävä massa
create table kemijärvi_5km_ruutu_01
select * from 77km_liuska11_ita -- liuskan 11 itäpuoli
union all
select * from 77km_liuska10_ita; -- liuskan 10 itäpuoli


create table kemijärvi_5km_ruutu_02 select count(*) as `kpl`, 'A1' as `laji`
from kemijärvi_5km_ruutu_01 where pkoord between '7410000' and '7414999'
                and ikoord between '518523' and '523522';

insert into kemijärvi_5km_ruutu_02 select count(*) as `kpl`, 'A2' as `laji`
from kemijärvi_5km_ruutu_01 where pkoord between '7405000' and '7409999'
                and ikoord between '518523' and '523522';

insert into kemijärvi_5km_ruutu_02 select count(*) as `kpl`, 'A3' as `laji`
from kemijärvi_5km_ruutu_01 where pkoord between '7400000' and '7404999'
                and ikoord between '518523' and '523522';

insert into kemijärvi_5km_ruutu_02 select count(*) as `kpl`, 'A4' as `laji`
from kemijärvi_5km_ruutu_01 where pkoord between '7395000' and '7399999'
                and ikoord between '518523' and '523522';

insert into kemijärvi_5km_ruutu_02 select count(*) as `kpl`, 'A5' as `laji`
from kemijärvi_5km_ruutu_01 where pkoord between '7390000' and '7394999'
                and ikoord between '518523' and '523522';


insert into kemijärvi_5km_ruutu_02 select count(*) as `kpl`, 'B1' as `laji`
from kemijärvi_5km_ruutu_01 where pkoord between '7410000' and '7414999'
                and ikoord between '523523' and '528522';

insert into kemijärvi_5km_ruutu_02 select count(*) as `kpl`, 'B2' as `laji`
from kemijärvi_5km_ruutu_01 where pkoord between '7405000' and '7409999'
                and ikoord between '523523' and '528522';

/*
** ... toistetaan sama muiden ruutujen osalta ...
*/
```

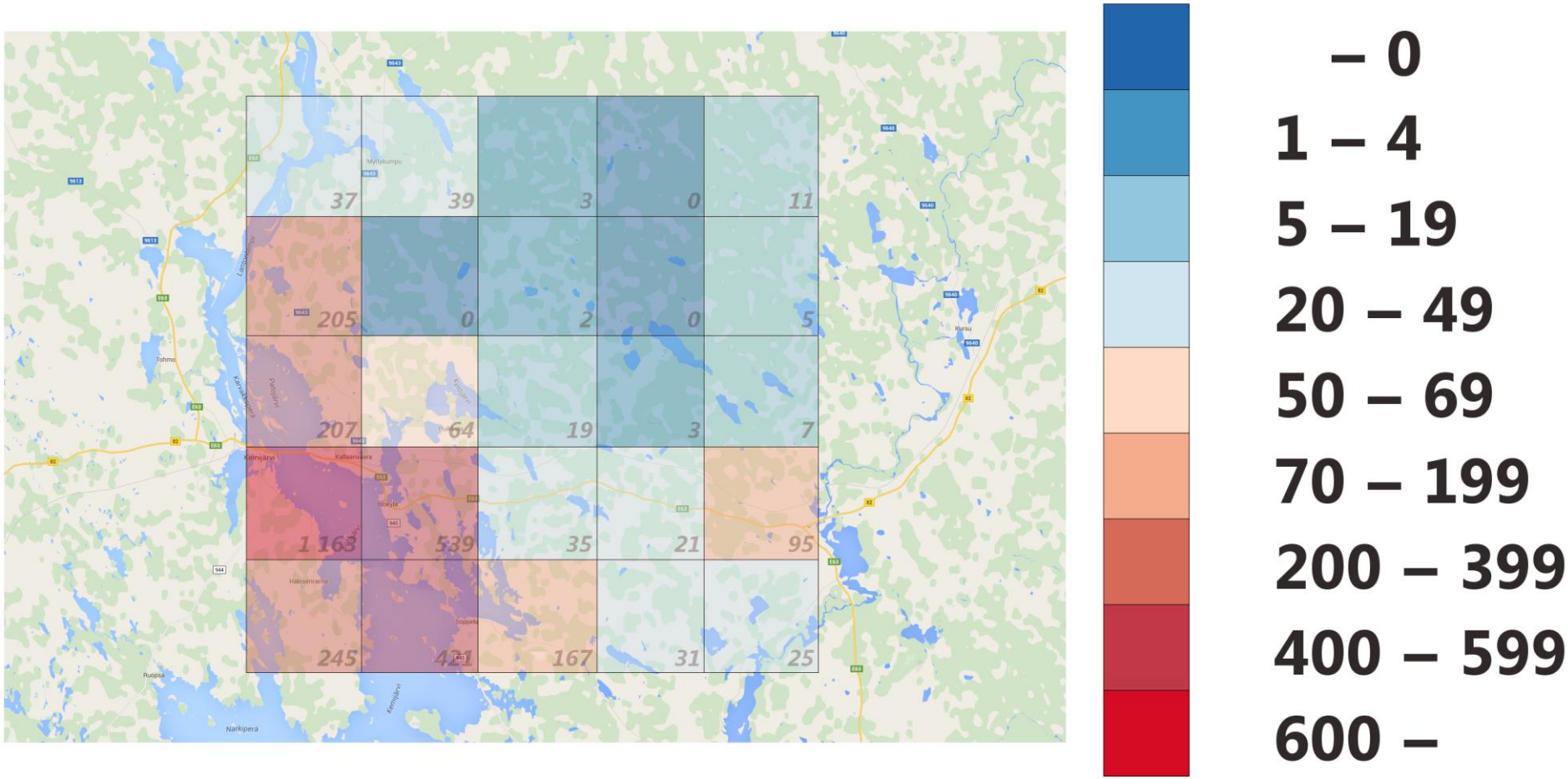
Liite 4. Visualisointi koko Suomen rakennuksista alueittain (liuskat)



Liite 5. Visualisointi koko Suomen rakennuksista alueittain (puolitetut liuskat)



Liite 6. Visualisointi Kemijärven alueelta ruuduittain



Liite 7. Visualisointi Kotkan alueelta ruuduittain