Suuret tietomassat ja koneoppiminen makrotaloustieteellisessä tutkimuksessa

Teemu Pekkarinen

Taloustiede on alkanut käyttää kasvavassa määrin suuria tietomassoja eli niin sanottua big dataa tutkimuksessaan. Erilaisten tilastolähteiden ja tilastomuotojen määrä on kasvanut valtavasti, mikä on vaatinut lisää työkaluja taloustieteelliseen tutkimukseen myös tilasto- ja tietojenkäsittelytieteiden puolelta. Tämä katsaus tarkastelee suurten tietomassojen käyttöä makrotaloustieteen tutkimuksessa ja ennustamisessa. Artikkelissa esitellään myös muutamia koneoppimisen menetelmiä suurten tietomassojen käsittelyyn ja analysointiin. Katsauksen keskeisin tavoite on tarkastella laaja-alaisesti, mitä uutta alati kasvavat tietomassat ja koneoppiminen ovat tuoneet makrotaloustieteeseen.

Tietokoneet ja internetsivustot keräävät tänä päivänä valtavia määriä informaatiota ihmisten toimista ja käyttäytymisestä. Lisäksi tietokoneet ovat useimpien taloudellisten transaktioiden välissä.[1] Tämä tarkoittaa sitä, että ihmisten ja yritysten ostoista, myynneistä, paikkatiedoista, kirjoituksista, klikkauksista, internethauista ja monista muista toimista jää jälki johonkin tiedostoon. Tämä informaatio varastoidaan tietokantaan. Sieltä sitä voidaan käyttää esimerkiksi markkinoinnin kohdentamiseen, tapahtumien todentamiseen tai tiedon analysointiin. Tämä uudenlainen tietomassojen järjestelmällinen varastoiminen onkin muodostanut niillä kerätyille suurille tietomassoille täysin uuden englanninkielisen kutsumanimen: *big datan*.

Big data ei kuitenkaan tarkoita pelkästään aineistoa, jossa on suuri määrä havaintoarvoja, vaan usein siihen luokitellaan myös aineistot, joita kerätään korkeilla frekvensseillä ja siten niitä voidaan analysoida tavallista nopeammin, jopa reaaliaikaisesti. Nämä aineistot voivat olla myös ei-numeerisia, esimerkiksi teksti-, kuva- tai videomuotoisia, tai niitä voidaan kerätä uusista lähteistä, kuten sosiaalisesta mediasta, internethauista tai biometrisistä sensoreista.

[1] *Tietokonevälitteisistä transaktioista kts. Varian (2010).*

Teemu Pekkarinen (teemu.pekkarinen@helsinki.fi) on taloustieteen tohtorikoulutettava Helsingin yliopistossa. Kiitän kahta anonyymia lausunnonantajaa sekä Juha Kilposta ja Antti Suvantoa hyödyllisistä ja rakentavista kommenteista.

Näitä uusia suuria aineistoja on käytetty makrotaloustieteellisessä tutkimuksessa esimerkiksi työttömyyden ennustamiseen internethakujen perusteella sekä inflaation ja internetin onlinehintojen vertailuun. Mikrotaloustieteessä puolestaan erilaisten internetaineistojen ja kuluttajan käyttäytymisen välisiä yhteyksiä on alettu tutkia kasvavassa määrin. Taloustieteen tarpeisiin uusia aineistoja on paljon ja niitä on käytetty innovatiivisesti – esimerkiksi tulotasoa on arvioitu satelliittikuvista (Henderson ym., 2012 sekä Donaldson ja Storeygard, 2016).

Kun aineistot sisältävät valtavan määrän informaatiota, joka ei välttämättä ole numeerisessa muodossa, perinteiset ekonometrian välineet ovat kaivanneet tuekseen myös tietojenkäsittelytieteen välineistöä. Aineistojen kasvaessa mallin, muuttujien ja parametrien valinta sekä ylisovittamisen ongelmat ovat entistä keskeisempiä haasteita tutkimuksessa, ja näihin on haettu apua myös koneoppimisen näkökulmasta.

Tässä artikkelissa tarkastellaan suurten tietomassojen ja koneoppimisen käyttöä makrotaloustieteellisessä tutkimuksessa ja ennustamisessa. Suuren suosionsa vuoksi uusia tutkimuksia sekä tutkimusprojekteja ilmaantuu jatkuvasti lisää. Tästä syystä moni mainitsemisen arvoinen julkaisu jää todennäköisesti huomiotta. Myös mikrotaloustieteessä on useita kiinnostavia *big data* -sovelluksia ja monet koneoppimisen välineet sopivat mikroaineistoille jopa paremmin kuin makromuuttujille. Näiden laaja-alainen esittely ansaitsisi kuitenkin oman tutkimuksensa yhtä lailla koneoppimisen menetelmien yksityiskohtaisemman tarkastelun kanssa. Tämän katsauksen keskeisin tavoite onkin antaa mahdollisesti ensimmäinen kipinä aiheeseen ja esitellä laveasti suurien tietomassojen ja muutamien koneoppimisen työkalujen käyttöä perinteisteisissä makrotaloustieteen sovelluskohteissa.

Artikkeli pyrkii vastaamaan kahteen kysymykseen: Millaista makrotaloustieteellistä tutkimusta suurten tietomassojen avulla on tehty? Entä mitä koneoppimisen työkalut tarjoavat makrotaloustieteeseen ja ennustamiseen? Artikkeli etenee näiden kahden kysymyksen mukaisessa järjestyksessä. Viimeisessä jaksossa pohditaan, kuinka alati kasvavat datamassat vaikuttavat yleisesti taloustieteeseen.

1. Suurten datamassojen käyttö makrotaloustieteessä

1980-luvun puoleen väliin asti suuri osa taloustieteellisestä tutkimuksesta oli teoreettista. Tietokoneiden kehittyessä empiiristen tutkimusten määrä alkoi kasvaa nopeasti. Tänä päivänä yli 70 prosenttia julkaistuista tutkimuksista perustuu havaittuun aineistoon, josta valtaosa on tutkijoiden itsensä keräämää. Lisäksi myös kokeellisten tutkimusten määrä on kasvanut. Parissa vuosikymmenessä on siis tapahtunut suuri harppaus empiirisen tutkimuksen suuntaan (Hamermesh 2013).

Suuret tietomassat ovat rantautuneet makrotaloustieteen tutkimukseen vasta viime aikoina. Valtaosa suuria tietomassoja käsittelevistä makrotaloustieteeseen ja ennustamiseen liittyvistä tutkimuksista käyttää joko sosiaalisen median dataa tai Googlen hakudataa.[2] Yksi syy, miksi juuri Googlen aineistot ovat

[2] *Google-hakuaineistolla katsauksessa tarkoitetaan Googlen keräämää aineistoa, joka sisältää tiedon jokaisesta Googlella tehdystä hausta. Kaikille avointa tilastoa on julkaistu viikkotasoisesti vuodesta 2004 alkaen Google Trends -palvelun kautta.*

suosittuja, on niiden helppo ja maksuton saatavuus. Yleisesti suosittuja tutkimus- ja sovelluskohteita suurille tietomassoille ovat olleet työttömyys, yksityinen kulutus, inflaatio sekä rahoitus- ja asuntomarkkinat. Seuraavaksi esitellään muutamia tutkimuksia näistä aiheista, samaisessa järjestyksessä.

