INFLAATION KASVUVAUHDIN ENNUSTE-MENETELMIEN VERTAILU

Taloustiede
pro gradu -tutkielma

Laatija:
Janne Flinck

Ohjaaja:
Heikki Kauppi

25.04.2017
Turku

Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -järjestelmällä.

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

Sisällysluettelo

Kuvaluettelo

Taulukkoluettelo

1 JOHDANTO

Inflaatiokehityksen ennustaminen on keskeinen tekijä pitkän aikavälin taloudellisissa päätöksissä, ei pelkästään keskuspankkien hintavakauden ja rahapolitiikan säätelyssä. Talouden kasvuoletukset saattavat muuttua oleellisesti ja siksi on pyrittävä arvioimaan niihin liittyvät epävarmuustekijät ja kehityssuunnat. Taloudessa ennustaminen perustuu yleensä aikasarjoihin, jotka koostuvat peräkkäisistä aikaan sidotuista havaintoaineistoista. Esimerkiksi nykyarvostamisen ongelmat, kuten myös suuri osa arvopapereiden hinnoittelusta, sisältävät ennusteongelman, joka liittyy tulevaisuuden rahavirtojen ennakoimiseen. (Elliott ja Timmermann 2008, 3 – 5)

Ennustetilanteet eroavat paljolti niiden aikahorisontista, valituista ennustemuuttujien vaikutuksista, havaintoaineiston määrästä ja sen sisältämistä riippuvuuksista. Tarkan ennusteen tekeminen edellyttää havaintoaineiston esikäsittelyä ja tämän pohjalta tehtävää tilastollista mallinnusta. Mallinnuksessa joudutaan tekemään joukko oletuksia esimerkiksi liittyen mallin funktionaaliseen muotoon ja siihen, mitkä taloudelliset ennustemuuttujat ovat oleellisia tämän ennusteen kannalta. Näiden oletusten ulkopuolelle jää helposti eksogeeniset tekijät kuten poliittiset, rakenteelliset ja teknologiset muutokset. Lisäksi mallinnus saattaa kärsiä liian vähästä aineiston määrästä ja sen heikosta laadusta. (Elliott ja Timmermann 2008, 3 – 5)

Philipsin-käyrä on perinteinen työkalu inflaation ennustamiseen. Tämä työkalu ei kuitenkaan näytä toimivan yhtä hyvin kuin ennen ja se näyttää tuovan hyvin vähän lisäarvoa inflaation ennustamiseen. (Atkeson ja Ohanian, 2001) Makrotaloudelliset muuttujat ovat hyvin usein toisistaan riippuvaisia ja sisältävät informaatiota toisistaan. (Jin-Lung ja Tsay 2005, 1) Usean ennustemuuttujan hyödyntäminen mahdollistaa monipuolisemman ja laajemman havaintoaineistojen käytön. Toisaalta oleellisten ennustemuuttujien valinta on vaikeaa ja usean ennustemuuttujan käyttö johtaa moniparametriseen ja monimutkaiseen malliin. Usean ennustemuuttujan käyttö johtaa ns. suuren dimensionaalisuuden ongelmaan. (Stock ja Watson 2004, 517) Tätä ongelmaa Stock ja Watson (2004) lähestyvät pääkomponenttiregression avulla muodostaen faktoreita ennustaakseen Yhdysvaltojen teollisuustuotannon indeksin kasvuvauhtia 130 ennustemuuttujalla. Tämän tutkielman tarkoituksena on verrata eri ennustemallien toimivuutta Yhdysvaltojen inflaatiolle käyttäen kuukausitason dataa vuodesta 1969 vuoteen 2015 asti. Tutkielman tavoitteena on seurata Stock ja Watson (1999) asetelmaa, jossa he tutkivat inflaation kasvuvauhdin ennustamista 168 makrotaloudellisella ennustemuuttujalla soveltamalla erilaisia metodeja.

Ennustemenetelmien toimivuutta voidaan arvioida vertaamalla ennustettujen havaintojen arvoja niitä vastaaviin toteutuneisiin arvoihin. Tämä palaute voi johtaa muutoksiin ennustemetodissa ja mitä nopeammin käsitys mallin toimivuudesta ja tarkkuudesta saadaan, sitä parempia taloudellisia päätöksiä voidaan tehdä. (Elliott ja Timmermann 2008, 32 – 33)

Taloustieteellisten ennustemallien muodostamisessa joudutaan tekemään valinta mallin monimutkaisuuden ja ennustetarkkuuden välillä. Monimutkaiset epälineaariset regressiomallit, joissa käytetään monia ennustemuuttujia, ovat yleensä tuottaneet huonompia ennusteita kuin yksinkertaiset mallit, jotka käyttävät vain pientä osaa käytössä olevasta informaatiosta. Menetelmiä, joissa käytetään useita ennustemuuttujia, mutta joissa yritetään kiertää korkean dimensionaalisuuden tuoma ongelma, on tutkittu viimeaikaisessa kirjallisuudessa. (Exterkate ym. 2012)

Myös aikasarjojen epälineaarisen riippuvuuden mahdollisuus on saanut huomiota. Neuroverkkoja on sovellettu tämänkaltaisen epälineaarisen suhteen mallintamiseen. Yleisesti ottaen nämä lähestymistavat ovat sopivia vain pienille määrille ennustemuuttujia eivätkä nämä mallit ole kuitenkaan tuoneet merkittävää parannusta ennusteiden tarkkuuksiin. (Stock 2001)

Tässä tutkielmassa sovellan lineaarisia ja epälineaarisia metodeja, jotka soveltuvat korkean dimension ongelmaan. Vertailen erilaisia menetelmiä käyttämällä mittarina ennusteen keskineliövirheen neliöjuurta. Kaikki ennustevertailut tapahtuvat käyttäen simuloitua otoksen ulkopuolista menetelmää. Mallien empiirinen vertailu keskittyy kahteen pääkysymykseen: mikä malli tuottaa parhaimman tuloksen ja ovatko tulokset samankaltaisia vai eroavatko niiden tarkkuudet merkittävästi toisistaan.

Tämän tutkielman rakenne on seuraava. Toisessa luvussa käyn läpi ennustamisen ongelmia. Kolmannessa luvussa esittelen ennustemuuttujien valintametodeja, kriteereitä ja ennustevertailut. Neljännessä luvussa kuvaan metodit, joita käytän empiirisessä analyysissä. Viidennessä luvussa esittelen empiirisen analyysin ja malleilla saadut tulokset arvioituna keskineliövirheen neliöjuurella. Kuudes luku päättää tutkielman tulosten pohdintaan ja ennustamisongelmien läpikäyntiin.

2 ENNUSTEONGELMA

2.1 Ennusteongelman määrittely

Tämän työn kannalta ennustaminen määritellään prosessiksi, jossa sovelletaan tilastollista mallia tai koneoppimisalgoritmia uuden havainnon ennakointiin. Keskityn tapaukseen, jossa tavoitteena on ennustaa aikasarjan tulevaa arvoa tietylle horisontille h. Vastemuuttujasta käytän merkintää y_{t+h} ja sitä vastaavasta ennustearvosta $\hat{y}_{t+h}$ silloin kun aikasarjasta on poistettu niille tyypillinen epästationaarinen rakenne, sillä käytetyt metodit olettavat, että aikasarjat ovat stationaarisia. Tämä ennuste tehdään, kun ennustemuuttujien F_t ja ennustettavan aikasarjan havainnon y_{t+h} aikaisemmat arvot $1, \dots, T$:hen asti ovat annettuja, missä T on otoskoko.

