Jaakko Orola

MASSADATA JA KONEOPPIMINEN MAKROTALOUSTIETEESSÄ

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2020

TIIVISTELMÄ

Orola, Jaakko
Massadata ja koneoppiminen makrotaloustieteessä
Jyväskylä: Jyväskylän yliopisto, 2020, 27 s.
Tietojärjestelmätiede, Kandidaatintutkielma
Ohjaaja: Halttunen, Veikko

Informaatioyhteiskunta tuottaa itsestään jatkuvasti kasvavalla nopeudella tietoa, jota on mahdollista hyödyntää uusien menetelmien, kuten koneoppimisen avulla. Taloustieteilijät ovat viimeisten vuosikymmenten aikana kehittäneet tapoja tehdä talouden ennusteita käyttäen useita erilaisia tiedonlähteitä samanaikaisesti. Tämä kirjallisuuskatsauksena toteutettu tutkielma vastaa kysymykseen, kuinka suuria datamassoja voidaan hyödyntää makrotaloustieteessä, ja kuinka koneoppimisen menetelmät soveltuvat korvaamaan ja täydentämään makrotaloustieteen perinteisesti käyttämiä ekonometrian menetelmiä ennustamisessa. Tutkimuksessa havaittiin, että prosessi hyödyntää koneoppimisen menetelmiä täysin on makrotaloustieteessä edelleen vaiheessa. Tutkimustulokset osoittavat, että esimerkiksi verkkoharavoinnilla hankittu data ja hakukonedata sisältävät informaatiota, jota perinteisistä tietolähteistä ei löydy. Hadoop ja NoSQL-tietokannat osoittautuvat tärkeiksi datanhallinnan työkaluiksi. Monet uudet tiedonlähteet sopivat reaaliaikaiseen ennustamiseen, sillä dataa on julkisesti tarjolla päivittäistasolla.

Asiasanat: Big Data, makrotaloustiede, reaaliaikainen ennustaminen, koneoppiminen

ABSTRACT

Orola, Jaakko
Big Data and machine learning in macroeconomics
Jyväskylä: University of Jyväskylä, 2020, 27 pp.
Information systems, Bachelor's thesis
Supervisor: Halttunen, Veikko

The modern information society creates data about itself at an ever-increasing pace. With emerging technologies like machine learning, it is possible to make use of this data. For the last three decades, economists have developed models that predict using a multiple data source approach. This literature review answers the question how Big Data can be utilized in macroeconomics and how machine learning technologies can complement or replace econometrical methods in prediction. The process of utilizing machine learning in macroeconomics was found to be incomplete at the time of this review. The results show that data gathered with web scraping and search engine statistics contain information that is not present in contemporary datasets. Apache Hadoop and NoSQL databases prove to be important tools in managing Big Data. Many new data sources that can be collected at a high frequency are useful in macroeconomic nowcasting.

Keywords: Big Data, Macroeconomics, Nowcasting, Machine learning

SISÄLLYS

1 JOHDANTO

Taloustieteessä suurten data- ja muuttujamäärien käsittely on ollut keskeinen ongelma jo kauan verrattuna moniin muihin tieteen aloihin (Bok, 2018). Tutkielma keskittyy massadatan hyödyntämiseen makrotaloustieteessä sekä ekonometriassa. Lähestyn aihetta tietojärjestelmätieteen näkökulmasta, sillä aiemmat tutkimukset ovat keskittyneet taloustieteelliseen ja tilastotieteelliseen näkökulmaan. Kirjallisuuskatsaukseen on koostettu massadatan mahdollistamia uusia tapoja laskea taloustieteen mittareita. Menetelmiä vertaillaan perinteisesti käytettyihin menetelmiin sekä toisiinsa. Tutkimuksen lähteiksi valitsin laadultaan julkaisufoorumin kriteerit täyttävien julkaisujen vertaisarvioituja tutkimuksia sekä keskuspankkien, erityisesti Euroopan keskuspankin ja Yhdysvaltojen keskusvarannon raportteja vertaisarvioidun tieteellisten vertaisarvioitujen julkaisulähteiden puuttuessa.

Taloustieteen termistö ja seurattavat tunnusluvut ovat vakiintuneita. Uusien menetelmien testaaminen on käytettävissä olevan historiallisen tiedon määrän takia mahdollista. Uusien menetelmien tehokkuutta on mahdollista verrata vallassa oleviin ekonometrian tilastotieteellisiin menetelmiin. Tämän vuoksi taloustiede sopii tieteenalana koneoppimisen kehittämiseen ja sen toimivuuden arviointiin. Lisäksi tietoyhteiskunnan uusista tiedonlähteistä kerätty data sisältää informaatiota, mikä voi parantaa ennusteiden tarkkuutta. Agrawalin (2019) mukaan tulevaisuudessa yksi tekoälyn suurimmista hyödyistä on ennustamisen kehittyminen.

Motivaatio alan tutkimukseen oli myös tarve arvioida automatisaation vaikutuksia taloustieteen asiantuntijatyöhön. Siitä huolimatta, että taloustiedettä on tutkittu pitkään, eivät taloustieteilijöiden ennusteet ja laskentamenetelmät ole olleet tarkkoja, eivätkä sekä pitkän ajan ennusteet että äkillisten talouskriisien ennakoiminen ole sen takia toimineet. Informaatioteknologian menetelmiä, joita käytetään taloustieteessä ei ole ennen koostettu yksittäiseen tutkimukseen näkökulmasta, joka erittelee ja arvioi samalla teknologioita suoraan sen taustalla olevasta IT-infrastruktuurista asti.

Keskeiseksi tutkimuskysymykseksi muotoutui heti tutkimuksen alusta alkaen:

- Miten massadataa ja koneoppimista voidaan hyödyntää makrotaloustieteessä?

Käytin tiedonhankintaan seuraavia hakutietokantoja: Google Scholar, Scopus ja JYKDOK:in kansainvälisten e-aineistojen haku. Olennaisiksi hakusanoiksi muodostuivat seuraavat sekä niiden yhdistelmät: Big Data (massadatan englanninkielinen kaupallinen ja yleisessä käytössä oleva synonyymi), Nowcasting, Macroeconomic ja Econometrics. Tutkimus suoritettiin kirjallisuuskatsauksena.

Useimmissa alan laadultaan kriteerit täyttävissä tutkimuksissa ei sen syvemmin eritellä niissä käytettyjä teknologioita tai IT-konfiguraatioita. Olemassa olevat kirjallisuuskatsaukset keskittyvät pikemminkin tuloksiin kuin teknologiaan. Tässä kirjallisuuskatsauksessa pyritään antamaan esimerkkejä myös käytetyistä teknologioista sekä selittämään niiden toimintaperiaatteita. Tutkielman ensimmäinen sisältöluku esittelee makrotaloustieteen viime vuosikymmeninä käyttämiä menetelmiä. Toinen sisältöluku keskittyy massadatan prosessin kuvaamiseen sekä erilaisten analysoinnin työkalujen ja koneoppimisen mallien esittelyyn. Kolmas sisältöluku esittelee tähän mennessä aiheesta kertynyttä tutkimuskirjallisuutta.

2 MAKROTALOUSTIETEEN EKONOMETRIA

Makrotaloustieteen ekonometria tasapainottelee mittaamisen ja teorian välillä. Diebold (1998) jakaa ekonometrian ennustamisen menetelmät karkeasti kahteen ryhmään: rakenteellisiin ja rakenteettomiin. Rakenteettoman ennustamisen juuret ovat 1920-luvulla, jolloin taloustieteen mallintamiseen ruvettiin soveltamaan lineaarisia differenssiyhtälöitä stokastisilla shokeilla. Näin pystyttiin ennustamaan taloudellisia aikasarjoja tehokkaasti. Näitä malleja kutsutaan autoregressiivisiksi malleiksi.