Ettredgen ym. (2005) artikkeli oli yksi ensimmäisistä tutkimuksista, jossa hakukoneilla tehtyjen hakujen määrää pyrittiin käyttämään ennustamiseen. Heidän saamien tulosten mukaan Yhdysvalloissa työnhakuun liittyvillä internethauilla ja virallisilla työttömyysluvuilla on ollut positiivinen ja tilastollisesti merkittävä yhteys. Vuonna 2005 Ettredgen ym. käyttämiä hakutuloksia oli kuitenkin julkaistu vain lyhyeltä ajalta, joten heidän tuloksensa antoivat vain pieniä viitteitä siitä, miten internethaut voisivat lisätä ennustevoimaa makromuuttujille. Antenucci ym. (2014) tarkastelivat työmarkkinoita Twitter-aineistolla muodostamalla signaaleja "menetin työni" -tyyppisistä kirjoituksista. Näillä signaaleilla, eli tarkoitukseen sopivien tviittien määrällä, he muodostivat indikaattoreita, jotka ennakoivat työttömyyslukuja. Gee ym. (2017a; 2017b) puolestaan käyttivät Facebook-aineistoa tutkiessaan, kuinka heikot ja vahvat siteet sosiaalisissa verkostoissa toimivat työn saannissa. Toisin kuin aikaisemmin oli tutkittu Geen ym. (2017a) tulosten mukaan vahvemmat sosiaaliset siteet ovat työn löytämisessä merkittävämpiä kuin heikot siteet. Lisäksi Geen ym. (2017b) käyttämä lähes 17 miljoonan sosiaalisen siteen sisältämä aineisto 55 eri maasta osoittaa, että on todennäköisempää päätyä sellaiseen työpaikkaan, missä joku henkilön ystävistä on entuudestaan töissä.

Googlen hakuaineistoa puolestaan on käytetty paljon työttömyyden seurantaan sekä lyhyen aikavälin ennustamiseen (*nowcasting*) että pidemmän aikavälin ennustamiseen (*forecasting*). Google-hakuaineistolla tehdyistä tutkimuksista ehkä tunnetuimpia ovat Yhdysvaltojen työttömyyttä tutkineet Choi ja Varian (2012) sekä Saksan työttömyyttä tarkastelleet Askitas ja Zimmermann (2009).[3] Tuhkuri (2014) oli ensimmäinen, joka testasi Google-aineiston toimivuutta Suomen työttömyysluvuille. Choin ja Varianin (2012) menetelmiä seuraten Tuhkuri (2014) muodosti Google-indeksin työttömyydelle käyttäen työttömyyteen liittyviä hakusanoja kuten "työttömyysetuudet". Hänen tuloksensa osoittavat, että kyseinen indeksi on merkitsevästi korreloitunut työttömyysasteen kanssa ja ennakoi verrattain hyvin työttömyyslukuja.

Myös yksityistä kulutusta ja kuluttajien käyttäytymistä on tutkittu Googlen aineistolla jonkin verran. Kholodilin ym. (2010) tekivät lyhyen aikavälin ennusteita Yhdysvaltojen yksityisen kulutuksen vuosimuutokselle. Heidän Google-hauilla tekemänsä yksityisen kulutuksen mallin ennustekyky oli tilastollisesti merkitsevästi parempi kuin yksinkertaisella autoregressiivisellä mallilla. Kholodilin ym. (2010) vertailivat Google-hakuihin perustuvia ennusteita myös malleihin, missä oli mukana kulut-

[3] *Lisäksi Google hakuaineistolla Yhdysvaltojen työttömyyttä ovat tutkineet muuan muassa Kuhn ja Skuterud (2004), Stevenson (2008), Choi ja Varian (2009), D'Amuri ja Marcucci (2012), Kuhn ja Mansour (2014), Tuhkuri (2015) ja Baker ja Fradkin (2017). D'Amuri (2009) on tutkinut työttömyyttä Italiassa, Anvik ja Gjelstad (2010) Norjassa, McLaren ja Shanbhodue (2011) Isossa-Britanniassa, Chadwick ja Sengul (2012) Turkissa, Fondeur ja Karamé (2013) Ranskassa ja Vicente ym. (2015) Espanjassa. Lisäksi Tuhkuri (2016) ennustaa työttömyysastetta kaikissa EU-28 maissa, ja Pavlicek ja Kristoufek (2015) tutkivat Tšekin, Unkarin, Puolan ja Slovakian työttömyyttä.*

tajatutkimuksia ja talouden indikaattoreita ja päätyivät lopputulokseen, että Google-haut auttavat parantamaan yksityisen kulutuksen ennusteita. Vosen ja Schmidt (2011; 2012) ennustivat yksityistä kulutusta Yhdysvalloissa ja Saksassa käyttämällä hyväksi Google-aineistoa. Heidän tuloksensa osoittivat, että suurin osa Google-indikaattoreihin perustuvista aineiston ulkopuolisista (*out-of-sample*) ja sisäpuolisista (*in-sample*) ennusteista toimivat paremmin kuin kyselytutkimuksiin perustuvat indikaattorit.

Massachusetts Institute of Technologyssa on käynnissä hanke, jonka nimi on *The Billion Prices Project.* Siinä kerätään sadoilta verkossa olevilta vähittäismyyjiltä päivittäin hintatietoja ympäri maailmaa. Aineisto tarjoaa siten lyhyen frekvenssin hintatietoja, joita voidaan hyödyntää esimerkiksi inflaation ja inflaatio-odotusten tutkimuksissa, minkä avulla tätä tietoa voidaan puolestaan käyttää muissa makrotaloustieteen kysymyksissä. Viimeisimpiä tutkimuksia hankkeelta ovat Cavallo ja Rigobon (2016) sekä Cavallo (2017), joissa tutkitaan onlinehintojen yhteyttä *offline*-hintoihin ja hintaindekseihin. Vaikka kyseinen projekti on vasta alussa, heidän tutkimuksensa on havainnut selkeitä yhtäläisyyksiä *online*- ja *offline*-hintojen muodostuksessa ja dynamiikassa. Myös muita inflaatioon liittyviä suurien tietomassojen tutkimuksia on tehty. Esimerkiksi Powell ym. (2017) käyttivät internetistä kerättyjä hintoja ennustaakseen hintaindeksejä. Heidän mukaansa erityisesti elintarvikkeiden (voi, viski, omenat, banaanit, jogurtti, ym.) hintatasoja pystytään ennakoimaan hyvin käyttämällä nopeasti päivittyviä *online*-hintoja. Guzman (2011) puolestaan vertaili Google-datalla muodostettuja inflaatio-odotuksia 36 eri inflaatio-odotusindikaattoriin. Hänen tulostensa mukaan *big datan* avulla muodostetuilla inflaatioennusteilla on pienin ennustevirhe. Koop ja Onorante (2016) testasivat, tuoko Google-data ennustevoimaa yhdeksälle Yhdysvaltojen makromuuttujalle, joissa mukana oli esimerkiksi kuluttajahinta- ja palkkainflaatio sekä raaka-aineiden hintaindeksi ja öljyn hinta. Heidän tulostensa mukaan Google-datan sisällyttäminen ennustemalliin paransi lyhyen aikavälin ennusteita, mutta se toimi parhaiten, kun Google-muuttujat oli lisätty malliin tietyillä painotuksilla ja todennäköisyyksillä.[4] Herääkin kysymys, voivatko nämä uudet tilastoinnit (esimerkiksi *The Billion Prices Projectin* onlinehinnat) korvata tai olla apuna perinteisemmälle taloustilastoinnille?

Osakekurssit ja niiden kaupankäyntivolyymit ovat olleet mielenkiintoinen kohde *big datan* soveltamiselle. Esimerkiksi Bollen ym. (2011) käyttivät Twitter-aineistoa ihmisten mielialojen seurantaan ja pyrkivät sitä kautta ennustamaan Dow Jonesin osakeindeksiä (*Dow Jones Industrial Index*). Heidän tuloksensa osoittavat Twitter-mielialaindeksin ja Dow Jonesin päivän päätöskurssien välillä olevan tilastollisesti merkitsevä korrelaatio. Preis ym. (2013) käyttivät Google-dataa tämän indeksin lyhyen ja pidemmän aikavälin ennustamiseen. Bordino ym. (2012) sovelsivat erilaisten hakukoneilla suoritettujen hakujen määriä NASDAQ-100:n kaupankäyntivolyymin ennustamiseen. Myös näissä kahdessa tutkimuksissa hakukoneaineistoissa havaittiin olevan tilastollisesti merkitsevä riippuvuussuhde selitettävään tekijään. Moat ym. (2013) ja Curme ym. (2014)

[4] *Koop ja Onorante (2016) käyttivät ennusteissaan dynaamista mallien keskiarvoistamista (Dynamic Model Averaging) ja dynaamisten mallien valintaa (Dynamic Model Selection) yhdessä muuttuvaparametristen regressioiden kanssa.*

puolestaan käyttivät Google-hakuaineistoa sekä finanssialaan liittyvien Wikipedia-sivustojen vierailukertoja osakekurssien ennustamiseen. Heidän mukaansa myös Wikipedia-aineisto saattaa parantaa sijoituskäyttäytymisen ennusteita.