Käytössä on datajoukko F_t, jota käytetään muodostamaan ennuste $\hat{y}_{t+h}$. F_t on kaikki käytössä oleva informaatio ajalla t, mukaan lukien y_t:n viiveet. Haluamme muodostaa ennusteen $\hat{y}_{t+h}$, joka minimoi ennusteen keskineliövirhettä $E[(y_{t+h} - \hat{y}_{t+h})^2 | F_t]$. Tämä keskineliövirhe minimoituu, kun ennuste on ehdollinen odotusarvo $E(y_{t+h}|F_t)$. (Stock 2001, 564)

Käytännössä $E(y_{t+h}|F_t)$ on tuntematon ja yleensä epälineaarinen. Ennusteet muodostetaan estimoimalla tätä tuntematonta ehdollista odotusarvoa ennustefunktiolla, joka voi olla parametriton tai parametrillinen. Parhaat parametrit ovat ne, jotka minimoivat keskineliövirheen neliöjuuren. (Stock 2001, 564) Keskeisiä osia ennusteiden laadinnassa ja niiden evaluoinnissa on ennustemuuttujien valinta, ennustemallin funktionaalisen muodon valinta ja tapa, jolla painotetaan vanhoja ja uudempia havaintoja. (Elliott ja Timmermann 2008, 8, 17)

2.2 Ennustemuuttujien valinta

Ennustemuuttujien (engl. variable selection tai feature selection) valinnassa on kysymys mallin kannalta merkittävien muuttujien valinnasta vaihtoehtoisten ennustemuuttujien joukosta. On mahdollista, että kaikki ennustemuuttujat liittyvät vastemuuttujaan eikä niistä mitään kannata poistaa. On kuitenkin todennäköisempää, että vastemuuttujaan vaikuttaa vain pieni osajoukko kaikista käytössä olevista ennustemuuttujista. Hyvin usein on niin, että ennustemuuttujien joukossa on turhia muuttujia ja ne voidaan poistaa ilman oleellista informaation menetystä. (Gareth ym. 2015, 78) Esimerkiksi, jos jotkin kaksi ennustemuuttujaa korreloivat keskenään vahvasti, niin toinen niistä on turha, sillä ne sisältävät samaa informaatiota.

Ennustemuuttujien valintamenetelmät ovat usein yhdistelmämenetelmiä hakumenetelmistä, jotka etsivät merkitsevää muuttuja-alijoukkoa yhdistettynä valintakriteeriin, joka pisteyttää vaihtoehtoiset muuttuja-alijoukot keskenään. Yksinkertaisin menetelmä on testata kaikki mahdolliset muuttujakombinaatiot ja valita sellainen muuttujajoukko malliin, joka tuottaa pienimmän keskineliövirheen havaintoaineistossa. Tämä on hyvin raskas lähestymistapa eikä sovellu kuin pienille muuttujamäärille ja havaintoaineistoille. (Gareth ym. 2015, 78; Guyon ja Elisseeff 2003)

Muuttujien valintamenetelmät voidaan luokitella kolmeen seuraavan luokkaan: paketointimenetelmät (eng. wrapping), suodatusmenetelmät (engl. filtering) ja sulautetutmenetelmät (engl, embedded). Paketointimenetelmissä valittujen muuttujien avulla muodostettujen mallien sopivuutta testataan validointiotosjoukossa, jota ei ole käytetty mallin muodostamiseen tai muuttujien valintaan. Suodatusmenetelmät käyttävät erilaisia kriteerejä löytääkseen ne muuttujat, joilla on suurin merkitys ennusteeseen. Esimerkki tällaisesta kriteeristä on Akaiken-informaatiokriteeri tai Bayesian-informaatiokriteeri. (Guyon ja Elisseeff 2003; Gareth ym. 2015, 78 – 79) Sulautetuissa menetelmissä muuttujien valinta on sisällytetty mallin muodostamiseen. Esimerkki tällaisesta menetelmästä on Breimanin (2001) esittelemä satunnaismetsä ja Tibshiranin (1996) esittelemä “Least absolute shrinkage and selection operator”.

Seuraavassa osassa kerron tarkemmin metodien taustalla olevasta teoriasta ja siitä, millä tavalla eri metodit potentiaalisesti parantavat ennustetarkkuutta.

3 OTOSMENETELMÄT JA VALINTAKRITEERIT

Kuten luvussa 2 kuvattiin, optimaalinen ennuste minimoi keskineliövirheen neliöjuuren. Keskineliövirhettä käytetään eri metodien keskinäisessä vertailussa sekä parhaimman mallin valinnassa. Keskineliövirheen neliöjuuren käyttö on suosittu vertailukriteeri makroekonometriassa ja jatkan tässä työssä tätä perinnettä.

Ennustemallin tiukka sovitus käytettävissä olevaan havaintoaineistoon ei takaa hyvää ennustettavuutta. Havaintoihin tiukasti sidottu malli yleistyy heikosti tulevaisuuden toteutumiin, sillä se ei pysty huomioimaan samanaikaisesti generoivan prosessin satunnaiskohinaa sekä havaintoaineiston säännöllisiä rakenteita erityisesti silloin, jos havaintoaineistoa on niukasti. Sellaisen mallin, joka seuraa tarkasti havaintoaineistoa, mutta ei pysty huomioimaan satunnaisuutta tulevissa toteutumissa sanotaan olevan ylisovittautunut (engl. overfitting) mallin muodostuksessa käytettyyn havaintoaineistoon. Tällaisen mallin vastakohtana on alisovittautunut malli, joka ei mallinna havaintoaineiston säännönmukaisuuksia riittävällä tarkkuudella. Tekemällä lisäoletuksia mallin muodostamisessa esim. käyttämällä korkeamman asteen polynomeja ja oletuksia havainnoista parantaa mallin sovittautumista havaintoaineistoon, mutta samalla malli monimutkaistuu. Malliin sanotaan syntyvän harhaa johtuen tehdyistä oletuksista. Toisaalta malli, johon on sisällytetty vähemmän oletuksia ei pysty samaan tarkkuuteen samalla havaintoaineistolla. Tämä ongelma tunnetaan mallinnuksessa harhavarianssidekompositiona (engl. bias-variance decomposition), joka johtuu siitä, että mallin valinnassa yritetään samanaikaisesti minimoida kahta eri virhelähdettä: mallia liikaa yksinkertaistavia oletuksia ja varianssin liiallista sitomista havaintoaineistossa olevaan satunnaiseen kohinaan. Ongelma on keskeinen erityisesti tilastollisilla oppimismenetelmillä muodostetuissa malleissa ja niiden muodostamisessa käytetyissä otoksissa. (Hastie ym. 2009, 33 – 34; Burnham ja Anderson 2002, 219 – 228).

Inoue ja Kilian (2006, 274) tutkivat artikkelissaan havaintoaineiston määrän ja siitä seuraavan harhavarianssidekomposition vaikutusta simuloidun otoksen ulkopuolisiin ennustevertailuihin ja informaatiokriteeriin. Suurilla datamäärillä informaatiokriteerin osoitetaan toimivan vähintäänkin yhtä hyvin, ellei paremmin kuin simuloidun otoksen ulkopuolisiin ennustevertailuihin perustuva mallin valinta. Mallinnuksessa minimoitava keskineliövirhe voidaan esittää harhan ja varianssin avulla:

$$E(y_{t+h} - \hat{y}_{t+h})^2 = [E(\hat{y}_{t+h}) - y_{t+h}]^2 + E[\hat{y}_{t+h} - E(\hat{y}_{t+h})]^2 + E[\varepsilon_{t+h}^2]. \qquad (1)$$