2.1 Vektoriautoregressiiviset mallit

Autoregressiiviset mallit ennustavat muuttujien historiallisten aikasarjojen sisältämän tiedon pohjalta. Yksinkertaisessa autoregressiossa muuttujan arvo hetkellä *t* pohjautuu sen historiallisiin arvoihin. Vektoriautoregressiivisessä mallissa (VAR) muuttujan arvo pohjautuu sen omien historiallisten arvojen lisäksi muiden datasetissä olevien muuttujien historiallisiin arvoihin. Historiallista arvoa tietyltä ajanjaksolta kutsutaan viiveeksi (Diebold, 1998). Stock (2001) luettelee kolme erilaista VAR-mallia:

- Supistetun muodon VAR *(reduced form)*
- Rekursiivinen VAR *(recursive)*
- Rakenteellinen VAR *(structural)*

Supistetun muodon VAR-malli esittää jokaisen muuttujan sen historiallisten arvojen määrittämänä lineaarisena funktiona, johon on lisätty näiden kanssa korreloimaton jäännöstermi *u*. Tämän lisäksi oletuksena on, että malliin valittujen eri muuttujien arvot sisältävät toisiinsa liittyvää informaatiota, joten funktioon sisällytetään vielä toisten muuttujien historiallisten arvojen vaikutusta estimoiva termi. Lineaaristen funktioiden arvot lasketaan pienimmän neliösumman menetelmällä. Rekursiivisen VAR -mallin jäännöstermit asetetaan toisis-

taan riippumattomiksi. Esimerkkinä 3 muuttujan VAR, jossa muuttujat käsitellään järjestyksessä inflaatio, työttömyys, korkotaso. Mallin ensimmäisessä yhtälössä vastemuuttujana (selitettävä muuttuja) on inflaatio. Selittäjinä ovat kaikkien muuttujien viiveet. Toisessa yhtälössä vastemuuttujana on työttömyys ja selittävänä muuttujana on kaikkien muuttujien viiveiden lisäksi inflaation nykyarvo. Kolmannessa yhtälössä vastemuuttujana on korkotaso ja selittäjinä ovat kaikkien muuttujien viiveet, inflaation nykyarvo sekä työttömyyden nykyarvo. PNS:llä lasketut yhtälöt tuottavat jokaiseen yhtälöön toisistaan riippumattomat jäännökset. Muuttujien järjestystä muuttamalla myös tulokset muuttuvat, joten n muuttujan rekursiivisella VAR-mallilla voi laskea n! eri tulosta (Stock, 2001). Rakenteellisessa VAR -mallissa muuttujien välisiä suhteita ja kausaalisuutta estimoidaan taloudellisen teorian kautta. Näitä estimaatteja voi lisätä malliin joko yksittäisiin lausekkeisiin tai esimerkiksi koko mallin laajuudelle.

VAR on todettu hyödylliseksi ja luotettavaksi työkaluksi (Stock, 2001). Sitä käytetään laajasti makrotaloustieteen ennustamisessa ja analyysissä. VAR-mallit pystyvät hyödyntämään muuttujien sisältämää informaatiota ja mallintamaan niiden välisiä suhteita suoraan ilman, että tarvitsisi asettaa mallille rajoitteita. Funktioiden määrän ja koon suuri kasvaminen uusia muuttujia lisättäessä kuitenkin pakottaa selittävien muuttujien määrän rajoittamiseen. Tyypilliseen VAR-malliin mahtuu yleensä n. 3–10 eri muuttujaa, sillä tämän jälkeen uusien rajoittamattomien muuttujien lisääminen malliin alkaa aiheuttaa ylisovittumista. Tätä ongelmaa on pyritty korjaamaan suoraan taloustieteen teoriaa käyttämällä, tavoitteena vähentää laskutoimituksiin tarvittavien muuttujien määrää: oletuksena on, että suurin osa yleisesti käytettyjen muuttujien välisistä suhteista on selitettävissä muutamalla yhdistävällä tekijällä. Lisäksi on pyritty rajoittamaan malleja esimerkiksi luokittelemaan muuttujia niiden välisten suhteiden laadun perusteella (Stock, 2001).

2.2 Bayes VAR

Littermanin (1986) mukaan perinteisen frekventistisen mallin heikkous piilee sen ennakko-oletuksissa. Rakentaessaan taloustieteellistä mallia frekventisti joutuu päättämään etukäteen, mitä muuttujia malliin sisällytetään. Muuten mallin tarkkuus kärsii edellisessä kappaleessa kuvattujen ongelmien takia. Näin valitaan teoriaan pohjautuen muutama tärkein muuttuja. Uskotaan, että parhain tulos on saatavilla ainoastaan pitämällä malli tarpeeksi yksinkertaisena. Tästä poikkeava bayesiläinen näkökulma olettaa, että informaatio on todennäköisesti jakautunut suurelle muuttujamäärälle. Bayesiläinen malli pyrkii sisällyttämään laskentakapasiteetin puitteissa kaikki muuttujat, joiden uskotaan sisältävän relevanttia informaatiota, sekä niiden suhteisiin liittyvä etukäteen tiedetty informaatio. Jos väite informaation jakautumisesta suurelle muuttujamäärälle on totta, tuottaa bayesiläinen malli paremman tuloksen, mitä frekventistinen. Bayesiläisen mallinnuksen etuja on, ettei tarvita asiantuntijan manuaa-

lista, mahdollisesti vinoutunutta muokkausta datasetille. Lisäksi bayes -malli tuottaa todennäköisyysjakauman ennusteille.

BVAR (bayes vektoriautoregressio) käsittelee suurta muuttujamäärää sisällyttämällä priorijakauma sen parametreille (Banbura, 2010). Priorijakaumaksi voidaan valita esimerkiksi normaalijakauma tai kaksipuoleinen eksponentiaalijakauma (eng. myös Laplace distribution). Normaalijakauma asettaa jokaiselle datasetissä olevalle selittävälle muuttujalle nollasta poikkeavan korrelaatiokertoimen. Se painottaa vastemuuttujan vaihtelua eniten selittäviä muuttujia, ja antaa vähän relevanttia informaatiota sisältäville muuttujille hyvin pienet painoarvot. Kaksipuoleinen eksponentiaalijakauma taas antaa hyvin suuret painoarvot eniten vastemuuttujaa selittäville muuttujille, mutta asettaa vähemmän informaatiota sisältävien muuttujien painoarvoiksi 0 (De Mol, 2008).

Normaalijakaumallisella priorijakaumalla posteriorijakauman tiheyden ratkaisuongelmaa kutsutaan ridge regressioksi, ja sen yksinkertainen ratkaisualgoritmi on penalisoitu PNS. Kaksipuoleisen eksponentiaalijakauman posterioritiheyden ratkaisuongelmaa taas kutsutaan Lasso regressioksi. Lasson ratkaisemiseksi on kehitelty useita tehokkaiksi luokiteltavia laskenta-algoritmeja, kuten LARS (pienimmän kulman regressio) (De Mol, 2008). Lasso regression toteuttava algoritmi suorittaa sekä muuttujien valinnan, että parametriestimoinnin. Kaksipuoleinen eksponentiaalijakauma on siis tehokas vaihtoehto priorijakaumaksi monille suoritusajaltaan vaativille algoritmeille (De Mol, 2008).

Bayesiläisen lähestymistavan etu tässä tilanteessa on, että se pystyy käsittelemään suuria, satojen muuttujien aineistoja ilman asiantuntijan asettamia rajoitteita mallille tai puuttumista muuttujien valintaprosessiin (De Mol, 2008; Litterman, 1986). Banbura (2010) testasi BVAR-malleja 7, 20 ja 131 muuttujalla ja vertasi niitä VAR – ja pääkomponenttiregressiomalleihin. Suuri 131 muuttujan malli tuotti tarkempia tuloksia kuin 7 muuttujan malli, mutta jo 20 muuttujan malli riitti tässä tapauksessa yhtä hyvän ennusteen sekä rakenneanalyysin tekemiseen.

2.3 Rakenteelliset mallit

Fernández – Villaverden (2010) mukaan uuskeynesiläiset mallit, kuten DSGE-mallit ovat muuttaneet rakenteellisten mallien makroekonometriaa hajanaisiin hypoteettisiin oletuksiin perustuvasta tieteenalasta systemaattiseksi ja nopeasti kehittyväksi. DSGE-malli rakentuu rakenteellisten oletusten, ratkaisualgoritmien sekä simulaatioiden varaan. DSGE-malleissa esimerkiksi rahan roolia talouden suhdannevaihtelussa mallinnetaan yksinkertaistamalla reaalimaailmasta, asettamalla rajoituksia tai sääntöjä sen käyttäytymiselle, lisäämällä lyhytaikaisia shokkeja ja lisäämällä muita elementtejä parantamaan mallin yhteensopivuutta. Käytännössä voitaisiin simuloida tilannetta, jossa kilpailu on säännöiltään monopolistista ja rahan kulutus perustuu ainoastaan käytettävissä olevan käteisen määrään. Tämän lisäksi markkinoihin kohdistuu keskuspankin

rahapolitiikan sekä finanssipolitiikan aiheuttamia lyhytaikaisia shokkeja, joihin markkinat reagoivat reaalimaailmaa vastaavalla viiveellä.