Myös asuntomarkkinoiden hintojen kehitystä on pyritty mallintamaan ja ennustamaan Google-aineistolla. Yhdysvaltojen asuntojen hintoja ovat tutkineet Kulkarni ym. (2009) sekä Wu ja Brynjolfsson (2015). McLaren ja Shanbhodue (2011) ovat tutkineet Britannian ja Widgrén (2016) Suomen asuntomarkkinoita. Näiden tutkimusten perusteella hakukoneaineistot näyttäisivät tarjoavan hyvin informaatiota markkinoiden reaktioista erilaisiin sokkeihin ja uutisiin.

On selvää, että uudet suuret tietomassat ovat suurelta osin mikrotaloustieteen aineistoja, jotka tarjoavat runsaasti erilaisia tutkimuskohteita niin kuluttajien käyttäytymisestä ja tuloista kuin myös yritysten kassavirroista ja tuloksista. Siksipä luultavasti valtaosa *big dataa* käyttäneistä tutkimuksista sijoittunee makrotaloustieteen ulkopuolelle. Näistä esimerkkinä voidaan mainita Uber-kuljettajien työmarkkinoita ja palkkoja tutkineet Cohen ym. (2016) sekä Hall ja Krueger (2016). Goel ym. (2010) puolestaan käyttivät eri internethakujen määriä ennustamaan ensi-iltaan tulevien elokuvien lippukassoja, uusien videopelien ensimmäisen kuukauden myyntejä sekä musiikkikappaleiden sijoitusta Billboard Hot 100 -listalla. Päivittäistavarakauppojen skanneridataan perustuvia tutkimuksia ovat julkaisseet muun muassa Kortelainen ym. (2016), Anderson ym. (2017) sekä Hong ja Li (2017). Kroft ja Pope (2014) tutkivat Craiglist-sivuston suosion kasvun vaikutusta asunto- ja työmarkkinoiden kohtaantoon.[5] Laouénan ja Rathelot (2017) tutkivat Airbnb-datalla myyjien etnisyyden vaikutusta majoituksen hintaan. Lendle ym. (2016) hyödyntävät eBay-aineistoa tutkiessaan maantieteellisen etäisyyden vaikutusta tavaran vaihdantaan.[6]

Kuten huomata saattaa, erilaisia aineistoja ja tutkimuskohteita on valtavasti. Siksipä kaikki nämä muutamat esimerkkinä annetut tutkimukset ovat vasta alkusoittoa tulevaan. Lisää kirjallisuudesta löytyy esimerkiksi katsausartikkeleista Einav ja Levin (2014) sekä Askitas ja Zimmermann (2015).

2. Koneoppimisen työkaluja (makro)taloustieteeseen

Aineistojen kasvaessa ja monipuolistuessa myös datan käsittelyn ja analysoinnin työkalut kehittyvät ja jotkin aikaisemmin kehitetyt menetelmät tulevat entistä suurempaan rooliin. Koneoppiminen on yleiskäyttöinen menetelmä, joka on tietojenkäsittelytieteen ja tekoälyn yksi osa-alue. Sen keskeisin tarkoitus on saada ohjelmisto ja sen algoritmit oppimaan omasta tekemisestään. Koneoppimisen menetelmät tarjoavat apua mallin, muuttujien ja parametrien valintaan sekä ylisovittamisen (*overfitting*) ongelmiin, jotka ovat keskeisiä haasteita empii-

[5] *Craiglist-verkkopalvelu tarjoaa ilmaisen internetsivuston pienille ilmoituksille ja mainoksille kuten esimerkiksi työpaikka- ja asuntoilmoituksille.*

[6] *Myös Ginsbergin ym. (2009) tutkimus influenssaepidemian ennustamisesta on paljon viitattu tutkimus taloustieteen ulkopuolelta. Lisäksi Stephens-Davidowitzin (2014) tutkimus etnisten taustojen vaikutuksista presidentti Barack Obaman kannatukseen tarjoaa mielenkiintoisen big datan sovelluskohteen.*

rikolle. Koneoppimisen näkökulmasta ylisovittamisella tarkoitetaan tilannetta, jossa ennustemalli saadaan toimimaan otoksen sisäisessä ennustamisessa moitteetta, eli mallilla pystytään ennustamaan hyvin jo toteutuneita havaintoja, mutta sama malli silti epäonnistuu otoksen ulkopuolisessa ennustamisessa.

Koneoppimisen käyttö mallintamisessa perustuu aineiston jakamiseen testi- ja opetusaineistoihin. Nimiensä mukaisesti opetusaineistolla muodostetaan malli ja tätä testataan testiaineistoon. Näin voidaan tehokkaasti testata esimerkiksi millä mallilla ja millä selittävillä tekijöillä pystytään tuottamaan pienin ennustevirhe. Yhtäältä, jos tutkijan mielenkiinnon kohteena on löytää sopivimmat parametrit malliinsa, ne voidaan etsiä optimoimalla koneoppimisen menetelmillä. Halutaanko ongelmaa lähestyä ohjatusti (*supervised learning*) vai annetaanko koneen itsensä päätyä johonkin lopputulokseen (*unsupervised learning*), on tutkijan itsensä päätettävissä. Erilaisten algoritmien ja sääntöjen avulla ohjelmoija voi laittaa koneen testaamaan erilaisia malleja ja siten päätymään johonkin lopputulokseen. Ohjelmoija voi myös antaa koneen itsensä kokeilla erilaisia sääntöjä, joiden perusteella ohjelmisto opettaa itse itseään päätymään parhaimpaan mahdolliseen tulokseen. Tämä jälkimmäinen on sanan varsinaisessa merkityksessä koneoppimista. Jos tätä tyyliä käytetään jollekin aineistolle esimerkiksi erilaisten riippuvuussuhteiden tai rakenteiden etsintään, puhutaan siitä silloin usein tiedonlouhimisena (*data mining*).

Koneoppimisen muodot voidaan jakaa ohjattuun oppimiseen, minkä alaluokkia ovat regressio ja luokittelu, ja ohjaamattomaan oppimiseen, johon esimerkiksi klusterointi kuuluu. Tässä katsauksessa keskitytään lähinnä ohjattuun oppimiseen, koska tavallisesti taloustieteilijän mielenkiinnon kohteena on testata jotakin ennalta harkittua mallia. Lisäksi makrotaloustieteessä ja ennustamisessa regression lisäksi eräät ohjatun oppimisen menetelmät ovat hiljalleen yleistymässä. On kuitenkin selvää, että liian pienellä aineistolla koneoppiminenkaan ei tarjoa tutkijalle lisäarvoa. Tällöin parametriestimaattien tarkkuus kärsii ja on vähemmän mahdollisuuksia jakaa aineisto erilaisiin testi- ja opetusaineistoihin. Tästä syystä *big data* ja koneoppiminen ovat toinen toisiaan tukevia trendejä, jotka ovat yleistyneet monilla tieteenaloilla. Kun aineistoa on paljon, voidaan myös opetusaineisto jakaa erilaisiin osiin ja sitä kautta testata erilaisten mallien toimivuutta testiaineistoon.