Tässä kaavassa y_{t+h} on toteutuma ajanhetkellä $t + h$ ja $\hat{y}_{t+h}$ on sitä vastaava ennuste käyttämällä informaatiota F_t. Ensimmäinen termi, eli $[E(\hat{y}_{t+h}) - y_{t+h}]^2$ on harhan (engl. bias) neliö, joka edustaa ennusteen odotusarvon, $E(\hat{y}_{t+h})$, ja toteutuneen havain-

non, y_{t+h}, välistä erotusta. Toinen termi, eli $E[\hat{y}_{t+h} - E(\hat{y}_{t+h})]^2$ on estimoidun ennusteen varianssi ja $E[\varepsilon_{t+h}^2]$ on y_{t+h} varianssi. Kaavasta nähdään, että monimutkainen, runsaasti oletuksia sisältävä ennustefunktio saadaan lähestymään todellista mallia otoksen sisällä, mutta varianssi kasvaa. (Fortmann-Roe 2012; Burnham ja Anderson 2002, 31 – 35, Hastie ym. 2009, 24, 37-38)

Tämä kaava tunnetaan harhavarianssidekompositiona, joka voidaan osoittaa lähtemällä liikkeelle keskineliövirheestä ryhmittelemällä sen termit seuraavalla tavalla (Geman 1992, 9):

$$E\left[\frac{1}{T-h} \times \sum_{t=1}^{T-h} (y_{t+h} - \hat{y}_{t+h})^2\right] = \frac{1}{T-h} \times \sum_{t=1}^{T-h} E[(y_{t+h} - \hat{y}_{t+h})^2], \qquad (2)$$

missä T on otoskoko ja h on valittu ennustehorisontti. Havainto y_t voidaan kirjoittaa todellisen generoivan funktion f avulla muotoon $y_{t+h} = f(F_t) + \varepsilon_{t+h}$. Tässä $E[\varepsilon_{t+h}] = 0$ ja F_t on kaikki käytössä oleva data ajanhetkellä t. Estimoimme todellista funktiota $f(F_t)$ käyttäen metodia $\hat{f}(F_t)$, joka tuottaa ennusteen $\hat{y}_{t+h}$. Olkoon $F_t = x_0$ jokin valittu otosarvo tietyllä ajanhetkellä (Hastie ym. 2009, 223). Tarkastellaan odotusarvoa summan sisällä seuraavasti (Geman 1992, 4, 10; Rudin 2012, 2-4):

$$E[(y_{t+h} - \hat{y}_{t+h})^2 | F_t = x_0] = E[(y_{t+h} + f(x_0) - f(x_0) - \hat{y}_{t+h})^2]$$

$$= E\left[\left(y_{t+h} - f(x_0)\right)^2 + (f(x_0) - \hat{y}_{t+h})^2 + 2\left((f(x_0) - \hat{y}_{t+h})\left(y_{t+h} - f(x_0)\right)\right)\right]$$

$$= E[\varepsilon_{t+h}^2] + E[(f(x_0) - \hat{y}_{t+h})^2] + 2(E[f(x_0)y_{t+h}] + E[f(x_0)^2] - E[\hat{y}_{t+h}y_{t+h}] + E[\hat{y}_{t+h}f(x_0)]).$$

Tässä yhtälössä

1. $E[f(x_0)y_{t+h}] = f(x_0)^2$, koska $y_{t+h} = f(x_0) + \varepsilon_{t+h}$.
2. $E[\hat{y}_{t+h}y_{t+h}] = E[\hat{y}_{t+h}(f(x_0) + \varepsilon_{t+h})] = E[\hat{y}_{t+h}f(x_0) + \hat{y}_{t+h}\varepsilon_{t+h}] = E[\hat{y}_{t+h}f(x_0)]$, koska $E[\varepsilon_{t+h}] = 0$

Eli nyt saadaan:

$$E[(y_{t+h} - \hat{y}_{t+h})^2 | F_t = x_0] = E[\varepsilon_{t+h}^2] + E[(f(x_0) - \hat{y}_{t+h})^2].$$

Tällä tavalla keskineliövirhe voidaan kirjoittaa harhan ja varianssin avulla seuraavasti lisäämällä ja vähentämällä ennustetun arvon odotusarvo, $E[\hat{y}_{t+h}]$:

$$E[\varepsilon_{t+h}^2] + E[(f(x_0) - \hat{y}_{t+h})^2] = E[\varepsilon_{t+h}^2] + E[(f(x_0) + E[\hat{y}_{t+h}] - E[\hat{y}_{t+h}] - \hat{y}_{t+h})^2]$$

$$= E[\varepsilon_{t+h}^2] + E[(f(x_0) - E[\hat{y}_{t+h}])^2 + E[(E[\hat{y}_{t+h}] - \hat{y}_{t+h})^2] + 2E[(E[\hat{y}_{t+h}] - \hat{y}_{t+h})(f(x_0) - E[\hat{y}_{t+h}])]]$$

$$= E[\varepsilon_{t+h}^2] + harha^2 + Var(\hat{y}_{t+h}) + 2\big(E\big[f(x_0)E[\hat{y}_{t+h}]\big] - E[E[\hat{y}_{t+h}]^2] - E[\hat{y}_{t+h}f(x_0)] + E\big[\hat{y}_{t+h}E[\hat{y}_{t+h}]\big]\big).$$

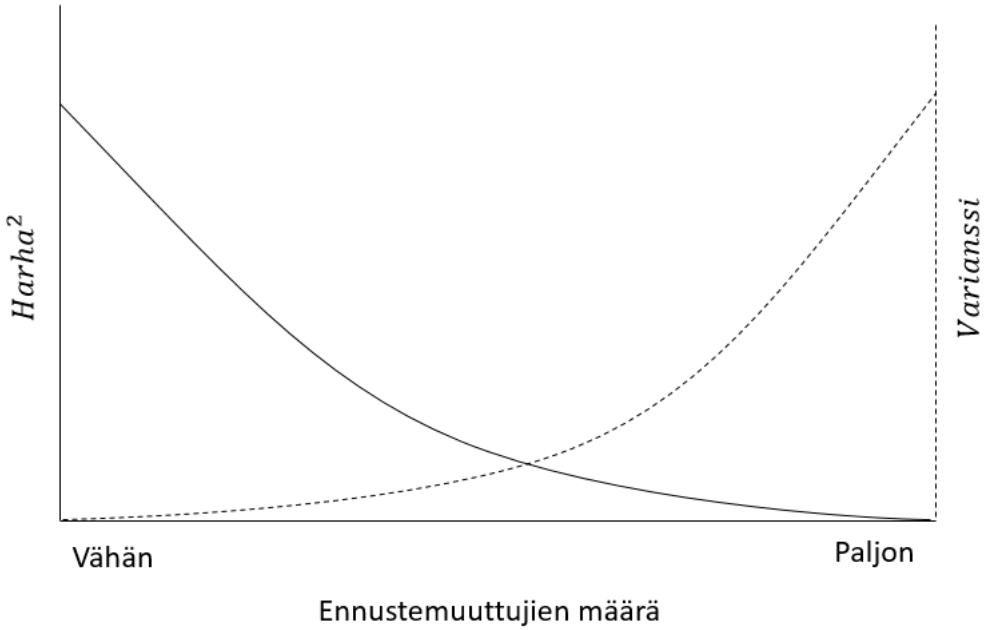
Tässä yhtälössä

1. $E\big[f(x_0)E[\hat{y}_{t+h}]\big] = f(x_0)E[\hat{y}_{t+h}]$,
2. $E[E[\hat{y}_{t+h}]^2] = E[\hat{y}_{t+h}]^2$,
3. $E[\hat{y}_{t+h}f(x_0)] = f(x_0)E[\hat{y}_{t+h}]$,
4. $E\big[\hat{y}_{t+h}E[\hat{y}_{t+h}]\big] = E[\hat{y}_{t+h}]^2$.