Usein DSGE-malleja estimoidaan pienellä noin 3–7 aikasarjan dataseteillä. Oletuksena on, että malliin liitetyt teoreettiset konseptit tuottavat sopivan mallin, eikä ylimääräinen data tekisi mallista paremmin sovittunutta. Lisäksi malleihin on asetettu 0–1 aikasarjaa per mallinnettu muuttuja (Boivin, 2006). Smets (2007) rakensi DSGE-mallin Yhdysvaltojen talouden estimointiin. Kyseiseen malliin on lisätty 7 rakenteellista shokkia vastaamaan 7 mallissa seurattavaa muuttujaa. Mallissa käytetään siis seitsemää neljännesvuosittain julkaistavaa aikasarjaa estimointiin: logaritminen todellinen BKT:n muutos, todellinen kulutus, reaali-investoinnit ja reaalipalkka, logaritminen työskennellyt tunnit, logaritminen BKT-deflaattori sekä inflaatio (federal funds rate). Mallin sisällä käytetään myös autoregressioita käsittelemään aikasarjoja. Lopullisen rakenteelliset algoritmit sisältävän mallin estimointiin käytetään bayesiläistä mallinnusta. Ensin lasketaan estimaatti posteriorijakaumasta. Tämän jälkeen jakaumaa simuloidaan Metropolis-Hastings-algoritmilla, eräänlaisella markovinketjualgoritmilla. Näin saadaan kuva jakaumasta ja sen uskottavuudesta.

Smetsin (2007) havaintojen perusteella DSGE-malli tuottaa tarkemman ennusteen verrattuna samalla datasetillä laskettuun VAR-malliin, tosin osa eroista saattaa johtua ylisovittuneisuudesta. Lisäksi DSGE-malli tuottaa yhtä tarkkoja tuloksia samalla datasetillä lasketun BVAR-mallin kanssa lyhyellä aikavälillä. Pitkän aikavälin ennusteissa 3 vuoteen asti DSGE-malli antoi jokaiselle lasketulle talouden indikaattorille paremmat ennusteet kuin verrokkimallit. DSGE-mallit ovat saaneet useilta tutkijoilta kritiikkiä siitä, että ne eivät väitetysti kykenisi ennustamaan taloudellisia kriisejä. Esimerkiksi kritiikin mukaan Smetsin (2007) malli ei pystyisi käsittelemään oikein ja ennustamaan vuoden 2008 finanssikriisiä (Del Negro, 2015). Del Negro (2015) käyttää kyseistä aiemmin tässä luvussa esitettyä mallia finanssikriisiä edeltävällä datasetillä. Tutkimus osoittaa, että malli ennustaa muut siihen liitetyt indikaattorit hyvin rajakustannuksia ja korkotasoa lukuun ottamatta. Kyseisten indikaattoreiden virhe kuitenkin johtui tutkimuksen mukaan rahapolitiikan aiheuttamasta vastakkaisreaktiosta mallissa.

3 MASSADATA MAKROTALOUSTIETEESSÄ JA SEN TEKNOLOGIAT

Varianin (2014) mukaan ekonometrian data-analyysissä on neljä vaihetta: ennustaminen, tiivistäminen, estimaatin tekeminen sekä hypoteesin testaus. Koneoppiminen keskittyy näistä ennustamiseen. Koneoppiminen sopii sen vuoksi juuri taloustieteen ennustamiseen hyvin, sillä se ottaa huomioon muuttujien taustalla olevia tekijöitä (Mullainathan, 2017). Koneoppimisen algoritmeilla, kuten päätöspuilla pystytään havainnollistamaan monimutkaisia suhteita paremmin, kuin perinteisesti käytetyillä malleilla (Varian, 2014). Internetissä on satoja miljoonia verkkosivustoja, ja niiden määrä nousee kymmenillä prosenteilla vuosittain (Edelman, 2012). Tämän takia verkkoharavointi on muodostumassa tärkeäksi massadatan keräämisen työkaluksi. Taloustieteilijät eivät ole yleisellä tasolla vielä saaneet lisättyä koneoppimista pääasialliseksi työkalukseen, ja sen täyden potentiaalin käyttöönotto on vielä kesken (Athey, 2018). Tämän vuoksi tämä luku keskittyy käsitteisiin, jotka ovat olennaisia käytetystä analysointimenetelmästä riippumatta sekä analysointimenetelmistä, joiden toimintaa on jo tieteellisesti tutkittu makrotaloustieteeseen liittyen.

3.1 Massadata

Massadatalla tarkoitetaan suuria datamassoja, joita voidaan hyödyntää erilaisilla tallennus- käsittely- ja esittämismenetelmillä. Massadatasta käytetty englanninkielinen termi Big Data on vakiintunut mutta kaupallisilta toimijoilta levinnyt. Tästä huolimatta se on käytännössä erittäin laajasti hyväksytty kuvaamaan merkitystään. Massadatan määritelmäksi on esimerkkinä kolme V:tä. Gandomi (2015) määrittelee seuraavasti:

- Volyme, eli määrä. Massadata tarkoittaa suuria, yleensä vähintään useiden teratavujen kokoisia datamassoja, joiden koko riippuu toimijasta ja sen toimialasta.
- Variety, eli monipuolisuus. Massadata koostuu useista erilaisista lukemiskelpoista ja -kelvotonta dataa esittävistä tallennusformaateista.
- Velocity, eli datan määrän kasvamisen nopeus. Massadataa syntyy jatkuvasti massiivisia määriä sekä tiedostetusti että sen lähteen tiedostamatta.
- Muita kolmen V:n inspiroimia määritelmiä, kuten: Veracity, eli osa lähteistä tuottaa dataa, josta on vaikeaa rakentaa toimivia malleja ja tehdä johtopäätöksiä; Variability, eli se, että dataa syntyy jatkuvasti muuttuvalla tahdilla, jossa on nousuja ja laskuja; Complexity, eli datan ja sen lähteiden monimuotoisuus; Value, eli massadata on pienissä määrin suhteellisesti arvotonta, mutta kun sitä analysoidaan suuria määriä, voidaan myös rakentaa suurta arvoa.

Massadata voidaan myös ainakin tämän tutkimuksen havaintojen perusteella jakaa sen saatavuuden mukaan: osa massadatasta on ostettavissa tai muuten hankittavissa kuten julkisille toimijoille lain oikeuttamana, esimerkiksi taloustieteilijöille hyödyllinen data pankkikorttiostoista. Osa datasta on täysin tai rajoitetusti jaettu sen kerääjän toimesta, esimerkkinä Google Insights. Massiivisen suuri määrä dataa kuitenkin löytyy vapaasti saatavilla internetistä. Internetin moninaista numero- ja tekstidataa on saatavilla verkkoharavoinnilla (eng. Web scraping). Data voidaan jakaa strukturoituun ja strukturoimattomaan dataan. Noin 5 % kaikesta datasta on strukturoitua. Strukturoimaton data sisältää strukturoimatonta tekstiä, videoita, kuvia sekä äänitteitä (Gandomi, 2015).

Taloustieteelle hyödylliseksi havaittu massadata on yleensä teksti- ja numeromuodossa. Numerodata voidaan Doornikin (2015) mukaan jakaa karkeasti kolmeen eri tyyppiin:

- Korkea; suuri määrä havaintoja suhteessa muuttujien määrään,
- Paksu; suuri määrä muuttujia suhteessa havaintojen määrään,
- Massiivinen; suuri määrä sekä havaintoja että muuttujia.

3.2 Massadatan käsittely

Gandomi (2015) luonnehtii prosessia, jossa massadatasta luodaan tietämystä viisivaiheiseksi prosessiksi. 1) Hankkiminen ja tallennus, 2) noutaminen, puhdistus ja metadatan luonti, 3) integraatio, aggregointi ja esittely, 4) Mallinnus ja analyysi sekä 5) tulkinta. Vaiheet 1–3 kuuluvat aliprosessiin datanhallinta, 4–5 kuuluvat aliprosessiin data-analyysi.