Makrotaloustieteessä niin sanottua ristiinvalidointia on käytetty ennustamisen apuna (esim. Utans ym. 1995, Wohlrabe ja Buchen, 2014 sekä Cheng ja Hansen, 2015). Ristiinvalidointi on tilastotieteen menetelmä, jota käytetään mallintai parametrien valintatilanteessa ennustevirheen arviointiin. Tämä menetelmä on helposti yhteensopiva monen koneoppimisen algoritmin kanssa ja siksi sitä näkee käytettävän paljon. Ristiinvalidointi perustuu opetus- ja testiaineistoiden jakoon, joka voidaan suorittaa monin eri tavoin. Esimerkiksi data voidaan jakaa *K:hon* yhtä suureen osaan ja toistaa ennustevirheen laskenta *K:sta* kertaa siten, että jokainen näistä *K* osajoukosta toimii vuorollaan testiaineistona (*K-fold cross-validation*). Yhtä lailla aineistosta voidaan jättää vuorotellen *i*:nnet arvot pois ja ennustaa näitä lopulla opetusaineistolla (*leave-one-out cross-validation*).

On kuitenkin muistettava, että makroaineistot ovat usein aikasarjoja ja tällöin ristiinvalidointi ei ole täysin suoraviivaista. Yksi helppo opetus- ja testiaineistoihin jako aikasarjamalleille on kasvava aikaikkuna. Toisin

sanoen, valitaan aluksi opetusaineistoksi *T* ensimmäistä arvoa ja testataan mallia *T+1* arvoihin. Tämän jälkeen otetaan aikasarjat ajanhetkeen *T+1* saakka ja ennustetaan *T+2* arvoja. Tällöin opetusaineiston koko kasvaa aina yhdellä, kunnes koko aikasarja-aineisto on käyty läpi.

Toisin sanoen ristiinvalidoinnilla voidaan laskea erilaisilla malleilla ja muuttujilla muodostettuja ennustevirheitä ja siten vertailla näitä keskenään. Tämä yleiskäyttöinen ristiinvalidointi onkin siten jo itsessään eräs ratkaisu mallin, muuttujien ja parametrien valintaan ja ylisovittamisen ongelmiin.

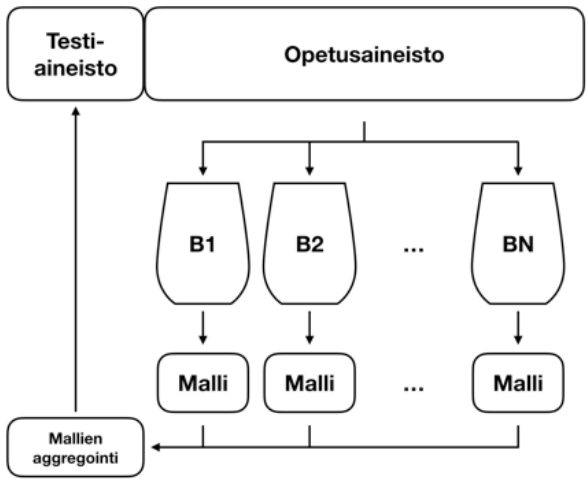
Toinen hieman eri näkökulmasta lähestyvä koneoppimisen metodi on tilastollinen luokittelu. Sen tarkoitus on jakaa aineisto osajoukkoihin käyttäen algoritmeja tai luokittelusääntöjä. Luokittelijana voi toimia jokin ennalta määrätty päätösfunktio (esimerkiksi logit tai probit), päätöspuu tai vaikkapa neuroverkko.[7] Yksinkertaisena esimerkkinä voidaan ottaa diskreetti valintateoria, missä kuluttajan valinta perustuu hyötyfunktioon. Ajatellaan, että kuluttaja on ostoksilla ruokakaupassa. Hän vertailee erilaisia hyödykkeitä keskenään ja valitsee ostoskoriinsa ne hyödykkeet, joista kokee saavansa eniten hyötyä. Tällöin kaupan tarjoama hyödykeavaruus tulee jaetuksi valittuihin ja ei-valittuihin hyödykkeisiin käyttäen hyötyfunktiota luokittelun algoritmina. Koneoppimisen työkalut ovat toisin sanoen jokseenkin helposti yhdistettävissä taloustieteelliseen ajattelutapaan. Kuten arvata saattaa, myös luokittelualgoritmit voidaan opettaa mahdollisimman hyväksi esimerkiksi ennustetarkoitukseen tai hyödyn maksimointia kuvaamaan.

Ehkäpä tunnetuin luokittelumenetelmä on yllä mainittu päätöspuun oppiminen (*decision tree learning*). Päätöspuun rakentaminen on eräs lähestymistapa ennustamiseen, missä voidaan käyttää mahdollisesti suuriakin tietomassoja. Tämä menetelmä voi perustua luokittelupuuhun tai regressiopuuhun taikka molempiin eli niin sanottuun CART-analyysiin (*Classification And Regression Tree*, Breiman ym. 1984). Päätöspuussa luokittelu tehdään peräkkäisten testien avulla. Puumalli tulee sanana siitä, että lajittelukriteerit ja -järjestys esitetään puumuotoisena, missä eri oksan haarat ovat luokittelun sääntöjä. Toisin sanoen puumalli pyrkii jakamaan aineiston selitettävän muuttujan mukaan sitä parhaiten kuvaaviin osajoukkoihin aloittaen esimerkiksi kahteen osaan jakamisesta, jonka jälkeen näiden kahden osan kahteen jakamisesta ja niin edelleen. Pienimmät osajoukot siten kertovat, mitkä selittävät tekijät ja niiden ominaisuudet ennustavat parhaiten selitettävää tekijää.

Bagging (*bootstrap aggregation*) ja *boosting* ovat eräitä luokittelutapoja, jotka muodostavat useamman mallin kokonaisuuden. Lyhyesti muotoiltuna, näissä kahdessa menetelmässä opetusaineisto jaetaan useampaan osaan, joista jokaisella estimoidaan erikseen jokin malli. *Baggingissa* nämä osajoukot (pussit) muodostetaan satunnaisotoksina takaisinpanolla opetusaineistosta (*bootstrapping*). Näillä jokaisella aineistolla opetetaan jotakin mallia, joista muodostetaan ennuste. Viimeiseksi, näistä ennusteista kootaan lopullinen ennuste, esimer-

[7] *Keinotekoinen neuroverkko on matemaattinen malli biologistisesta neuroverkosta, kuten esimerkiksi ihmisen aivoista. Neuroverkot oppivat (tai oikeastaan ne laitetaan oppimaan) samalla tavalla kuin ihmisetkin. Onnistuminen vahvistaa suoritusta ja sen todennäköisyyttä, kun taas epäonnistumisen jälkeen neuroverkko pyrkii korjaamaan virheettään siten, että seuraavalla kerralla virheen todennäköisyys on pienempi. Neuroverkkojen käytöstä taloustieteessä löytyy lisää esimerkiksi artikkelista Kaastra ja Boyd (1996).*

Kuvio 1. Bagging-algoritmin kulkukaavio



kiksi ottamalla keskiarvo kaikkien eri regressiomallien ennusteista. Jos mallina *baggingissa* käytetään luokittelijaa, niin lopullisesta mallien aggregoinnista puhutaan äänestämisenä (*voting*). Yksi suosittu luokittelumalli *baggingissa* on päätöspuu. Tälle useista puista koostuvalle algoritmille on annettu oma nimikin, satunnainen metsä (*random forest*).

Boosting on ikään kuin "viritetty" versio *baggingista*. *Boostingissa* ensimmäinen osajoukko otetaan satunnaisesti ja tällä estimoidaan malli. Tämän jälkeen tuloksista katsotaan, mille alkioille ennuste onnistui heikoiten. Tämän jälkeen näiden heikoiten menestyneiden alkioiden todennäköisyyttä joutua seuraavaan osajoukkoon lisätään. Seuraavaksi algoritmi ottaa satunnaisotoksen tästä painotetusta opetusaineistosta. Tällä otoksella jälleen muodostetaan malli ja katsotaan sen ennustekykyä eri alkioille ja muodostetaan painotukset. Tätä prosessia jatketaan niin kauan, kunnes jokaisella *N*:llä osajoukolla on muodostettu malli. Tämän jälkeen ennusteista tehdään jälleen mallien aggregointi, esimerkiksi painotettuna keskiarvona regressiomallien ennusteista. Kansantajuisesti sanottuna *boosting*-algoritmi opettelee ennustamaan myös "vaikeimmat" havainnot. Siksipä sitä on testattu esimerkiksi taantumien ennustamiseen Ngin (2014) ja Döpken ym. (2017) toimesta.