Tästä seuraa se, että viimeinen termi $2\big(E\big[f(x_0)E[\hat{y}_{t+h}]\big] - E[E[\hat{y}_{t+h}]^2] - E[\hat{y}_{t+h}f(x_0)] + E\big[\hat{y}_{t+h}E[\hat{y}_{t+h}]\big]\big)$ menee nollaksi, eli saamme:

$$E[\varepsilon_{t+h}^2] + harha^2 + Var(\hat{y}_{t+h}) = E[\varepsilon_{t+h}^2] + [E(\hat{y}_{t+h}) - y_{t+h}]^2 + E[\hat{y}_{t+h} - E(\hat{y}_{t+h})]^2.$$

Kuva 1 havainnollistaa harhavarianssidekomposition vaikutusta mallin valintaan. Harha pienenee ja varianssi kasvaa, kun malliin lisätään parametreja. Parhaan ennustemallin ei kuitenkaan tarvitse olla juuri harha- ja varianssikäyrien leikkauspisteessä (Burnham ja Anderson 2002, 31).



Kuva 1 Harhavarianssidekomposition vaikutus muuttujien määrään

Seuraavaksi paneudun mallinnuksessa käytettävien aineistojen muodostamiseen ja mallin valinnan perusteisiin.

3.1 Ristiinvalidointi

Viime aikoihin asti talousennusteita on arvioitu ilman parametrien estimointivirheitä ja mallin määrittelyvirheitä. On hyvin tunnettua, että ennustavilla mallinnusmenetelmillä, jotka hakevat sopivaa mallia useista eri vaihtoehdoista, on taipumus ylisovittaa malli havaintomateriaaliin, jota on käytetty mallin muodostamiseen. Tämä johtaa optimistisiin arvioihin ennustemallin suorituskyvystä silloin, kun käytetään samaa havaintomateriaalia sekä parametrien estimointiin että evaluointiin. Tyypillinen käytäntö ennustavassa mallinnuksessa onkin, että osaa havaintomateriaalista ei käytetä mallien ja niiden parametrien estimointiin vaan ne säästetään mallin evaluointivaiheeseen, jossa käytetään havaintojoukon ulkopuolisia näytteitä. (Elliott ja Timmermann 2008, 25)

Ristiinvalidoinnissa on kaksi tavoitetta: löytää malli, jolla on pienin testausvirhe ja hienosäätää mallin parametrien arvot. Ristiinvalidoinnin avulla saadaan realistisempi kuva mallin opetusvirheestä ja siten myös tarkempi käsitys ennustemallin tarkkuudesta. (Hastie ym. 2009, 241 – 247)

Tilastollisiin oppimismenetelmiin (esim. harjanneregressio ja pääkomponenttiregressio) perustuvassa ennustemallinnuksessa havaintomateriaali jaetaan opetusdataan ja testausdataan, jota kutsutaan myös validointidataksi. Opetusdata on se osa havaintomateriaalista, jota käytetään mallien ja niiden parametrien muodostamiseen. Testausdata on se osa havaintomateriaalia, joka jätetään pois opetusdatasta ja jota käytetään opetusdatalla

opetettujen mallien valintaan ja niiden suorituskyvyn arviointiin. Ennustemallinnuksessa syntyy periaatteessa kahdentyyppistä virhettä: opetusvirhettä ja testausvirhettä. Keskimääräinen opetusvirhe (tai opetusvirhetaajuus) voidaan laskea helposti mallien tuottamien ennusteiden keskineliövirheillä opetusdatassa. Tyypillisesti näin saatu virhearvo on merkittävästi pienempi kuin testausdatalla saatava testausvirhearvo. Testausvirhetaajuus (engl. test error rate) on keskimääräinen virheiden määrä, jonka tilastollinen oppimismenetelmä tuottaa testausdatalle, jota ei ole käytetty mallin opettamisvaiheessa. (Gareth ym. 2015, 176)

Ristiinvalidoinnissa on kaksi tavoitetta: löytää malli, jolla on pienin testausvirhe ja hienosäätää mallin parametrien arvot. Ristiinvalidointia käytän keskineliövirheen neliöjuuren minimin löytämiseksi mallinnuskappaleessa tiettyjen mallien kohdalta. (Gareth ym. 2015, 175 – 183; Hastie ym. 2009, 241 – 247)

Yksinkertaisimmillaan opetusdatan ja testausdatan muodostaminen ja käyttö on hyvin suoraviivaista. Jaetaan käytettävissä olevat havainnot kahteen osaan – opetusdataan ja testausdataan. Jako voi olla tasajako tai satunnaisesta kohdasta tehty kahtiajako. Vaihtoehtoiset mallit parametreineen sovitetaan opetusdataan ja näin saatuja malleja käytetään ennustamaan testausdatan havaintoja. Tällä yksinkertaisella menetelmällä on kuitenkin ongelmia. Testausvirhe vaihtelee paljon sen mukaan, miten jako on tehty. Vain osaa kaikista havainnoista hyödynnetään mallien ja niiden parametrien estimointiin, jolloin mallista ei tule niin tarkka kuin se voisi olla ja näin testausvirhearvio on todellista suurempi. (Gareth ym. 2015, 176 – 178)

Koko aineisto voidaan jakaa kahden osan sijaan useaan osaan ja osia käytetään opetukseen ja testaukseen. ”Jätetään yksi ulos” (engl. leave one out) -menetelmässä datajoukko jaetaan d:hen osaan, joista ensimmäistä $1/d$-osaa käytetään testaukseen ja loppuja osia ($d - 1$ kpl) mallin opetukseen. Tätä voidaan toistaa siten, että otetaan seuraava osa ja käytetään sitä testaukseen ja jäljellä olevia osia opetukseen jne. (Gareth ym. 2015, 178 – 181; Hastie ym. 2009, 241 – 247)

3.2 Simuloidut otoksen ulkopuoliset ennustevertailut

Ennustemetodien tuottamia ennusteita samasta aikasarjasta voidaan arvioida käyttämällä keskineliövirhettä. Otoksen ulkopuolista suoritusta voidaan mitata käyttämällä joko todellisia otoksen ulkopuolisia ennusteita tai simuloituja otoksen ulkopuolisia ennusteita. Molemmilla metodeilla on samat tavoitteet, eli tarkastella metodin suorituskykyä. (Stock 2001, 566)

Kandidaattimallien arviointi otoksen ulkopuolisilla ennustevertailuilla (engl. simulated out-of-sample) perustuu siihen, että otosaikasarjan aikaisempia arvoja käytetään mal-

lin rakentamiseen ja ennuste tehdään otosaikasarjan seuraaviin arvoihin. Nämä myöhäisemmät arvot, joita ei siis ole käytetty mallin rakentamiseen, ovat simuloituja otoksen ulkopuolisia arvoja. (Inoue ja Kilian 2006, 274 – 275)

Yleinen tapa vertailla kandidaattimalleja on implementoida simuloidun otoksen ulkopuolinen menetelmä käyttäen liukuvaa regressiota. Liukuvan regression tapauksessa kandidaattimallit sovitetaan havaintoihin jotka ovat käytettävissä periodilla $1, \dots, T$. Nämä havainnot jaetaan ikkunoihin $S - W + 1, \dots, S$, missä S on S ensimmäistä havaintoa ja W on liukuvan ikkunan pituus. Ennusteita verrataan havaintoihin $S + h$, missä $S = W, W + 1, W + 2, \dots, T - h, h$ on horisontti ja T on koko aikajakso. (Inoue ja Kilian 2006, 274 – 275) Simuloidun otoksen ulkopuolinen menetelmä käyttäen liukuvaa regressiota on verrattavissa ”Jätetään yksi ulos” -menetelmään.