Kerätyssä massadatassa voi olla gigabittien verran strukturoimatonta dataa, joten tavanomaisella relaatiokantakokoonpanolla jo sen varastoiminen aiheuttaa ongelmia. Suurille datamäärille on kehitetty yksinkertaisemmilla toiminnallisuuksilla mutta nopeampaan datavirtaan valmiita NoSQL-tietokantoja. Google haravoi läpi jopa 20 miljoonaa internetosoitetta päivittäin. Se on kehittänyt erillisiä työkaluja, kuten Google file system sekä MapReduce, jotka soveltuvat näin massiivisen datamäärän käsittelyyn. MapReduce on ohjelmointityökalu, joka ajaa tehtäviä jakamalla ne useisiin samanaikaisesti suoritettaviin osiin eri prosessointiyksiköille. Tämänkaltainen järjestely saattaa jopa nopeuttaa konfiguraation osana olevien yksittäisten prosessointiyksiköiden kykyä käsitellä dataa. Näiden työkalujen pohjalta on myös tehty open-source ratkaisuja kuten Apache Hadoop. Työkalut toimivat pilvipalveluiden avulla; tilaa ja käsittelytehoa voi vuokrata skaalautuvasti, mikä säästää resursseja ja antaa pienemmille toimijoille mahdollisuuden analysoida massadataa (Gunther, 2015; Glushkova, 2019; Varian, 2014).

Työkalun datasta käsittelemä lopputulos on mahdollisesti jo suoraan luettavaa tietoa tai mahtuu järkevään tilaan lopullista analyysiä varten. Usein datan määrä on silti niin suuri, että siitä kannattaa ottaa vain osa tietoaineistoksi. Tällainen otos riittää kuvaamaan koko aineistoa tarpeeksi tarkasti tehokasta statistista analyysiä varten. Otoksesta on tässä vaiheessa järkevää tehdä testaavaa analyysiä, suodattaa dataa, joka ei ole olennaista esimerkiksi Open Refine -puhdistustyökalun avulla sekä testata datan loogisuutta ennen varsinaista käsittelyä ja mallintamista (Varian, 2014).

Taloustieteessä massadatan mallintamiseen käytetään ennakoivaa analyysiä. Ennakoiva analyysi pyrkii ennustamaan joko tulevaa tai jotakin jo mitattua muuttujaa, jonka arvon pystyy estimoimaan sen suhteista muihin muuttujiin. Ennakoivan analyysin tekniikoita ovat regressioalgoritmit sekä koneoppiminen (Gandomi, 2015).

3.3 Web Scraping

Verkkoharavointi on tiedon noutamisen kokonaisuuksien havaitsemiseen keskittyvä tekniikka. Verkkoharavoinnin algoritmit tunnistavat erilaisia tietoja ja niiden yhteyksiä, ja rakentavat näin metadataa esimerkiksi tietystä verkko-osoitteesta löytyneestä tekstistä ja numeroista. Verkkoharavointi on makrotaloustieteen kannalta olennaista, sillä se on yksi helpoimpia ja suhteellisesti halpoja menetelmiä kerätä dataa (Cavallo, 2016).

Verkkoharavointi vaati vielä muutama vuosi sitten ohjelmointitaitoja esimerkiksi PHP:ssä tai pythonissa, mutta nykyään on tarjolla käyttöliittymällä varustettuja ohjelmistoja pelkästään tätä varten. Niillä voidaan suoraan maalata tietyt asiat tietystä verkko-osoitteesta, jolloin verkkoharavoinnin algoritmi oppii hakemaan samankaltaista dataa muista verkko-osoitteista. Teknologian käyttöönotto ei vaadi käyttäjältä paljoa teknistä tietämystä. (Cavallo, 2016). Hara-

vointityökalu ei kuitenkaan pysty aina karsimaan halutun ulkopuolisia tietoja tehokkaasti. Tällöin voidaan käyttää ohjatun oppimisen algoritmeja puhdistamaan haravointityökalun keräämä data. Algoritmi opetetaan kategorisoimaan tietosisältöjä automaattisesti oppimisaineiston avulla (Breton ym., 2016).

Verkkoharavointi voidaan toteuttaa kertaluontoisesti tai esimerkiksi tietylle kohteelle tietyllä intervallilla. Makrotaloustieteen käyttötarkoituksiin kertaluontoisesta verkkoharavoinnista ei ole tähänastisissa tutkimuksissa löydetty hyötyä. Sopivan tietoaineiston luomiseksi vaaditaan pidempiaikaista datanseurantaa, jos halutaan luoda toimiva malli, sillä osa menetelmistä vaatii suuria tietoaineistoja, yksittäisen tekijän merkityksellisyys muuttuu ajan kanssa ja yleensä muutokset indikaattoreihin lasketaan muutoksista ja niiden vaikutuksista, kuten Cavallo (2016) esittää. Verkkoharavoinnin tuloksena syntyneestä datasta voidaan luoda tietoaineisto, jolla kyseisen tutkimuksenkohteen muodostaman parametrin vaikutusta jonkin indikaattorin tarkkuuteen pyritään arvioimaan. Jos aineistosta tehty tutkimus osoittaa, että tämä voi toimia selittävänä muuttujana indikaattorille, voidaan sitä ruveta seuraamaan halutuin välein (Edelman, 2012).

Verkkoharavoinnin suurin hyöty on se, että sen tuottamalla tiedolla voidaan korvata luotettavampaa mutta vanhentunutta tietoa kannattavaan hintaan, jolloin esimerkiksi bruttokansantuotteen muutoksen arvio perustuu tietoon, eikä asiantuntijan (harhaiseen) arvioon. Verkkoharavoinnin ongelmaksi voi muodostua se, että sivustojen ja verkkokauppojen ylläpitäjät kieltävät datankeruun sivustoltaan tai alkavat veloittamaan oikeuksista kerätä tietoa (Breton ym., 2016).

3.4 Koneoppiminen

Ekonometria on yksinkertaistettuna taloustieteen laskemista, joka on polveutunut tilastotieteestä. Ekonometrian perinteiset työkalut eivät aina pysty käsittelemään suuria monimutkaisia datamääriä järkevästi. Koneoppimisen tieteenala alkoi muodostua, kun tutkijat alkoivat siirtyä ajattelutapaan, että malli täytyy johtaa suoraan empiirisistä havainnoista (Bok, 2018). Massadata sisältää usein paljon epäoleellista tietoa, joka pitää karsia pois. Koneoppimisen menetelmät ovat parempia kuvaamaan useiden muuttujien välisiä monimutkaisia riippuvuussuhteita. Koneoppimisen menetelmiä ovat esimerkiksi hakupuut, tukivektorikoneet, neuroverkot ja syväoppiminen (Varian, 2014). Taloustieteessä käytetään koneoppimisen menetelmiä, jotka koostavat tai tiivistävät epälineaarisia suhteita käytettäväksi tiedoksi. Koneoppimisen algoritmien käyttö on Mullainathanin (2017) mukaan muuttunut teknisesti helpoksi. R ja Python tukevat valmiita paketteja kaikkiin tässä tutkielmassa esiteltyihin koneoppimisen menetelmiin. Oikean luokittelumenetelmän valitseminen ja sen tehokas hyödyntäminen vaativat kuitenkin kokemusta tai tietoa useista eri menetelmistä. Massa-

datan datanhallinnassa ja data-analyysissä tarkkoihin tuloksiin pääsee vain kokemuksen kautta.

Varianin (2014) mukaan tilastollisia ennusteita pyritään yleistettynä selvittämään jonkin selitettävän tai selitettävien muuttujien y jakaumana toisten selittävien muuttujien x funktiona. Koneoppimisessa x-muuttujia kutsutaan ennustaviksi muuttujiksi tai ominaisuuksiksi. Usein y:stä ja x:stä on jo havaittua dataa, mutta halutaan tehdä hyvä ennuste y:stä uusilla x:n arvoilla. Hyvällä ennusteella tarkoitetaan, että se minimoi jonkun tappiofunktion kuten jäännösneliösumman. Ekonometriassa ongelmaa lähdettäisiin ratkaisemaan jollakin lineaarisella menetelmällä. Epälineaariset funktiot voivat olla hyödyllisiä, varsinkin kun dataa on paljon tarjolla. Massadatan käsittelyyn sopivia menetelmiä ovat esimerkiksi:

- Päätöspuut, kuten luokittelu- ja regressiopuut (CART),
- Satunnaismetsät,
- Penalisoitu regressio, kuten LASSO ja Ridge regressio.