Molemmissa malleista voidaan tehdä myös yhdistetty malli, jota voidaan testata testiaineistoon. Tutkija voi siten käyttää *baggingia* ja *boostingia* erilaisilla opetus- ja testiaineistojaoilla (esimerkiksi käyttäen ristiinvalidointia) ja testata näin erilaisten mallien toimivuutta aineiston ulkopuolisessa ennustamisessa. Tämä tuo oppimiseen niin sanotusti yhden kerroksen lisää.

Baggingia ja *boostingia* on käytetty paljon makromuuttujien ennustamiseen. Inoue ja Kilian (2008) käyttivät *baggingia* Yhdysvaltojen kuluttajahintaindeksin ennustamisessa. Rapach ja Strauss (2010) hyödynsivät sitä työttö-

myyden kasvun ennusteissa, Hillebrand ja Medeiros (2010) osakevolatiliteetin ennusteissa ja Audrino ja Medeiros (2011) lyhyiden korkojen ennusteissa. Stock ja Watson (2012) esittelivät yleisiä kutistamismenetelmiä (*shrinkage methods*), joissa on mukana myös *bagging*. Jordan ym. (2017) käyttivät *baggingia* osaketuottojen ennustamiseen makromuuttujilla. Jokaisessa näissä tutkimuksissa havaittiin *baggingin* olevan varteenotettava työkalu ennustamistarkoitukseen.

Bai ja Ng (2009) käyttivät *boostingia* valitsemaan ennustavat muuttujat USA:n inflaation, Fed Funds -koron, teollisuustuotannon kasvun ja työttömyyden ennusteisiin. Heidän tuloksensa osoittivat, että eräät *boosting*-mallit tuottavat parempia ennusteita kuin autoregressiiviset mallit. Shafikin ja Tutzin (2009) työttömyyden ennustevirheet pienenivät, kun he hyödynsivät *boostingia* epälineaaristen mallien valintaan. Buchenin ja Wohlraben (2011; 2014) mukaan *boostingilla* tehdyt mallien identifiointi ja muuttujien valinta tuottivat "kilpailukykyisiä" ennusteita useille makromuuttujille. Kim ja Swanson (2014) ennustivat 11 makromuuttujaa ja vertailivat näillä suurta valikoimaa erilaisia malleja ja supistamismenetelmiä, kuten *boostingia* ja *baggingia* sekä seuraavaksi esiteltäviä elastista verkkoa ja *ridge*-regressiota. Heidän saamiensa tulosten mukaan näillä koneoppimisen menetelmillä tehdyt ennusteet omasivat pienimmät ennustevirheet.

Muita erilaisia mallin ja muuttujien valinta -algoritmeja on paljon, kuten Akaiken informaatiokriteeri (AIC) ja bayesiläinen informaatiokriteeri (BIC) tai muut bayesiläiset luokittelijat ja keskiarvoistukset.[8] Näistä algoritmeista voi myös muodostaa oman algoritmin, joka tekee jokaisella menetelmällä ennusteensa testiaineistolle ja sen jälkeen keskiarvoistaa nämä menetelmät. Tätä menetelmää kutsutaan kokonaisuusoppimiseksi (*ensemble learning*). *Bagging* ja *boosting* ovat siten myös eräitä kokonaisuusoppimisen menetelmiä. Koneoppimisessa on siis monta kerrosta lähtien aineiston jaosta erilaisiin osajoukkoihin ja päätyen algoritmien yhdistämiseen. Saattaa kuulostaa monimutkaiselta, mutta tietokone saadaan tekemään tämä kaikki yllättävänkin nopeasti ja yksinkertaisesti ohjelmoiden.[9]

Koneoppimisen menetelmiä voidaan käyttää myös hieman tutummissakin tilastotieteen ympäristöissä. Muuttujien valinnan sijaan mallia voidaan myös niin sanotusti silottaa (*smoothing*) ja siten lähestyä ylisovittamisen ongelmia. Toisin sanoen mallia yksinkertaistetaan siten, että heikosti selittävien muuttujien roolia pienennetään. Näistä menetelmistä ehkäpä tunnetuimpia ovat Hoerlin (1962) kehittämä *ridge*-regressio ja Tibshiranin (1996) kehittämä LASSO-regressio.[10] Usein näiden kahden yhdistelmästä puhutaan elastisen verkon regressiona, jonka avulla voidaan tietyllä parametrisoinnilla muodostaa *ridge*-, LASSO- tai pienimmän neliösumman estimaattorit. Elastinen verkko, ridge ja LASSO eroavat pienimmän neliösumman menetelmästä siten, että virhetermien neliösummaa minimoitaessa otetaan mukaan myös niin sanottu rangaistuster-

[8] *Näitä on esitellyt kattavasti esimerkiksi Castle ym. 2009, jotka vertailevat 21 erilaista algoritmia.*

[9] *R on tilastolliseen laskentaan ja grafiikan tuottamiseen tarkoitettu vapaa ohjelmistoympäristö. R-ohjelmisto on suosittu koneoppimisen menetelmien käytössä. Esimerkiksi paketista adabag löytyy valmiit funktiot baggingiin ja boostingiin.*

[10] *LASSO on lyhennys ilmaisusta Least Absolute Shrinkage and Selection Operator.*

mi (*penalty term*), jonka suuruus riippuu selittävien tekijöiden määrästä. Rangaistustermin avulla osa regressiokertoimista supistuu kohti nollaa. Elastisen verkon rangaistustermin optimointiin voidaan käyttää esimerkiksi ristiinvalidointia. Tällä keinolla perinteisessä lineaarisessa regressioanalyysissa pystytäänkin kutistamaan mallia siten, että vain merkittävimmät selittäjät jäävät jäljelle. Esimerkiksi Hofmarcher ym. (2011) sekä Schneider ja Wagner (2012) käyttävät tätä menetelmää talouskasvun ajureiden selvittämiseen.

Spike ja *slab* -regressio on myös yksi käyttöönotettu valintamenetelmä mallin muuttujille. Tätä bayesiläistä lineaarisen mallin muuttujien valintaa esittivät ensimmäisenä Mitchell ja Beauchamp (1988). Sittemmin sitä ovat viimeistelleet muun muassa Ishwaran ja Rao (2005). Yksinkertaisuudessaan menetelmän tarkoitus on antaa priori-jakauma todennäköisyyksille, millä muuttujat ovat mukana mallissa (*spike*) ja priori-jakauma kertoimien suuruuksille (*slab*). Näitä prioreita ja uskottavuusfunktiota hyväksikäyttäen pystytään tavanomaisin bayesiläisin keinoin simuloimaan posteriori-jakaumat todennäköisyyksille, millä muuttuja on mukana mallissa sekä kertoimien suuruuksille.[11] Esimerkiksi Scott ja Varian (2015) käyttivät tätä menetelmää yhdessä mallin keskiarvoistuksen ja Kalman-suodattimen kanssa lyhyen aikavälin ennustamiseen käyttäen Google-hakuaineistoa. Tätä menetelmää Scott ja Varian (2014; 2015) kutsuvat rakenteelliseksi bayesiläiseksi muuttujan valintamenetelmäksi aikasarjoille (*Bayesian Structural Time Series*).

[11] *Simuloinnissa voidaan käyttää esimerkiksi Markov Chain Monte Carlo -menetelmää (MCMC).*

Nämä muutamat työkalut ovat vain murto-osa valtavasta koneoppimisen algoritmien skaalasta. Chakraborty ja Joseph (2017) tarjoavat laajan ja yksityiskohtaisen katsauksen koneoppimisen välineistä makrotaloustieteessä ja keskuspankkitoiminnassa. Myös Hal Varianin tutkimukset *big datan* ja ekonometrian parissa ovat olleet monelta osin suosittuja (Varian 2010; 2014; Choi ja Varian 2012; Scott ja Varian 2014; 2015). Näistä *Journal of Economic Perspectives* -lehdessä 2014 julkaistu katsausartikkeli "Big Data: New Tricks for Econometrics" käy läpi ekonometrisia työkaluja koneoppimisen ja *big datan* perspektiivistä. Varianin mukaan koneoppiminen tarjoaa uusia tulokulmia aineiston tutkimiseen ja sitä kautta löytyvien mallien spesifiointiin.