Ennustemalleja voi kuitenkin myös valita vertailemalla eri mallien antamia keskineliövirheiden neliöjuuria käyttäen otosjoukon sisäisiä (engl. in sample) arvoja. Tässä on vaarana mallin ylisovittaminen opetusdataan. Tämä voidaan estää asettamalla mallille jokin kriteerifunktio, joka rankaisee ylisovittamisesta. Tällaista informaatiokriteerimenetelmää on käytetty valitsemaan parhaita ennustemuuttujia osakkeiden tuotoille. Mallien valinta informaatiokriteerimenetelmää käyttäen tarkoittaa sitä, että kaikki kandidaattimallit estimoidaan otosperiodeilla $1, \dots, T - h$ ja tämän jälkeen valitaan ennustemalli, joka minimoi etukäteen valitun kriteerifunktion. (Inoue ja Kilian 2006, 274 – 275)

Simuloidun otoksen ulkopuolisella menetelmällä on taipumus olla epäjohdonmukainen parhaimman mallin suhteen, sillä tämä menetelmä päätyy valitsemaan yliparametrisoidun mallin suurella todennäköisyydellä kuten Inoue ja Kilian (2006) ovat osoittaneet. Tämä taipumus suosia yliparametrisoituneita malleja johtaa suurempaan keskineliövirheen neliöjuureen todellisessa ennustustilanteessa kuin jos mallin valinnassa käytettäisiin informaatiokriteeriä, joka valitsee “todellisen” mallin tai sitä parhaiten asymptoottisesti lähenevän mallin kandidaattimallien joukosta, mikäli informaatiokriteerin rangaistusparametri on valittu hyvin. Tästä seuraa se, että informaatiokriteeri valitsee pienemmän virheen mallin äärellisellä havaintomäärällä simuloidun otoksen ulkopuolisissa ennusteissa. Informaatiokriteerimenetelmä valitsee siis tarkemman mallin arvioituna keskineliövirheen neliöjuurella kuin simuloidun otoksen ulkopuolinen menetelmä sellaisissa tilanteissa, joissa molemmat menetelmät ovat käytettävissä. (Inoue ja Kilian 2006, 275)

Tämän kaltaiset ennusteongelmat ovat hyvin tyypillisiä makroekonometriassa ja rahoituksessa. Esimerkiksi ennustettaessa valuuttakursseja tai inflaatiota malleilla, joilla on paljon makrotaloudellisia ennustemuuttujia, törmätään juuri tähän kysymykseen, eli mitä havaintojoukkoa tulee käyttää mallin valintakriteerin yhteydessä. (Inoue ja Kilian 2006, 274)

3.3 Informaatiokriteeri

Muuttujat valitaan yleensä automaattisella prosessilla, missä vertaillaan eri ennustemuuttujien vaikutusta tietyllä testiotannalla (engl. test statistic). Näin voidaan epäsuorasti estimoida mallin tuottamaa virhettä ja tehdä tarpeelliset korjaukset opetusdatan virheelle huomioonottamalla mahdollisen ylisovituksen tuoma vääristymä. Näitä testiotantoja käytetään esimerkiksi kahden yleisesti käytetyn informaatiokriteerin yhteydessä, eli Akaike-informaatiokriteeri ja Bayesian-informaatiokriteeri (Stock 2001, 566; Gareth ym. 2015, 78 – 79).

Mallin ylisovitusta tapahtuu, kun mallissa on liian monta parametria ja mallilla on huono ennustuskyky ja se reagoi opetusdatassa oleviin merkityksettömiin muutoksiin. Tällainen malli muistaa opetusdatansa, mutta ei pysty sovittautumaan uuteen dataan. Molemmat edellä mainituista informaatiokriteereistä yrittävät välttää liiallisista parametreista johtuvan ylisovituksen käyttämällä rangaistustermiä, joka rankaisee mallin kompleksisuudesta, eli mallin muuttujien määrästä. (Burnham ja Anderson 2004)

Akaike-informaatiokriteeri suosii vähemmän kompleksista mallia mallin sovitustarkkuuden kustannuksella. Tällä informaatiokriteerillä on kiinteä rangaistuskerroin kompleksisuudelle. Akaike-informaatiokriteeri on tilastollisten mallien relatiivinen vertailumittari ja näin ollen toimii myös mallinvalintatilastona. Akaike-informaatiokriteeri arvioi mallin laatua vertaamalla sitä toisiin malleihin. Se antaa siis arvion informaation menetyksestä, kun jotakin mallia käytetään edustamaan generoivaa prosessia, joka tuotti datan. (Verbeek 2004, 58) Aion käyttää informaatiokriteereitä viiveiden määrän valintaan autoregressiivisessä mallissa. Näistä viiveistä käytän merkintää p. Esittelen autoregressiivisen mallin tarkemmin kappaleessa 4.1.

Akaike-informaatiokriteeri voidaan kirjoittaa seuraavaan muotoon:

$$AIC = log\hat{\sigma}^2 + \frac{2(p+1)}{W}, \tag{3}$$

missä $\hat{\sigma}^2$ on estimoidun satunnaisprosessin virhetermin varianssi, $p + 1$ on vapaiden parametrien määrä ja W on liukuva ennusteikkuna (Verbeek 2004, 285). Paras malli on se, jolla on pienin Akaike-informaatiokriteerin arvo. Tämä informaatiokriteeri palkitsee mallin sopivuudesta (engl. goodness of fit), mutta rankaisee mallin muuttujien lukumäärästä (Gareth ym. 2015, 212; Verbeek 2004, 58).

Käytännössä aloitamme malleista, joissa on eri määrä ennustemuuttujia ja etsimme mallin, joka minimoi Akaike-informaatiokriteerin arvon. Täydellistä mallia ei ole ja informaatiota menetetään lähes aina, joten tavoitteena on valita malli, joka minimoi informaation menetyksen.

Bayesian-informaatiokriteeri (BIC) suosii mallin yhteensopivuutta sen kompleksisuuden kustannuksella. Malli, jolla on pienin Bayesian-informaatiokriteerin arvo, on suosituin malli eri vaihtoehtojen joukosta. Tämä informaatiokriteeri rankaisee myös mallin kompleksisuudesta. (Verbeek 2004, 58) Bayesian-informaatiokriteeri voidaan esittää seuraavalla kaavalla:

$$BIC = log\hat{\sigma}^2 + \frac{(p+1)}{W} logW, \tag{4}$$

missä $\hat{\sigma}^2$ on estimoidun satunnaisprosessin virhetermin varianssi, $p + 1$ on vapaiden parametrien määrä ja W on liukuvan ennusteikkunan koko (Verbeek 2004, 285). Pienempi informaatiokriteerin arvo viittaa joko malliin, jossa on vähemmän muuttujia tai parempi sopivuus tai molempiin. (Gareth ym. 2015, 212; Verbeek 2004, 58) Näiden informaatiokriteerien kaavoista näemme, että Akaike- ja Bayesian-informaatiokriteereillä on sama viitekehys mallin sopivuudelle.

4 KÄYTETTYJEN MENETELMIEN ESITTELY

Metodit, joita käytän tässä tutkielmassa, voidaan jakaa kahteen kategoriaan: parametrisiin ja parametrittomiin metodeihin. Parametriset metodit pitävät sisällään oletuksia todellisen funktion f muodosta ja ne yleensä sovitetaan pienimmän neliösumman menetelmällä. Parametriset mallit vaativat luonnollisesti estimaatit myös mallin parametreille tai kertoimille. Tällaisten mallien mahdollinen haittapuoli on se, että malli ei yleensä vastaa todellista mallia f. Tämä taas johtaa siihen, että ennusteiden tarkkuus on heikko. Parametrittomat mallit eivät tee minkäänlaisia oletuksia funktion f muodosta, minkä tähden parametrittomilla malleilla on yleensä selkeä etu parametrisiin malleihin nähden. Ne vain pyrkivät estimoimaan funktion f mahdollisimman tarkasti annetuilla havainnoilla. Parametrittomat mallit vaativat suuren määrän havaintoaineistoa, jotta saadaan tarkka estimaatti $\hat{y}_{t+h}$. (Gareth ym. 2015, 17 – 24) Käytän merkintää F_t käytössä olevasta informaatiosta ajalla t, mukaan lukien y_t:n viiveet ja merkintää $F_{i,t}$, joka on i:s ennustemuuttuja F_t:n joukosta, missä $i = 1, \dots, n$ ja n on ennustemuuttujien lukumäärä.