Koneoppimisen menetelmien ongelmana voi olla ylisovittuminen. Ylisovittuminen tarkoittaa sitä, että rakennetun mallin ennustava muuttuja on liian monimutkainen. Tällöin malli toimii sitä rakennettaessa käytetyn otoksen ulkopuolisen aineiston kanssa huonosti, vaikka se mallintaisi alkuperäistä aineistoa virheettömästi. Koneoppimiseen on tämän vuoksi kehitetty regularisaatioksi kutsuttuja menetelmiä, joilla malleja voidaan rankaista liiallisesta monimutkaisuudesta. On käytännöllistä käyttää eri tietoaineistoja opetukselle, validoinnille sekä testaukselle. Opetuksessa estimoidaan malleja, joista valitaan validoinnissa parhaiten soveltuva, jonka tarkkuutta testataan. Mallin monimutkaisuutta voidaan säätää k-kertaisella ristivalidoinnilla. Siinä käsittelemätön data jaetaan k osajoukkoon, joista otetaan yksi osajoukko s sivuun. Tämän jälkeen valitaan niin sanotulle säätöparametrille arvo, sovitetaan malli toisilla osajoukoilla, ja testataan, saadaanko osajoukolla s sama tulos kuin koko tietoaineistolla. Tämän jälkeen nostetaan säätöparametrin (Lambda) arvoa ja aloitetaan sovittaminen uudestaan niin, että valitaan toinen osajoukko s-osajoukoksi. Näin löydetään useita toimivia malleja, joista valitaan se, joka sisältää eniten olennaista informaatiota, eli on todennäköisesti tarkin (Varian, 2014). Yksi yleinen tapa suorittaa ristivalidointia on 10-kertainen ristivalidointi, jossa otetaan datasta kymmenen osajoukkoa. Osajoukkoja voi myös käyttää muissa suhteissa kuin suhteessa k-1.

3.4.1 CART ja satunnaismetsät

Luokittelupuu on päätöspuu, joka on yleensä binäärisesti jakautuva, eli puu jakautuu kahteen alipuuhun, jotka taas jakautuvat kahteen alipuuhun pysäytykseen saakka. Luokittelupuu luokittelee eri lopputulemien mukaan, kun taas regressiopuu luokittelee jatkuvien selittävien muuttujien mukaan. Muita käytet-

tyjä puita on esimerkiksi ehdolliset puut (conditional inference tree). Puita kasvatettaessa keskeisenä tavoitteena on päästä hyvään otoksen ulkopuoliseen tarkkuuteen. Päätöspuut toimivat hyvin, kun: 1) Aineistossa on havaittavissa selvää epälineaarisuutta ja useita tekijöitä. 2) Tietoaineiston koko on suuri. 3) Aineisto sisältää informaatiota, jota on vaikea havaita lineaarisilla menetelmillä (Varian, 2014).

Kuten muissakin koneoppimisen menetelmissä, päätöspuihin liittyy ylisovittumista. Puuta voi karsia asettamalla sille lopetuskriteereitä. Ristivalidoinnissa lehtisolmujen, eli puun alaosan päätesolmujen määrää käytetään monimutkaisuuden mittarina. 10 osajoukon ristivalidoinnissa 9 osajoukkoa käytetään opetuksessa rakennettaessa puuta tietyllä lehtisolmujen määrällä. Tämän jälkeen kymmenettä osajoukkoa käytetään testauksessa. Tavoitteena on löytää lehtisolmujen määrä, jolla saadaan paras luokittelutarkkuus testauksessa käytetyllä otoksella (Varian, 2014).

Luokittelutarkkuutta voidaan myös parantaa menetelmillä, jotka käyttävät useita puita yhden toimivan puun rakentamiseksi. Tehostamisessa (boosting) algoritmi rakentaa malleja yksitellen niin, että uusi malli painottaa edellisten mallien väärin luokiteltuja muuttujia. Tulokset määräytyvät joko enemmistövalinnalla (vote) tai keskiarvona (Varian, 2014). Bootstrap tarkoittaa sitä, että tietoaineistosta otetaan satunnainen osajoukko alkioita, jolle rakennetaan malli. Boostrap-aggretoinnissa (bagging) rakennetaan malleja ottamalla tietoaineistosta satunnaisotannalla halutun kokoinen osajoukko aineistoksi niin, että satunnaisotanta tapahtuu aina koko alkuperäisestä tietoaineistosta. Näin saadaan enemmistön tulokset valitsemalla luokittelu, jota on satunnaistettu (Breiman, 1996).

Breimanin (2001) mukaan satunnaismetsäksi kutsutaan edellä mainituilla tavoilla satunnaistettua prosessia, jonka tulos syntyy menettelytavalla, jossa kasvatetut puut äänestävät jokaisen yksittäisen alkion luokittelusta enemmistöperusteisesti. Satunnaismetsässä kuitenkin myös puiden jakamisprosessi satunnaistetaan käyttämällä jokaiseen luokittelemistapahtumaan osajoukon satunnaisesti valittua osajoukkoa. Satunnaismetsän puita ei kuitenkaan karsita. Varian (2014) esittelee seuraavanlaisen esimerkin satunnaismetsä-tyyppisestä algoritmista: Valitaan bootstrap-osajoukko. Valitaan satunnaiset ennustavat muuttujat jakamista varten. Nämä toistetaan useita kertoja. Viimeiseksi luokitellaan tietoaineisto enemmistövalinnan mukaan.

3.4.2 Regressioalgoritmit ja luokittelu

Luokittelussa regressioalgoritmit on todettu hyödyllisiksi. Kun x-muuttujia on suuri määrä, perinteinen lineaarinen regressio ei riitä estimoimaan tarkkaa mallia, varsinkaan jos kaikki muuttujat eivät sisällä uutta informaatiota mallia varten. Muuttujien valinnassa sopivaksi on havaittu niin sanottu elastinen verkko (Elastic Net), joka karsii mahdolliset epäolennaiset muuttujat mallista. Elastinen verkko minimoi aineistosta vastemuuttujan y suhteen selittävien muuttujien x

jäännösneliösumman, ja lisää siihen rankaisevan algoritmin. Algoritmi sisältää sekä LASSO-osan että ridge regressio -osan. Nämä algoritmit ovat osoittautuneet sekä suhteellisen kevyiksi laskea, että tarkoiksi ennusteen tekijöiksi (Varian, 2014).

Jos suurin osa selittävistä muuttujista sisältää selitettävien muuttujien suhteen olennaista tietoa, toimii ridge regressio parhaiten, jolloin Lasso-menetelmä karsiutuu pois. Jos suuri osa muuttujista on epäolennaisia, pystytään LASSO:n avulla poistamaan epäolennaisten muuttujien vaikutus malliin täysin, ts. karsimaan ne mallista (Varian, 2014). Muuttujien kertoimia β manipuloidaan pienemmiksi penalisoivilla termeillä λ_1 (Ridge regressio) sekä λ_2 (LASSO). Termien arvoa muuttamalla etsitään parhaiten y:n suhdetta x:ään kuvaama malli. Jos toinen tai molemmat λ-kertoimet saavat arvon 0, se tarkoittaa, että kyseinen funktio on eliminoitu algoritmista mallin tuottamiseksi. LASSO:n ongelma on, että jos x-muuttujien välillä on paljon korrelaatiota, se saattaa karsia myös olennaisia muuttujia. Jos muuttujien välillä on paljon korrelaatiota, elastinen verkko pystyy karsimaan epäolennaiset muuttujat pitäen kuitenkin suurimman osan olennaista tietoa sisältävistä muuttujista mallissa. LASSO:n ja Ridge regression erona on, että ridge regressiossa x-muuttujien kertoimista otetaan neliöjuuret mutta Lasso regressiossa niistä otetaan itseisarvot. Tämän vuoksi Ridge regressiossa kertoimet pystytään minimoimaan erittäin pieniksi, mutta ne eivät voi saavuttaa nollaa, toisin kuin LASSO:ssa. De Mol (2008) osoittaa, että sekä Ridge- että LASSO regression tuottamat tulokset korreloivat esimerkiksi manuaalista muuttujien valintaa vaativan pääkomponenttiregression kanssa. Banbura (2010) osoittaa, että tilanteessa, joka vaatii suurta muuttujamäärää, bayesiläiset menetelmät tuottavat tarkkoja ennusteita.

Spike-and-Slab regressio on Bayesiläinen regressioalgoritmi. Bayesiläinen tapa analysoida tietoa perustuu frekventistisistä menetelmistä käänteiseen oletukseen, että muuttujien arvot ovat vakioita. Tämä johtaa siihen, että esimerkiksi Spike-and-Slab-regressiossa valitaan Bernoulli-jakauma, jolla pyritään ennakoimaan sitä, kuinka todennäköisesti tietty muuttuja sopii rakennetun mallin osaksi (Varian, 2014). Arvo 1 tarkoittaa, että tietty muuttuja sopii malliin, arvo 0, ettei se sovi. Lasketaan monen mallin muuttujalle antama keskiarvo, joka on numero väliltä 0-1. Tämän jälkeen valitaan normaalijakauma, jolla ennakoidaan kyseisten selittävien muuttujien mahdollisia kertoimia. Näin saadaan kaksi arvoa jokaiselle muuttujalle: Sen todennäköisyys olla mukana mallissa (uskottavuus) sekä arvio jokaisen muuttujan kertoimen mahdollisista arvoista esitettynä keskiarvona normaalijakaumasta.