Yleisesti koneoppimisen malleja näkyy makrotaloustieteellisessä tutkimuksessa ja ennustamisessa suhteellisen vähän. Koneoppimisen soveltaminen ei aina ole mutkatonta, vaikka sitä kautta saisikin varteenotettavia vaihtoehtoja mallin ja muuttujien valintaan sekä ylisovittamisen ongelmiin. Koneoppimisen mallit kohtaavat samat tilastollisten menetelmien haasteet kuten esimerkiksi puuttuvan muuttujan harhan. Tästä syystä kausaalipäättely on vaikeaa myös koneoppimisen näkökulmasta. Lisäksi on mahdollista, että selittävien tekijöiden määrän kasvaessa algoritmit saattavat tarjota harhaanjohtavia yhteyksiä muuttujien välille. Jos malliin valitaan useita, mahdollisesti satoja erilaisia selittäviä tekijöitä, voi olla vain sattumaa, että joku näistä on tilastollisesti merkitsevä selittävä tekijä. Tähän voi tosin olla jälleen ratkaisuna ristiinvalidointi. Koneoppimisen malleissa ja erityisesti tiedonlouhinnassa on mahdollista, ettei koneen tarjoamille malleille ja muuttujille välttämättä löydy intuitiivista selitystä tai teorian tukea.

Lisäksi moniulotteisten epälineaaristen mallien tulkinta on useimmiten vaikeaa ja siksi koneoppimismallit tarvitsevat usein rinnalleen kontekstisidonnaista, kontrolloitua testaamista. Muutamia koneoppimisen ennustemalleja vertailevia tutkimuksia ovat esimerkiksi Swanson ja White (1997), Hand ja Henley (1997), Ahmed ym. (2010), Bontempi ym. (2013) ja Taieb ym. (2012). Näistä tutkimuksista löytyy myös koneoppimisen mallien rakenteiden yksityiskohtaisempia tarkasteluja.

3. Pohdintaa

Big data yhdessä perinteisten aineistojen kanssa nähdään yleisesti asiana, joka voi johtaa syvempään ymmärrykseen taloudellisista ilmiöistä ja niiden vuorovaikutuksista. Niin sanottu informaatiovallankumous tarjoaa uusia mahdollisuuksia datan havaitsemiseen ja analysointiin, mutta se vaatii myös uudenlaista osaamista, teknologiaa ja määrätietoisia organisaatioita, jotka pyrkivät hyödyntämään alati kasvavan informaation. Kuten tässä katsauksessa esitetty kirjallisuus antaa ymmärtää, taloustieteessä mahdollisuuksia ja sovelluskohteita on monia.

Etenkin ennustamisen osalta on tapahtunut edistystä. Uusia koneoppimisen menetelmiä on alettu testata ja uusia *big data* -aineistoja on otettu käyttöön. Yksi syy, miksi ennustemallien suosio ja tarve on kasvanut, on se, että lyhyen aikavälin ennustemalleilla pystytään ennakoimaan harvakseltaan julkaistavia tilastoja, joiden perusteella pystytään harjoittamaan politiikkaa. Esimerkkinä voidaan ottaa Suomen Pankin uusi bayesiläinen vektoriautoregressiivinen (BVAR) Suomen bruttokansantuotteen ennustemalli, joka hyödyntää lähes 50 muuttujaa ja ennuste päivittyy reaaliaikaisesti muuttujien uusien tilastojen myötä. Ennusteen lisäksi mallilla voidaan arvioida erikseen jokaisen eri muuttujan uusien tilastojulkistusten vaikutusta ennusteeseen. Tätä ja muita Suomen Pankin lyhyenaikavälin ennustemalleja on kattavasti esitellyt Itkonen (2016) sekä Itkonen ja Juvonen (2017).

Keskuspankin viestintä on tärkeä osa rahapolitiikkaa. Tästä syystä myös viestinnän sisältöä on alettu analysoimaan uusilla menetelmillä, joissa tekstistä pyritään löytämään sisältöä parhaiten kuvaavat piirteet. Tämän avulla pystytään tulkitsemaan keskuspankin ennakoivaa viestintää tai vastaavasti markkinoiden reaktioita uutisoinnin perusteella. Eräitä mielenkiintoisia julkaisuja aiheesta ovat tehneet Hansen ja McMahon (2016), Hansen ym. (2017) ja Tobback ym. (2017).

Myös valvontaviranomaiset hyötyvät suuremmassa mittakaavassa uusista suurista aineistoista pankkien vakavaraisuuden, riskien ja talouden vakauden monitoroinnissa ja tutkimuksessa. Siksipä esimerkiksi koneoppiminen saattaa esittäytyä hyödyllisenä työkaluna markkinoiden ja valvottavien ongelmien tunnistamisessa, missä nopea ja reaaliaikainen monitorointi on ensisijaisen tärkeää (Flood ym. 2016).

Erityisesti finanssikriisin myötä keskuspankkien uusien yksityiskohtaisempien tilastojen tarve makro- ja mikrovakaudesta on kasvanut voimakkaasti. Monet keskuspankit, kuten Euroopan keskuspankki, Yhdysvaltojen Federal Reserve ja Englannin keskuspankki, ovatkin alkaneet tehdä selkeitä organisaation muutoksia, joissa on tarkoitus järjestelmällisesti ohjata keskuspankkitoimintaa siten, että se huomioisi paremmin kasvavien tilastojen uudet mahdollisuudet. Näillä osastoilla henkilöstön osaamista tarvitaan niin talous- ja tilastotieteen saralta, kuin myös tietojenkäsittelytieteestä.

Suuret datamassat ovat taloustieteelle paljon muutakin kuin Google-hakuaineistot ja sosiaalisen median data. Erilaisten datalähteiden ja -muotojen määrä on kasvanut valtavasti. Ohjelmistokehittäjä ja yrittäjä sekä nykyään Googlen hallituksen puheenjohtajana toimiva Eric Schmidt (2010) väittikin seuraavaa: "Vuoteen 2003 mennessä ihmiskunta oli tuottanut 5 exabittiä (5 miljardia gigabittiä) informaatiota. Nykyisin tämä sama määrä informaatiota tuotetaan kahdessa päivässä, ja sen vauhti kiihtyy edelleen." Tämä kasvanut datamassa pystyy varmasti tarjoamaan lisää informaatiota niin mikro- kuin makrotaloustieteenkin tutkimukseen. Yliopistoista esimerkiksi MIT ja Brown tarjoavat nykyisin talous- ja tietojenkäsittelytieteen yhdistettyjä koulutusohjelmia.

Suurien tietomassojen kasvaessa ekonometrian menetelmien jatkuva kehittäminen on entistä tärkeämmässä roolissa taloustieteen tutkimuksen kannalta. Tämä todennäköisesti vaatii entistä tiiviimpää yhteistyötä muiden tieteenalojen, kuten tilasto- ja tietojenkäsittelytieteen kanssa. □

Kirjallisuus

Ahmed, N. K., Atiya, A. F., Gayar, N. E. ja El-Shishiny, H. (2010), "An empirical comparison of machine learning models for time series forecasting", *Econometric Reviews*, 29(5-6): 594–621.

Anderson, E., Malin, B. A., Nakamura, E., Simester, D. ja Steinsson, J. (2017), "Informational rigidities and the stickiness of temporary Sales", *Journal of Monetary Economics*, 90: 64–83.

Antenucci, D., Cafarella, M., Levenstein, M. ja Shapiro, M. (2014), "Using Social Media to Measure Labor Market Flows", NBER Working Paper 20010.

Anvik, C. ja Gjelstad, K. (2010), "*Just Google it: Forecasting Norwegian Unemployment Figures with Web Queries*", CREAM Publication 11/2010, Norwegian Business School.