Olettaen, että suhde vastemuuttujan y_{t+h} ja ennustemuuttujien F_t on suunnilleen lineaarinen, pienimmän neliösumman sovituksella on pieni harha. Jos $T > n$, eli havaintojen määrä on suurempi kuin ennustemuuttujien määrä, niin pienimmän neliösumman sovituksella on usein myös pieni varianssi ja malli toimii hyvin otoksen ulkopuolisilla havainnoilla. Jos T, eli havaintojen määrä, on vain vähän suurempi kuin n, eli ennustemuuttujien määrä, niin tämä johtaa ylisovittamiseen ja huonoon ennustetarkkuuteen ulkopuolisissa havainnoissa. Jos $n > T$, niin pienimmän neliösumman sovitusta ei voida enää tehdä, sillä varianssi on ääretön. Tämän kaltaisia ongelmia voi ratkaista askeltavilla menetelmillä, dimension pienentämisen menetelmillä, tai regularisointimenetelmillä. Nämä menetelmät usein johtavat suureen parannukseen ennustettaessa otoksen ulkopuolisia havaintoja. Usein myös kaikki käytössä olevat muuttujat eivät sisällä informaatiota vastemuuttujasta. Jos tällaisia muuttujia otetaan mukaan regressioon, mallista tulee monimutkaisempi ja turhien muuttujien lisääminen lisää mallin sitomista havaintoaineistossa olevaan satunnaiseen kohinaan. (Gareth ym. 2015, 203 – 204)

4.1 Autoregressiivinen prosessi

Ekonometristen aikasarjamallien vertailussa käytetään tyypillisesti yksinkertaista lineaarista mallia vertailukohteena muille malleille. Autoregressiivinen malli (AR) on tällainen yleisesti käytetty malli, jota käyttävät mm. Stock ja Watson (1999), Stock (2001), Biau ja D'Elia (2010) ja Exterkate ym. (2012) tutkimuksissaan. Autoregressiivistä metodia voidaan ajatella eräänlaisena lineaarisena regressiona, jossa aikasarjan edelliset havainnot ovat ennustemuuttujina. AR(p)-malli voidaan määritellä lausekkeella:

$$y_{t+h} = \varphi_0 + \sum_{i=1}^{p} \varphi_i y_t + \varepsilon_{t+h}, \quad (5)$$

missä φ_i, ..., φ_p ovat vakioita, p on autoregressiivisen prosessin aste, joka valitaan käyttäen aikaisemmin määriteltyä Bayesian-informaatiokriteeriä tai Akaike-informaatiokriteeriä ja ε_{t+h} on satunnaisprosessin virhetermi. Autoregressiivisen prosessin aste määrittelee käytettyjen viiveiden määrän. Tämä viiveiden määrä valitaan aina uudelleen jokaisessa ikkunassa W minimoimalla valittu informaatiokriteeri.

AR(p)-malli on stationaarinen, kun sen karakteristisen yhtälön:

$$z^p - \varphi_1 z^{p-1} - \cdots - \varphi_p = 0, \quad (6)$$

juurten itseisarvot ovat pienempiä kuin yksi, eli ne ovat yksikköympyrän sisäpuolella. Viivepolynomia $L^p y_t = y_{t-p}$ käyttäen stationaarisuus ehto voidaan kirjoittaa muotoon:

$$1 - \varphi_1 L - \cdots - \varphi_p L^p = 0, \quad (7)$$

jonka juuret tulee olla yksikköympyrän ulkopuolella. (Shumway ja Stoffer 2011, 84 – 86; Cowpertwait ja Metcalfe 2009, 79 – 80) Tämän tyyppistä mallia on yleisesti käytetty aikasarja-analyysissä ja käytän sitä myös tässä tutkielmassa vertailukohtana muille tutkituille malleille.

4.2 Harjanneregressio

Ennustaminen on haasteellista, kun käytössä on suuri määrä korreloituneita muuttujia. Näissä tapauksissa estimoidut kertoimet ovat yleensä epästabiileja ja muuttuvat suuresti, kun lisätään uusia havaintoja tai muuttujia. Tästä seuraa se, että otoksen ulkopuoliset ennusteet ovat epätarkkoja, vaikka malli olisi teoreettisesti harhaton. Yksi vaihtoehto on lisätä rajoite perinteiseen pienimmän neliösumman regressioon. (Exterkate ym. 2012, 3 – 4) Jos suurta määrää muuttujia käytetään rajoituksettomassa regressiossa, niin tämä johtaa myös ylisovittamiseen ja näin huonoihin ennustetuloksiin (Stock ja Watson 1999, 314).

Harjanneregressio (RR) (engl. ridge regression) on regularisoitu versio pienimmän neliösumman menetelmästä, jossa regularisoinnilla suositaan pienempiä kertoimia. Harjanneregressio pienentää kertoimia asettamalla rangaistuksen tai rajoituksen kertoimien kokoon ja pienentämällä niitä kohti nollaa. Tämän metodin ideana on korjata multikollineaarisuuden tuomia ongelmia. (Jin-Lung ja Tsay 2005, 6)

Lineaarinen yhtälö voidaan kirjoittaa seuraavassa muodossa:

$$y_{t+h} = \beta_0 + \sum_{i=1}^{n} F_{i,t}\beta_i + \varepsilon_{t+h}, \tag{8}$$

missä β_0 on vakiotermi, $\beta_1, \dots, \beta_i$ ovat $F_{i,t}$ kertoimia, jotka estimoidaan datasta, n on ennustemuuttujien lukumäärä ja ε_{t+h} on virhe termi. (Shumway ja Stoffer 2011, 48 – 50)

Harjanneregression kertoimet minimoivat regularisoidun version pienimmän neliösumman menetelmästä ja se voidaan esittää seuraavassa muodossa, missä tavoitteena on estimoida β_0 ja $\beta_1, \dots, \beta_i$ minimoimalla yhtälö:

$$\sum_{t=1}^{W} (y_{t+h} - (\beta_0 + \sum_{i=1}^{n} F_{i,t}\beta_i))^2, \tag{9}$$

missä W on liukuvan ikkunan pituus. Tällä yhtälöllä on rajoite:

$$\lambda \sum_{j=1}^{n} \beta_j^2 \leq s, \tag{10}$$

Yhtälö saa siis muodon:

$$\sum_{t=1}^{W} (y_{t+h} - (\beta_0 + \sum_{i=1}^{n} F_{i,t}\beta_i))^2 + \lambda \sum_{j=1}^{n} \beta_j^2 . \tag{11}$$

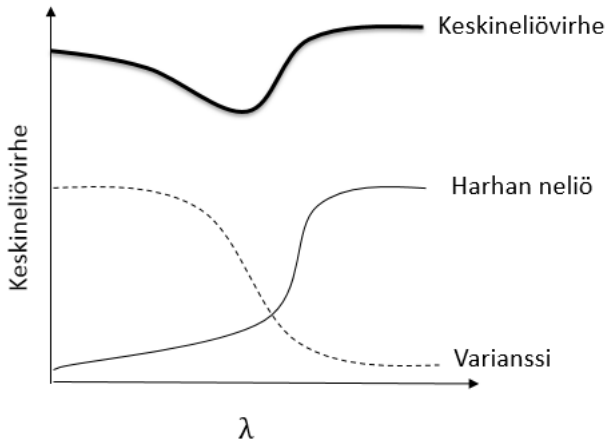
Parametria λ käytetään ohjaamaan rangaistustermin vaikutusta. Suuremmalla rajoitteen kertoimella on suurempi vaikutus parametrien pienenemiseen. Kun minimoidaan yhtälö regressiokertoimien suhteen ja jos rangaistustermi λ on yhtä suuri kuin nolla, niin käytännössä teemme perinteisen pienimmän neliösumman sovituksen. Kun λ suurenee, niin kertoimet lähestyvät nollaa. (Gareth ym. 2015, 215 – 220) Toisin kuin perinteinen pienimmän neliösumman metodi, joka tuottaa yhdet estimaatit kertoimille, harjanneregressio tuottaa eri estimaatit kertoimille $\beta_1, \dots, \beta_i$ jokaiselle λ:n arvolle. (Jin-Lung ja Tsay 2005, 6)