3.5 Nowcasting

Esimerkiksi Bruttokansantuotteen ennustaminen on aiemmin perustunut neljännesvuosittaisiin ennusteisiin, sillä siihen kerätty tieto on perustunut suurelta osin erilaisiin raportteihin, eikä ole ollut teknologioita, joilla voidaan ottaa uusi

informaatio huomioon automaattisesti. Makrotaloudellinen ennustaminen on aina ollut tasapainottelua tarkkuuden ja tuoreuden tai frekvenssin välillä (Bok, 2018). Ennusteiden tarkkuuteen on pyritty vaikuttamaan useita asiantuntijoiden manuaalisesti valitsemia malleja yhdistelemällä. Nykyaikainen nowcasting, eli lyhyen aikavälin talouden ennustaminen yhdistää eri aikoihin ja eri intervallilla julkaistavaa dataa reaaliaikaiseksi ennusteeksi. (Giannone, 2008; Bok, 2018). Termi tulee meteorologian lyhyen ajan ennustamisesta, ja sitä on viime vuosina alettu käyttää myös muilla tieteen ja talouden aloilla. Halutun talouden mittarin eri komponenttien dataa korvataan tai täydennetään yhdistämällä monesta lähteestä tulevaa dataa. Data voi olla erilaisissa muodoissa: kovaa informaatiota kuten tietyn alan yritysten liikevaihdon muutosta tai pehmeää informaatiota kuten kyselyt (Banbura, 2012).

Reaaliaikainen ennustaminen käyttää massadataa, sillä jo pelkästään siinä tietyllä ajanhetkellä käytettyjen muuttujien määrä lasketaan sadoissa. Viimeisen noin kahdenkymmenen vuoden aikana aikasarja-analyysin sovelluksia on kehitetty toimimaan näin korkeiden tietoaineistojen kanssa. Eri aikoihin saataville tulevan tiedon ongelmana on, että osa tarjolla olevasta tiedosta on jopa kuukausia vanhaa, osa samana päivänä kerättyä. Tätä kutsutaan jagged edgeksi, ja se ratkaistaan käyttämällä esimerkiksi Kalmanin filtteriä tai MIDAS-tyyppistä algoritmia (Giannone, 2008; Lahiri, 2013).

Reaaliaikaisen ennustamisen perusta on dynaaminen faktorimalli. Faktorimalli yhdistää kaikkien selittävien muuttujien sisältämästä informaatiosta muodostuvien faktorien, eli vaikuttavien tekijöiden, vaikutukset indikaattoriin. Kalmanin filtteri painottaa datanlähteitä niiden ajantasaisuuden ja laadukkuuden mukaan. Tämän lisäksi käytetään odotusarvon maksimoivaa algoritmia, eli EM-algoritmia. Näitä yhdistämällä muodostetaan suurimman uskottavuuden menetelmä. Pyritään siis etsimään malli, jolla saavutetaan suurin uskottavuus (Bok, 2018). Toinen lähestymistapa reaaliaikaiseen ennustamiseen on naiivi bayesilainen vektoriautoregressio. Se perustuu oletukseen, että jokainen mallin parametri on muista itsenäinen ja seuraa satunnaiskulkua (Bok, 2018).

Reaaliaikaiset ennusteet ovat tärkeitä, sillä niiden tarkkuus voi olla jopa 65 % parempi kuin pelkän kuukausittain julkaistavan tiedon sisältävä malli (Galbraith, 2018). Lisäksi ennusteet, joita pystytään nykyään luomaan automaattisesti ovat korreloituneita instituutioiden ja asiantuntijoiden laskemien ennusteiden kanssa (Bok, 2018). Ne sisältävät siis paljon samaa informaatiota perinteisesti käytettyjen menetelmien kanssa ja enemmän. Erityisesti finanssimarkkinoiden toimijat hyötyvät reaaliaikaisen ennustamisen kehityksestä, sillä ne tarvitsevat mahdollisimman tarkkaa tietoa päivittäin seuratessaan markkinatilannetta.

3.6 Massadatan käsittelyn ja reaaliaikaisen ennustamisen ongelmat

Massadatan käsittelyyn käytetyissä tilastollisissa menetelmissä tulee ottaa Doornikin (2015) mukaan huomioon se, että jos rakennettu malli ei sisällä kaikkea olennaista dataa, ottaa muu data sen paikan mallissa. Tämä tekee vääristymän malliin. Mallista tulee karsia osat, jotka eivät tuo relevanttia informaatiota esille datasta. Mallista tulee pystyä havainnoimaan, että se on luotettavasti toimiva. Valinta-algoritmin tulee pystyä samalla sekä käsittelemään suuren määrän muuttujia, että muodostamaan käyttötarkoituksessaan tehokkaan mallin. Näihin ongelmiin toimivaksi havaittuja ratkaisuja on esimerkiksi Autometrics, automaattinen mallinvalinta-algoritmi.

Paksulle tietoaineistolle Autometrics valitsee mallin hakupuu-teknologian avulla. Se rakentaa automaattisesti operaattoreita, epälineaarisia funktioita sekä indikaattoreita. Tämän jälkeen se valitsee oleelliset muuttujat ja testaa mallin sopivuuden. Näin rakennetaan useita malleja, joista valitaan parhaiten soveltuva. Korkean tietoaineiston tapauksessa muuttujat jaetaan osajoukkoihin, joista yhdestä valitaan relevantit muuttujat. Näihin lisätään aina seuraavan osajoukon relevanteiksi todetut muuttujat. Alkuperäisestä osajoukosta polveutuvaa joukkoa iteroidaan, kunnes lopulta jäljelle jää käsittelyä varten sopivan kokoinen joukko muuttujia. Osa muuttujien sisältämästä informaatiosta tulee esille vasta, kun uusia muuttujia lisätään malliin. Autometrics tutkii eniten toistensa suhteen korreloituneita muuttujia yhdessä. Jotta algoritmin yhteenlaskettu suoritusaika olisi realistinen, täytyy algoritmia kuitenkin optimisoida esimerkiksi turhiksi havaittuja indikaattoreita poistamalla ja käyttäen muuttujien poistamista joukoittain (Doornik, 2015).

4 MASSADATAN HYÖDYNTÄMINEN KÄYTÄNNÖSSÄ

Tässä luvussa on eritelty eri käyttökohteita massadatalle ja koneoppimisen menetelmille. On havaittavissa, että ensimmäisessä sisältöluvussa eriteltyjä massadatan eri muotoja on tutkittu monipuolisesti. Koneoppimisen menetelmistä satunnaismetsille on löydetty käyttöä. Verkkoharavointi osoittautui erittäin hyödylliseksi työkaluksi hintaindeksien laskemisessa. Reaaliaikaisen ennustamisen tutkimustuloksia ei ole erikseen määritelty tässä luvussa, sillä pankit usein perustavat mallinsa pienellä variaatiolla Giannonen (2008) rakentamaan malliin.

4.1 Pankkikorttidata

Galbraithin (2018) mukaan elektronisten maksujen dataa voidaan käyttää kulutuksen arvioimiseen, kun lasketaan bruttokansantuotetta. Pelkkä pankkikorttidata sisältää suurimman osan hyödyllisestä informaatiosta, ja se on helpommin saatavilla kuin luottokorttidata. Luottokorttidataa oli saatavissa vain osalle observoidusta ajasta. Luottokorttiyhtiöiden data on muutenkin vaikeammin saatavilla tai kallista hankkia. Pankki -ja luottokorteista periaatteessa jokainen yksittäinen osto on eriteltävissä. Käteismaksuista ei ole saatavissa tietoa, ja nostojen datasta on vaikeaa saada luotettavaa informaatiota todellisesta maksujen määrästä. Siispä Galbraith (2018) on jättänyt ne otoksen ulkopuolelle.

Maksujen data koostetaan kuukausittaistasolle ja yhdistetään komposiittiindeksin sekä työttömyystilastojen kanssa. Muutoksia voidaan seurata myös päivittäisellä tasolla, esimerkiksi jos halutaan ennakoida makrotalouden kriisejä Galbraith (2018).