Askitas, N. ja Zimmermann, K. (2009), "Google Econometrics and Unemployment Forecasting", *Applied Economics Quarterly* 55: 107–120.

Askitas, N. ja Zimmermann, K. (2015), "The internet as a data source for advancement in social sciences", *International Journal of Manpower*, 36(1): 2–12.

Audrino, F. ja Medeiros, M. C. (2011), "Modeling and forecasting short-term interest rates: The benefits of smooth regimes, macroeconomic variables, and bagging", *Journal of Applied Econometrics* 26.6: 999–1022.

Bai, J. ja Ng, S. (2009), "Boosting diffusion indices", *Journal of Applied Econometrics*, 24(4): 607–629.

Baker, S. ja Fradkin, A. (2017), "The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data", *Review of Economics and Statistics* (hyväksytty).

Bollen, J., Mao, H. ja Zeng, X.-J. (2011), "Twitter Mood Predicts the Stock Market", *Journal of Computational Science* 2: 1–8.

Bontempi, G., Taieb, S. B. ja Le Borgne, Y. A. (2013), "Machine learning strategies for time series forecasting", *In Business Intelligence*, Springer Berlin Heidelberg: 62–77.

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A. ja Weber, I. (2012), "Web Search Queries Can Predict Stock Market Volumes", *PloS One* 7: e40014.

Breiman, L., Friedman, J. H., Olshen, R. A. ja Stone, C. J. (1984), "Classification and Regression Trees", Wadsworth and Brooks / Cole, Monterey.

Buchen, T. ja Wohlrabe, K. (2011), "Forecasting with many predictors: Is boosting a viable alternative?", *Economics Letters* 113.1: 16–18.

Castle, J. L., Qin, X. ja Reed, W. R. (2009), "How to Pick the Best Regression Equation: A Review and Comparison of Model Selection Algorithms", Working Paper 13/2009, Department of Economics and Finance, University of Canterbury, Christchurch.

Cavallo, A. (2017), "Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers", *American Economic Review*, 107(1): 283–303.

Cavallo, A. ja Rigobon, R. (2016), "The Billion Prices Project: Using Online Prices for Measurement and Research", J*ournal of Economic Perspectives* 30(2): 151–178.

Chadwick, M. ja Sengul, G. (2012), "Nowcasting Unemployment Rate in Turkey: Let's Ask Google", Central Bank of the Republic of Turkey Working Paper 12/18.

Chakraborty, C. ja Joseph, A. (2017), "Machine Learning at Central Banks", Bank of England Working Paper No. 674.

Cheng, X, ja Hansen, B. (2015), "Forecasting with factor-augmented regression: A frequentist model averaging approach", *Journal of Econometrics* 186.2: 280–293.

Choi, H. ja Varian, H. (2009), "Predicting Initial Claims for Unemployment Benefits", Google https://static.googleusercontent.com/media/research.google.com/en//archive/papers/initialclaimsUS.pdf (viitattu 11.9.2017).

Choi, H. ja Varian, H. (2012), "Predicting the Present with Google Trends", *Economic Record* 88: 2–9.

Cohen, P., Hahn, R., Hall, J., Levitt, S. ja Metcalfe, R. (2016), "Using Big Data to Estimate Consumer Surplus: The Case of Uber", *NBER Working Paper Series*, 42.

Curme, C., Preis, T., Stanley, H. ja Moat, H. (2014), "Quantifying the Semantics of Search Behavior Before Stock Market Moves", *Proceeding of the National Academy of Science of the United States of America* 111: 11600–11605.

D'Amuri, F. (2009), "Predicting Unemployment in Short Samples with Internet Job Search Query Data", MPRA Paper No. 18403, Munich Personal RePEc Archive, http://mpra.ub.uni-muenchen.de/18403/ (viitattu 11.9.2017).

D'Amuri, F. ja Marcucci, J. (2012), "The Predictive Power of Google Searches in Forecasting Unemployment", Bank of Italy Working Paper 891.

Donaldson, D. ja Storeygard, A. (2016), "The view from above: Applications of satellite data in economics", *Journal of Economic Perspectives* 30.4: 171–198.

Döpke, J., Fritsche, U. ja Pierdzioch, C. (2017), "Predicting recessions with boosted regression trees", *International Journal of Forecasting* 33.4: 745–759.

Einav, L. ja Levin, J. (2014), "Economics in the age of big data", *Science*, 346(6210): 1243089.

Ettredge, M., Gerdes, J. ja Karuga, G. (2005), "Using Web-Based Search Data to Predict Macroeconomic Statistics", *Communications of the ACM* 48: 87–92.

Flood, M., Jagadish, H. V. ja Raschid, L. (2016), "Big Data Challenges and Opportunities in Financial Stability Monitoring", *Financial Stability Review* 20: 129–142.

Fondeur, Y. ja Karamé, F. (2013), "Can Google Data Help Predict French Youth Unemployment?", *Economic Modelling* 30: 117–125.

Gee, L. K., Jones, J. J. ja Burke, M. (2017a), "Social Networks and Labor Markets: How Strong Ties Relate to Job Finding On Facebook's Social Network", *Journal of Labor Economics*, 35(2): 485–518.

Gee, L. K., Jones, J. J., Fariss, C. J., Burke, M. ja Fowler, J. H. (2017b), "The paradox of weak ties in 55 countries", *Journal of Economic Behavior and Organization*, 133: 362–372.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M. ja Brilliant, L. (2009), "Detecting Influenza Epidemics Using Search Engine Query Data", *Nature* 457(7232): 1012–1014.

Goel, S., Hofman, J., Lahaie, S., Pennock, D. ja Watts, D. (2010), "Predicting Consumer Behavior with Web Search", *Proceedings of the National Academy of Science of the United States of America* 107: 17486–17490.

Guzman, G. (2011), "Internet Search Behavior as an Economic Forecasting Tool: The Case of Inflation Expectations", *Journal of Economic and Social Measurement* 36: 119–167.

Hall, J. V. ja Krueger, A. B. (2016), "An Analysis of the Labor Market for Uber's Driver-Partners in the United States", *NBER Working Paper No. 22843*.

Hamermesh, D. (2013), "Six Decades of Economics Publishing: Who and How?", *Journal of Economic Literature* 51: 162–172.

Hand, D. J. ja Henley, W. E. (1997), "Statistical classification methods in consumer credit scoring: a review", *Journal of the Royal Statistical Society: Series A* (Statistics in Society), 160(3): 523–541.

Hansen, S. ja McMahon, M. (2016), "Shocking language: Understanding the macroeconomic effects of central bank communication", Journal of International Economics 99: S114–S133.

Hansen, S., McMahon, M. ja Prat, A. (2017), "Transparency and deliberation within the FOMC: a computational linguistics approach", The Quarterly Journal of Economics (qjx045).

Henderson, J. V., Storeygard, A. ja Weil, D. N. (2012), "Measuring economic growth from outer space", *American economic review* 102.2: 994–1028.

Hillebrand, E. ja Medeiros, M. C. (2010), "The Benefits of Bagging for Forecast Models of Realized Volatility", *Econometric Reviews*, 29(5–6): 571–593.

Hoerl, A. E. (1962), "Application of Ridge Analysis to Regression Problems", *Chemical Engineering Progress Symposium Series* 1958: 54–59.

Hofmarcher, P., Crespo, J. C., Grun, B. ja Hornik, K. (2011), "Fishing Economic Growth Determinants Using Bayesian Elastic Nets", Research Report Series 113, Department of Statistics and Mathematics, Vienna University of Economics and Business.

Hong, G. H. ja Li, N. (2017), "Market structure and cost pass-through in retail", *The Review of Economics and Statistics*, 99(1): 151–166.

Inoue, A. ja Kilian, L. (2008), "How useful is bagging in forecasting economic time series? A case study of US consumer price inflation", *Journal of the American Statistical Association* 103.482: 511–522.

Ishwaran, H. ja Rao, S. J. (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies", *The Annals of Statistics*, 33: 730–773.