Tämän minimointitehtävän ratkaisuna saadaan harjanneregression kertoimet $\boldsymbol{\beta} = (\beta_1, \dots, \beta_i)'$ seuraavan kaavan mukaisesti:

$$\widehat{\boldsymbol{\beta}_\lambda} = \left[\sum_{t=1}^{W} F_{i,t}F_{i,t}' + \lambda \boldsymbol{I}_n\right]^{-1} \left(\sum_{t=1}^{W} F_{i,t}y_{t+h}\right), \tag{12}$$

missä $\boldsymbol{I}_n$ on nxn identiteettimatriisi, n on ennustemuuttujien lukumäärä, $i = 1, \dots, n$ ja λ on rangaistustermi, joka valitaan liukuvassa ikkunassa W. (Jin-Lung ja Tsay 2005, 6)

Harjanneregression etu suhteessa perinteiseen pienimmän neliösumman metodiin liittyy harha-varianssidekompositioon. Kun λ suurenee, niin harjanneregression joustavuus pienenee, joka johtaa varianssin pienenemiseen, mutta harhan suurenemiseen. Tämä on havainnollistettu kuvassa 2, missä on ennusteiden harhan neliö, ennusteiden varianssi ja ennusteiden keskineliövirhe. λ:n ollessa nolla varianssi on suuri ja harha on pieni. Kun λ:n arvo nousee, harjanneregression parametrit pienenevät ja tämä johtaa ennusteiden varianssin pienenemiseen ja harhan kasvamiseen. (Gareth ym. 2015, 217 – 220)



Kuva 2 Ennusteiden harhan neliö, varianssi ja keskineliövirhe eri rangaistustermin arvoilla

Harjanneregressiossa muuttujat pitää standardisoida ennen kuin lähdetään ratkaisemaan edellä mainittua yhtälöä. Standardointi tarkoittaa sitä, että jokaisella alkuperäisellä ennustemuuttujalla on keskiarvo 0 ja varianssi 1. (Hastie ym. 2009, 63 – 64)

4.3 Pääkomponenttiregressio

Jos suurta määrää ennustemuuttujia käytetään rajoituksettomassa regressiossa, niin tämä johtaa myös ylisovittamiseen ja näin huonoihin ennustetuloksiin (Stock ja Watson 1999, 314). Yksi tapa lähestyä tämän kaltaista korkean dimension ongelmaa on estimoida faktoreita ennustemuuttujien joukosta. Faktorit ovat täten F_t:n pääkomponentteja, jotka on muodostettu käyttämällä havaintoja ajassa t ja aikaisemmin. (Stock ja Watson 1999, 314 – 316) Pääkomponenttianalyysi on menetelmä datan dimensionaalisuuden pienentämiseksi. Pääkomponenttianalyysillä supistetaan ennustemuuttujien määrää pienemmäksi,

mutta kuitenkin niin, että tämä pienempi supistettu muuttujajoukko sisältää kaiken oleellisen informaation alkuperäisistä ennustemuuttujista. Tämä metodi sopii hyvin tutkielman asetelmaan, sillä ennustemuuttujien määrä on suuri.

Pääkomponenttiregressiossa (PCR) muodostetaan p kappaletta pääkomponentteja, $z_{1,t}, \dots, z_{p,t}$, jotka ovat lineaariyhdistelmiä alkuperäisistä muuttujista F_t. Näitä pääkomponentteja käytetään sitten ennustemuuttujina lineaarisessa mallissa, joka sovitetaan pienimmän neliösumman menetelmällä otosdataan. Tavoitteena on, että pieni määrä pääkomponentteja riittää selittämään suurimman osan aineiston vaihtelevuudesta ja sen, miten vastemuuttuja riippuu niistä. (Gareth ym. 2015, 233 – 236)

Pääkomponentit muodostetaan seuraavasti. Ensin kaikki ennustemuuttujat tulee normalisoida seuraavasti:

$$\sum_{t=1}^{W} F_{i,t} = 0 \tag{13}$$

ja

$$\sum_{t=1}^{W} F_{i,t}^2 = 1, \tag{14}$$

missä W on liukuvan ikkunan pituus. Tämä normalisointi tehdään jokaiselle ennustemuuttujalle $F_{i,t}$, missä $i = 1, \dots, n$. (Jin-Lung ja Tsay 2005, 3)

Normalisoitujen ennustemuuttujien avulla muodostetaan varianssi-kovarianssi -matriisi $\boldsymbol{S}_W = \sum_{t=1}^{W} F_{t,} F_t'$, josta saadaan pääkomponentit ominaisvektori-matriisihajoitelmalla:

$$\boldsymbol{S}_W = \boldsymbol{P}\, \boldsymbol{\Lambda}\, \boldsymbol{P}', \tag{15}$$

missä $\boldsymbol{\Lambda}$ on diagonaalimatriisi ominaisarvoista siten, että ne on järjestetty suuruusjärjestykseen $\boldsymbol{\lambda}_1 > \boldsymbol{\lambda}_2 > \boldsymbol{\lambda}_3$ jne. Matriisi $\boldsymbol{P} = [\boldsymbol{e}_1, \dots, \boldsymbol{e}_n]$, missä n on ennustemuuttujien lukumäärä, on muodostettu ominaisvektoreista siten, että $\boldsymbol{e}_i$ vastaa ominaisarvoa $\boldsymbol{\lambda}_1$. Pääkomponentit saadaan näiden ominaisvektorien ja ennustemuuttujavektorin matriisitulona, eli i:s pääkomponentti saadaan:

$$z_{i,t} = \boldsymbol{e}_i' F_{i,t}, \tag{16}$$

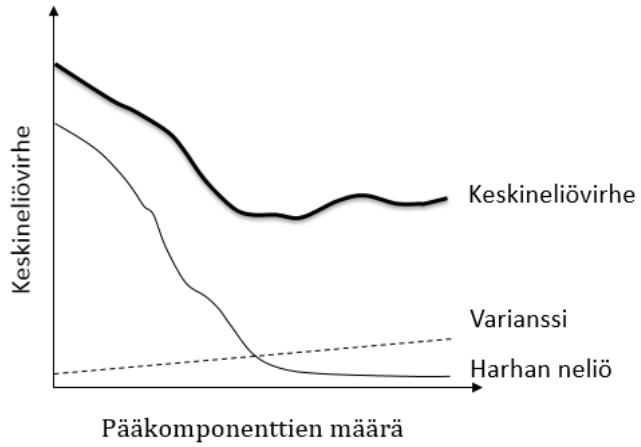
missä $z_{i,t}$ on ennustemuuttujan $F_{i,t}$ i:s pääkomponentti ja $\boldsymbol{e}_i$ on i:s ominaisvektori. (Jin-Lung ja Tsay 2005, 3)

Näitä estimoituja pääkomponentteja käytetään regressiossa, jonka kaava voidaan kirjoittaa muotoon:

$$y_{t+h} = \beta_0 + \sum_{i=1}^{p} \beta_i z_{i,t} + \varepsilon_{t+h}, \quad (17)$$

missä $z_{1,t}, \dots, z_{p,t}$ ovat pääkomponentteja ja p on valittu pääkomponenttien määrä. (Jin-Lung ja Tsay 2005, 3 – 4)