4.2 Verkkokauppadata

Verkkokauppojen hintatietoja voi käyttää datana, kun lasketaan hintaindeksejä. Cavallo (2016) hyödyntää verkkoharavointimenetelmiä. Data jalostetaan ja siitä lasketaan yksinkertaisia indikaattoreita. Edelmanin (2012) mukaan myös kulutusta voidaan mahdollisesti mitata käyttämällä verkkokauppadataa. Osa verkkokaupoista nimittäin näyttää tuotteiden varastojen arvot sivustolla vierailevile. Kohtalaisella nopeudella tapahtuvaa varaston pienentymistä voidaan pitää kulutuksena, ja suurta varaston täyttymistä suoraan varastojen täyttämisenä. Verkkokauppojen itse julkaisemista myyntiranking-tilastoista voidaan myös mallintaa kulutusta, jos ranking-sijoituksista onnistutaan johtamaan tuotteiden myyntimääriä.

Kivijalkakauppojen hintakehitystä laskiessa veisi liikaa resursseja seurata jokaisen tuotteen ja mallin hintaa. Siispä palkatut skannaajat keräävät tiedot vain tuotekategorian suosituimmista malleista. Yleensä hintoja päivitetään kuukausittain, ja kun esimerkiksi elektroniikkatuotteiden mallit vanhentuvat aletaan seuraamaan uudempaa vastaavaa tuotetta. Hintarobotilla voidaan havaintojen intervalli vähentää esimerkiksi yhdeksi päiväksi. Tämän lisäksi voidaan seurata verrattain pienillä resursseilla kategorian jokaista tuotetta. Indeksiä ei tarvitse tuotetta vaihtaessa säätää todelliselle tasolle (Cavallo, 2016).

Vaikka verkkokauppadataa voidaan seurata päivittäistasolla ja sen hankkimisen kustannukset ovat pienet, siinä on joitakin rajoitteita. Ostovoimapariteetin laskemiseen tarvittavaa dataa saadaan maakohtaisesti ainoastaan suurimmilta myyjiltä. Verkkokaupoissa näytetyt hinnat saattavat erota myymälän hinnoista. Myös maansisäiset hintaerot jäävät huomioimatta, kun verkkomyynnissä hinnat ovat samat maanlaajuisesti. Saatu data sisältää tietoa ainoastaan noin 13–23 % populaatiosta (Cavallo, 2018). Hinta saattaa siis vääristyä vastaamaan kalliimman tason alueiden mukaiseksi.

Skanneridataa keräävät vähittäismyyjät eivät yleensä ole valmiita luovuttamaan tietoja analysoitavaksi indeksiä varten, minkä vuoksi edes toimivan tiedonkäsittelymenetelmän rakentaminen on vaikeaa (Breton ym., 2016). Verkkokauppadata sen sijaan on yleensä luettavissa vapaasti. Lisäksi on mahdollista pyytää kaupoilta muutakin dataa, kuin sen, mikä on julkisena vapaasti noudettavissa. Verkkosivustot ja niiden käyttäjät ovat paljon oletettua valmiimpia jakamaan niitä koskevaa tietoa.

4.3 Hakukonedata

Mclaren (2011) tutki internetin hakukoneista saatavaa tietoa ja sen avulla rakennettavia taloudellisia indikaattoreita Isossa-Britanniassa. Google on selvästi johtava hakukone länsimaissa, ja se tarjoaa tietoa trendaavista hakusanoista ja niiden hakumääristä. Yksittäisten manuaalisesti valittujen hakusanojen käyttä-

minen aineistona on todettu luotettavaksi menetelmäksi, kunnes löydetään keino verrata ja yhdistää suurien hakusanamäärien ja niiden kombinaatioiden sisältämää informaatiota toisiinsa. Mclarenin (2011) mukaan hakukonedata sisältää hyödyllistä informaatiota, josta osa on sellaista, jota muut datan lähteet eivät sisällä. Työttömyyttä mitatessa hakusana ”JSA”, joka viittaa työttömyystukihakemukseen (Job-seeker’s Allowance) oli hakumääriltään tarpeeksi suuri ja tarkkuudeltaan sopiva, ja sillä oli positiivista kovarianssia työttömyyslukujen kanssa, joten se valittiin indikaattoriksi. Koska virallisia työttömyystilastoja julkaistaan viiveellä, pystyy Googlen datasta rakennetulla indikaattorilla parantamaan reaaliaikaisen ennustamisen tarkkuutta. Kuitenkin itse työttömyystukihakemusten määrä oli tutkimuksen mukaan Googlen hakukonedataa marginaalisesti tarkempi indikaattori yksittäin tarkasteltuna.

Toinen Mclarenin (2011) tutkima indikaattori oli asumiskiinteistöjen hinnat. Sopivaksi hakusanaksi valikoitui kiinteistönvälittäjät (estate agents). Hakusana valikoitui sen vuoksi, että verrokkihakusanojen hakumäärä pomppi päiväkohtaisesti, joten niitä ei voinut pitää luotettavana pohjana indikaattorille. Tutkimuksen tuloksena oli, että hakukonedata oli yksittäisenä indikaattorina kaikkia verrokki-indikaattoreita tarkempi yhden kuukauden ennustetta laskiessa. Lisäksi muiden indikaattoreiden data julkaistaan viiveellä. Voidaan vetää johtopäätös, että hakukonedata hyvällä valitulla hakusanalla sisältää uutta informaatiota, joka ei ole perinteisemmillä lähteillä ollut käytettävissä. Lisäksi hakukonedatassa on potentiaalia reaaliaikaisen ennustamisen parametrinä.

4.4 Muita havaintoja

Muitakin kuin teksti- ja numeropohjaisia sovellutuksia koneoppimiselle on spekuloitu. Satelliittidatasta voi muodostaa tietoa. Pelloista otetuista kuvista pystytään arvioimaan maatalouden vuosituotantoa. Kaupunkien valovoiman suhdetta tuottavuuteen voidaan tutkia. Näiden tehtävien ainut tällä hetkellä järkeenkäypä ratkaisu on hyödyntää koneoppimista (Mullainathan, 2017). Baker (2015) kehitti uuden indeksin, EPU:n, joka mittaa luottamusta talouspolitiikkaan. Indeksi mittaa tiettyjen uutisaiheiden esiintyvyyttä uutismediassa. Indeksi lasketaan vektoriautoregressio-mallilla.

Perinteisempiä tiedonlähteitä, jotka voivat hyödyntää koneoppimista on esimerkiksi ISM:n (Institute for Supply Management) yrityskartoitukset, kuten PMI (Purchase Managers’ Index), jotka ovat sisältäneet tärkeää informaatiota BKT:n ja suhdannevaihtelun ennustamisessa. Lahirin tutkimuksen (2013) mukaan tämänkaltaiset kartoitukset ovat relevantteja myös koneoppimista hyödyntävien nowcasting -menetelmien syötedatan osana. Behrens (2018) hyödynsi satunnaismetsiä arvioidessaan saksalaisten makrotalouden indikaattoreiden ennustamisen tehokkuutta. Tutkimuksessa tuli ilmi, että varsinkin lyhyen aikavälin indikaattorien tehokkuudessa ilmeni ongelmia.

5 YHTEENVETO

Makrotaloustieteen ennustamisen tarkimpia menetelmiä ovat 2010-luvulla olleet vektoriautoregressiot, bayes vektoriautoregressiot sekä DSGE-mallit. Varsinkin bayes VAR sopii kirjallisuuskatsauksen lähteiden mukaan näistä suurten datasettien käsittelyyn. Massadata on kooltaan suurta ja monimuotoista dataa, joka on käyttökelpoista joko suoraan tai yleensä käsittelyprosessin jälkeen. Koneoppimisen menetelmiä on kehitetty luokittelemaan ja ennustamaan näitä suuria datamääriä. Massadataa on ruvettu käyttämään makrotaloustieteessä varsinkin hintatasoon liittyvien indikaattoreiden ennustamisessa, mutta prosessi hyödyntää koneoppimisen menetelmiä täysin on makrotaloustieteessä edelleen vaiheessa. Tutkimustulokset osoittavat, että esimerkiksi verkkoharavoinnilla hankittu data ja hakukonedata sisältävät informaatiota, jota perinteisistä tietolähteistä ei löydy.