Itkonen, J. (2016), "Mistä tiedämme, miten taloudessa menee tänään", *Euro & Talous* 3/2016.

Itkonen, J. ja Juvonen, P. (2017), "Nowcasting the Finnish economy with a large Bayesian vector autoregressive model", Bank of Finland Economics Review 6/2017.

Jordan, S. J., Vivian, A. ja Wohar. M. E. (2017), "Forecasting market returns: bagging or combining?", International Journal of Forecasting 33.1: 102–120.

Kaastra, I. ja Boyd, M. (1996), "Designing a neural network for forecasting financial and economic time series", *Neurocomputing*, 10(3): 215–236.

Kholodilin, K., Podstawski, A. ja Siliverstovs, B. (2010), "Do Google Searches Help in Nowcasting Private Consumption? A Real-Time Evidence for the US", DIW Berlin Discussion Paper No. 997.

Kim, H. H. ja Swanson, N. R. (2014), "Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence", Journal of Econometrics 178: 352-367.

Koop, G. ja Onorante, L. (2016), "Macroeconomic Nowcasting Using Google Probabilities", First International Conference on Advance Research Methods and Analytics CARMA 2016, Universitat polytècnica de València, heinäkuu 2016, https://www.researchgate.net/publication/305672191 (viitattu 11.9.2017).

Kortelainen, M., Raychaudhuri, J. ja Roussillon, B. (2016), "Effects of carbon reduction labels: Evidence from scanner data", *Economic Inquiry*, 54(2): 1167–1187.

Kroft, K. ja Pope, D. G. (2014), "Does Online Search Crowd Out Traditional Search and Improve Matching Efficiency? Evidence from Craigslist", *Journal of Labor Economics*, 32(2): 259–303.

Kuhn, P. ja Mansour, H. (2014), "Is Internet Job Search Still Ineffective?", *Economic Journal* 124: 1213–1233.

Kuhn, P. ja Skuterud, M. (2004), "Internet Job Search and Unemployment Durations", *American Economic Review* 94: 218–232.

Kulkarni, R., Haynes, K., Stough, R. ja Paelinck, J. (2009), "Forecasting Housing Prices with Google Econometrics", GMU School of Public Policy Research Paper No. 2009–10.

Laouénan, M. ja Rathelot, R. (2017), "Ethnic Discrimination on an Online Marketplace of Vacation Rentals", *University of Warwick*, http://rolandrathelot.com/wp-content/uploads/Laouenan.Rathelot.Airbnb.pdf (viitattu 9.10.2017).

Lendle, A., Olarreaga, M., Schropp, S. ja Vézina, P. L. (2016), "There Goes Gravity: eBay and the Death of Distance", *Economic Journal*, 126(591): 406–441.

McLaren, N. ja Shanbhogue, R. (2011), "Using Internet Search Data as Economic Indicators", *Bank of England Quarterly Bulletin* 2011/Q2: 134–140.

Mitchell, T. J. ja Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression", *Journal of the American Statistical Association* 83: 1023–1032.

Moat, H., Curme, C., Avakian, A., Kennett, D., Stanley, H. ja Preis, T. (2013), "Quantifying Wikipedia Usage Patterns Before Stock Market Moves", Scientific Reports 3: 1–5.

Ng, S. (2014), "Boosting recessions", *Canadian Journal of Economics/Revue canadienne d'économique* 47.1: 1–34.

Pavlicek, J. ja Kristoufek, L. (2015), "Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries", PLoS ONE 10(5): e0127084.

Powell, B., Nason, G., Elliott, D., Mayhew, M., Davies, J. ja Winton, J. (2017), "Tracking and modelling prices using web-scraped price microdata: towards automated daily consumer price index forecasting", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Preis, T., Moat, H. ja Stanley, H. (2013), "Quantifying Trading Behavior in Financial Markets using Google Trends", *Scientific Reports* 3: 1–6.

Rapach, D. E. ja Strauss, J. K. (2010), "Bagging or combining (or both)? An analysis based on forecasting US employment growth", *Econometric Reviews* 29.5–6: 511–533.

Schmidt, E. (2010), "A New Philosophy of Progress", The Techonomy Conference, Lake Tahoe, California, August 6, 2010, http://techonomy.com/tag/eric-schmidt/ (viitattu 23.09.2017).

Schneider, U. ja Wagner, M. (2012), "Catching Growth Determinants with the Adaptive LASSO", *German Economic Review* 13: 71–85.

Scott, S. ja Varian, H. (2014), "Predicting the Present with Bayesian Structural Time Series", *International Journal of Mathematical Modelling and Numerical Optimisation* 5: 4–23.

Scott, S. ja Varian, H. (2015), "Bayesian Variable Selection for Nowcasting Economic Time Series", teoksessa Goldfarb, A., Greenstein, S. ja Tucker, C. (toim.), *Economic Analysis of the Digital Economy*, Chicago University Press: 119–136.

Shafik, N. ja Tutz, G. (2009), "Boosting nonlinear additive autoregressive time series", *Computational Statistics & Data Analysis* 53.7: 2453–2464.

Stephens-Davidowitz, S. (2014), "The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data", *Journal of Public Economics* 118: 26–40.

Stevenson, B. (2008), "The Internet and Job Search", *NBER Working Paper* 13886.

Stock, J. H. ja Watson, M. W. (2012), "Generalized shrinkage methods for forecasting using many predictors", *Journal of Business & Economic Statistics* 30.4: 481-493.

Swanson, N. R. ja White, H. (1997), "A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks", *The Review of Economics and Statistics*, 79(4): 540–550.

Taieb, S. B., Bontempi, G., Atiya, A. F. ja Sorjamaa, A. (2012), "A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition", *Expert systems with applications*, 39(8): 7067–7083.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society* B 58: 267–288.

Tobback, E., Nardelli, S. ja Martens, D. (2017), "Between hawks and doves: measuring central bank communication", *ECB Working Paper No. 2085*.

Tuhkuri, J. (2014), "Big Data: Google Searches Predict Unemployment in Finland", *ETLA Reports 31*.

Tuhkuri, J. (2015), "Big Data: Do Google Searches Predict Unemployment?", Helsingin yliopisto, https://helda.helsinki.fi/handle/10138/155258 (viitattu 11.9.2017).

Tuhkuri, J. (2016), "A Model for Forecasting with Big Data – Forecasting Unemployment with Google Searches in Europe", *ETLA Reports 54*.

Utans, J., Moody, J., Rehfuss, S., ja Siegelmann, H. (1995), "Input variable selection for neural networks: Application to predicting the US business cycle", In *Computational Intelligence for Financial Engineering, Proceedings* of the IEEE/IAFE, 118–122.

Varian, H. (2010), "Computer Mediated Transactions", *American Economic Review: Papers & Proceedings* 100: 1–10.

Varian, H. (2014), "Big data: New tricks for econometrics", *Journal of Economic Perspectives* 28: 3–36.

Vicente, M., López-Menéndez, A. ja Pérez, R. (2015), "Forecasting Unemployment with Internet Search Data: Does it Help to Improve Predictions When Job Destruction is Skyrocketing?", *Technological Forecasting & Social Change* 92: 132–139.

Vosen, S. ja Schmidt, T. (2011), "Forecasting Private Consumption: Survey Based Indicators vs. Google Trends", *Journal of Forecasting* 30: 565–578.

Vosen, S. ja Schmidt, T. (2012), "A Monthly Consumption Indicator for Germany based on Internet Search Query Data", *Applied Economics Letters* 19: 683–687.

Widgrén, J. (2016), "Google-haut Suomen asuntojen hintojen ennustajana", *ETLA Raportit – Reports 63*.

Wohlrabe, K., ja Buchen, T. (2014), "Assessing the macroeconomic forecasting performance of boosting: evidence for the United States, the Euro area and Germany", *Journal of Forecasting*, 33(4): 231–242.

Wu, L. ja Brynjolfsson, E. (2015), "The future of prediction: How Google searches foreshadow housing prices and sales", *In Economic analysis of the digital economy*, University of Chicago Press: 89–118.