Vertaisarvioitua tieteellistä tutkimusta on joskus vaikea löytää massadatan aihealueelta, sillä monesti sen asiantuntijat keräävät tietämyksensä julkaistaviksi kirjoiksi (Gandomi, 2015). Nowcasting-menetelmistä suuri osa tarjolla olevasta tiedosta taas on keskuspankkien raporteissa. Tarkkuuden lisäksi on otettava huomioon myös tiedon tuottamisen kustannukset. Jos ei ole varaa tarkkaan analyysin, on halvempia vaihtoehtoja. Tulevaisuudessa reaaliaikaista finanssimarkkinatietoa sekä reaaliaikaista makrotaloudellista tietoa voidaan seurata rinnakkain (Bok, 2018). Automaattiset rahoitusmarkkinoiden algoritmit voivat saada jatkuvan datavirran muokkaamasta automaattiselta reaaliaikaisen ennustamisen algoritmilta tietoa jatkuvalla syötöllä.

Tässä katsauksessa lueteltujen tiedonlähteiden sekä perinteisillä tavoilla tuotettu tieto sisältävät jokainen omalta osaltaan merkittävää informaatiota. Makrotaloustiede on Bokin (2018) mukaan ottanut johtavan aseman massadatan tutkimuksessa ja koneoppimisen kehityksessä. Voidaan päätellä, että erilaisten tiedonlähteiden koostamismenetelmien kehityksestä on tulevaisuudessa hyötyä myös muiden alojen tietojenkäsittelyssä. Koneoppimisessa voitaisiin myös ruveta käyttämään joitakin menetelmiä, joita on kehitetty ja opittu hyödyntämään ekonometriassa.

Tutkimus koneoppimisen käytöstä taloustieteessä on yliopistojen lisäksi keskuspankkien sekä tulevaisuudessa pitkälti yksityisten ekonomistien perustamien yritysten hartioilla. Yliopistojen rahoitus ei riitä kattavan tiedon ostamiseksi esimerkiksi luotto- ja pankkikorttiyrityksiltä tai pitämään yllä tarvittavaa laskentatehoa monivektorimallien päivittäislaskentaa varten niin, että toiminta olisi kannattavaa. Se, että massadatan analysointia taloustieteen tarkoituksiin voi harjoittaa yksityisyrityksenä (Billion prices project), antaa jo järkevän syyn siirtää datan prosessointia markkinoiden harteille. Ekonomistit, joilla on tietoteknistä ja koneoppimisen osaamista voivat keksiä tapoja kaupallistaa tietämystään. Tämä voi johtaa yhdessä itsenäisesti toimivien mallien ja teknologian kokonaisvaltaisen kehityksen kanssa siihen, että ekonomistit pystyvät ostamaan monipuolisia tietoaineistoportfolioita yhä tarkempien ennusteiden tekemiseksi.

Ekonomisteja palkataan teknologiayrityksiin yhä enemmissä määrin. Jos taloustieteen opiskelijat lisäisivät tietotekniikan osuutta tutkintoihinsa, voisi taitoja käyttää myös muiden sekä talousmaailman sisäisten että sen ulkoisten haasteiden ratkomiseen sekä yhteistyöhön tietojenkäsittelyn ammattilaisten kanssa (Athey, 2018; Varian, 2014). Taloustieteen ratkaisuja voisi käyttää esimerkiksi reaaliaikaisessa päätöksenteossa sekä liiketoimintatiedustelussa.

Mielenkiintoisiksi jatkotutkimusaiheiksi ja kysymyksiksi ehdottaisin kirjallisuuskatsauksen ja näiden yhteenvetoon koottujen tietojen perusteella seuraavia:

- Kuinka koneoppiminen ja massadata muuttavat asiantuntijatyön luonnetta nyt ja tulevaisuudessa?
- Kuinka digitaalinen johtaminen voi ottaa huomioon koneoppimisen menetelmistä johtuvaa yritystransformaatiota?
- Voiko reaaliaikaisen ennustamisen dataa yhdistää pankkien reaaliaikaisen päätöksenteon ja liiketoimintatiedustelun tueksi?

LÄHTEET

Agrawal, A. (2019). Artificial intelligence: The ambiguous labor market impact of automating prediction †. Journal of Economic Perspectives, 33(2), 31-50. doi:10.1257/jep.33.2.31

Athey, S. (2018). Economists (and economics) in tech companies. NBER Working Paper Series, , 25064. doi:10.3386/w25064

Baker, S. (2015). Measuring economic policy uncertainty. NBER Working Paper Series, , 21633. doi:10.3386/w21633

Banbura, M. (2012). Now-casting and the real-time data flow. IDEAS Working Paper Series from RePEc, Haettu osoitteesta https://jyu.finna.fi/PrimoRecord/pci.proquest1698259666

Bańbura, M. (2010). Large bayesian vector auto regressions. Journal of Applied Econometrics, 25(1), 71-92. doi:10.1002/jae.1137

Behrens, C. (2018). A test of the joint efficiency of macroeconomic forecasts using multivariate random forests. Journal of Forecasting, 37(5), 560-572. doi:10.1002/for.2520

Boivin, J. (2006). DSGE models in a data-rich environment. NBER Working Paper Series, , 12772. doi:10.3386/w12772

Bok, B. (2018). Macroeconomic nowcasting and forecasting with big data. doi:10.1146/annurev-economics-080217-053214

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140. doi:10.1007/BF00058655

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. doi:1010933404324

Breton, R., Flower, T., Mayhew, M., Metcalfe, E., Milliken, N., Payne, C., . . . Woods, A. (2016). Research indices using web scraped data: May 2016 update. Newport: Office for National Statistics.Available from Www.Ons.Gov.Uk/File,

Cavallo, A. (2016). The billion prices project: Using online prices for measurement and research †. Journal of Economic Perspectives, 30(2), 151-178. doi:10.1257/jep.30.2.151

Cavallo, A. (2018). Using online prices for measuring real consumption across countries. NBER Working Paper Series, , 24292. doi:10.3386/w24292

De Mol, C. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? Journal of Econometrics, 146(2), 318-328. doi:10.1016/j.jeconom.2008.08.011

Del Negro, M. (2015). Inflation in the great recession and new keynesian models †. American Economic Journal: Macroeconomics, 7(1), 168-196. doi:10.1257/mac.20140097

Diebold, F. X. (1998). The past, present, and future of macroeconomic forecasting. Journal of Economic Perspectives, 12(2), 175-192. doi:10.1257/jep.12.2.175

Doornik, J. A. (2015). Statistical model selection with "Big data". Cogent Economics & Finance, 3(1) doi:10.1080/23322039.2015.1045216

Edelman, B. (2012). Using internet data for economic research. Journal of Economic Perspectives, 26(2), 189-206. Haettu osoitteesta https://jyu.finna.fi/PrimoRecord/pci.aea10.1257%2Fjep.26.2.189

Fernández - Villaverde, J. (2010). The econometrics of DSGE models. SERIEs, doi:10.1007/s13209-009-0014-7

Galbraith, J. W. (2018). Nowcasting with payments system data. International Journal of Forecasting, 34(2), 366-376. doi:10.1016/j.ijforecast.2016.10.002

Gandomi, A. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144. doi:10.1016/j.ijinfomgt.2014.10.007

Giannone, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. Journal of Monetary Economics, 55(4), 665-676. doi:10.1016/j.jmoneco.2008.05.010

Glushkova, D. (2019). Mapreduce performance model for hadoop 2.x. Information Systems, 79, 32-43. doi:10.1016/j.is.2017.11.006

Gunther, N. (2015). Hadoop superlinear scalability. Communications of the ACM, 58(4), 46-55. doi:10.1145/2719919

Lahiri, K. (2013). Nowcasting US GDP: The role of ISM business surveys. International Journal of Forecasting, 29(4), 644-658. doi:10.1016/j.ijforecast.2012.02.010

Litterman, R. B. (1986). Forecasting with bayesian vector autoregressions-five years of experience. Journal of Business & Economic Statistics, 4(1), 25-38. doi:10.1080/07350015.1986.10509491

Mclaren, N. (2011). Using internet search data as economic indicators. Bank of England.Quarterly Bulletin, 51(2), 134-140. Haettu osoitteesta https://jyu.finna.fi/PrimoRecord/pci.proquest877039452

Mullainathan, S. (2017). Machine learning: An applied econometric approach. The Journal of Economic Perspectives, 31(2), 87-106. doi:10.1257/jep.31.2.87

Smets, F. (2007). Shocks and frictions in US business cycles: A bayesian DSGE approach. American Economic Review, 97(3), 586-606. doi:10.1257/aer.97.3.586

Stock, J. H. (2001). Vector autoregressions. Journal of Economic Perspectives, 15(4), 101-115. doi:10.1257/jep.15.4.101

Varian, H. R. (2014). Big data: New tricks for econometrics †. Journal of Economic Perspectives, 28(2), 3-28. doi:10.1257/jep.28.2.3