Kuvassa 3 on esitetty pääkomponenttiregressioon liittyvä ennusteiden harhan neliön, varianssin ja keskineliövirheen kehitys. Pienellä määrällä pääkomponentteja ennusteiden keskineliövirhe ja harhan neliö ovat suuria ja varianssi on pieni. Lisäämällä pääkomponentteja regressioon ennusteiden harhan neliö ja ennusteiden keskineliövirhe laskevat nopeasti ja ennusteiden varianssi nousee hitaasti. Lisäämällä komponentteja ennusteiden keskineliövirhe ei enää laske vaan se tasoittuu. (Gareth ym. 2015, 234)



Kuva 3 Ennusteiden harhan neliö, varianssi ja keskineliövirhe eri pääkomponenttien määrillä

Jos pääkomponenttiregressiomallin oletukset pitävät, niin sovittamalla pienimmän neliösumman menetelmä lineaarisiin yhdistelmiin alkuperäisistä ennustemuuttujista johtaa tarkempaan ennustukseen kuin alkuperäisten ennustemuuttujien käyttö. On helppo nähdä, että jos pääkomponenttien lukumäärä on sama kuin alkuperäisten ennustemuuttujien lukumäärä, niin pääkomponenttiregressio tuottaa pienimmän neliösumman sovituksen kaikilla ennustemuuttujilla F_t. Jos estimoidaan vain $p < n$ pääkomponenttia, niin mallilla ei ole yhtä suurta ylisovittamisen vaaraa kuin pienimmän neliösumman menetelmällä kaikilla alkuperäisillä ennustemuuttujilla, koska mallin sovittamislähestymistapa on joustavampi. (Gareth ym. 2015, 233 – 236)

Kun käytetään pääkomponenttiregressiota, niin ennustemuuttujat täytyy standardisoida ennen kuin luodaan pääkomponentit. Standardointi tarkoittaa sitä, että jokaisella

alkuperäisellä ennustemuuttujalla on keskiarvo 0 ja varianssi 1. Tämä standardisointi tuo kaikki muuttujat samalle mitta-asteikolle, eli minkään muuttujan varianssi ei dominoi pääkomponentteja. (Hastie ym. 2009, 79 – 80)

PCR on yleisesti käytetty tapa dimension pidentämiseen, mutta sillä on myös haittapuolensa. Suuren variaation pääkomponentit pidetään ja pienen variaation pääkomponentit hylätään. Vastemuuttujasta ei pääkomponenttien muodostuksessa ole lainkaan informaatiota. Eli käytetyt pääkomponentit eivät välttämättä sisällä relevanttia informaatiota vastemuuttujasta ja hylätyt pääkomponentit voivat ollakin hyödyllisiä. (Jin-Lung ja Tsay 2005, 3 – 4)

4.4 Osittaisen pienimmän neliön regressio

Osittaisen pienimmän neliön regressio (PLS) (engl. ”partial least squares regression”) on myös dimension pienentämisen menetelmä. Tämä menetelmä on erityisen hyödyllinen silloin, kun datassa on havaittavissa paljon multikollineaarisuutta, eli muuttujat korreloivat vahvasti keskenään. (Gareth ym. 2015, 237)

PLS-metodissa uudet ennustemuuttujat muodostetaan ohjatulla tavalla. Tämä tarkoittaa sitä, että vastemuuttujaa käytetään löytämään uudet ennustemuuttujat, jotka estimoivat sekä vanhoja ennustemuuttujia että myös vastemuuttujaa. Tämä menetelmä yrittää siis löytää suuntia, jotka selittävät sekä vastemuuttujan että ennustemuuttujat. Toisin sanoen menetelmä etsii lineaarisen regressiomallin projisoimalla ennustetut arvot ja havaitut ennustemuuttujien arvot uuteen dimensioavaruuteen. PLS-metodi antaa suurimman painoarvon niille muuttujille, joista vastemuuttuja riippuu voimakkaimmin. Se etsii korrelaatiota vastemuuttujan ja ennustemuuttujien välillä painottaen suurinta korrelaatiota lineaarikombinoidussa uudessa ennustemuuttujassa. (Gareth ym. 2015, 237 – 243)

Toisin kuin PCR, PLS ei käytä ennustemuuttujien välistä kovarianssia. Ensimmäinen askel on projisoida kaikki ennustemuuttujat pieneen määrään komponentteja, käyttäen korrelaatiota vastemuuttujan ja ennustemuuttujien välillä. Näitä uusia komponentteja käytetään myöhemmin ennustukseen. Ensimmäinen komponentti $z_{1,t}$ muodostetaan seuraavalla kaavalla:

$$z_{1,t} = \sum_{i=1}^{n} Cov(y_t, F_{i,t})\, F_{i,t}, \tag{18}$$

missä $F_{i,t}$ on alkuperäisten ennustemuuttujien joukko ja n on ennustemuuttujien lukumäärä. Seuraavaksi regressoidaan y_t ja $F_{i,t}$ komponenttiin $z_{1,t}$. Tämän regression residuaalit ovat $\tilde{y}_t$ ja $\tilde{F}_{i,t}$. Seuraavaksi muodostetaan toinen komponentti $z_{2,t}$. (Jin-Lung ja Tsay 2005, 4)

$$z_{2,t} = \sum_{i=1}^{n} Cov(\tilde{y}_t, \tilde{F}_{i,t})\, \tilde{F}_{i,t}, \tag{19}$$

missä n ennustemuuttujien lukumäärä. Näitä vaiheita iteroidaan, kunnes on muodostettu haluttu määrä komponentteja. Näitä estimoituja komponentteja käytetään regressiossa, joka estimoidaan käyttäen pienimmän neliösumman menetelmää:

$$y_{t+h} = \sum_{i=1}^{p} \beta_i z_{i,t} + \varepsilon_{t+h}, \tag{20}$$

missä $z_{1,t}, \dots, z_{p,t}$ ovat PLS komponentteja ja p on komponenttien lukumäärä. (Jin-Lung ja Tsay 2005, 4)

Tämä metodi vaatii muuttujien standardisointia. Tämä standardisointi takaa sen, että kaikki muuttujat ovat samalla skaalalla. Jos standardisointia ei tehtäisi, niin muuttujilla, joilla on suurin varianssi, olisi suurin vaikutus muodostetussa mallissa. (Gareth ym. 2015, 238) Aion tutkia eri komponenttien määriä PCR- ja PLS-metodeissa ja vertailla niiden keskineliövirheen neliöjuurta otoksen ulkopuolisilla havainnoilla.

4.5 Satunnaismetsä

Satunnaismetsämenetelmä (RF) (engl. random forest) on yksi menestyksekkäimmistä metodeista koneoppimisessa ja sitä pidetään tarkimpana yleiskäyttöisenä koneoppimismenetelmänä, jossa ei tehdä minkäänlaisia oletuksia ennustemuuttujien jakaumista tai merkitsevyydestä. Se sopii erityisesti tilanteisiin, joissa on hyvin suuri määrä ennustemuuttujia ja suuri dimensionaliteetti. (Biau ja D'Elia 2011, 1 – 4) Alun perin se kehitettiin luokitusmenetelmäksi (Breiman 2001), mutta menetelmä soveltuu hyvin myös regressioon ja ennustemuuttujien valintaan. Satunnaismetsämenetelmä on laajasti käytössä tiedon louhinnassa (engl. data mining) eri tieteenaloilla, mutta makrotaloustieteessä sitä on sovellettu vasta viime aikoina (Biau ja D'Elia 2011, 1).

Baiu ja D'Elia (2011) tarkastelevat satunnaismetsän soveltamista taloustieteessä ennustamalla BKT:n kasvuvauhtia euro-alueelle. He käyttävät tähän European Union Business and Consumer Survey -dataa eri sektoreilta. Lopputulos on, että satunnaismetsä ei tuottanut yhtä tarkkoja ennusteita kuin lineaarinen malli, mutta satunnaismetsästä saadut ensimmäiset kymmenen tärkeintä muuttujaa lineaarisessa mallissa toimivat hyvin.

Satunnaismetsä on yhdistelmämenetelmä, jossa yhdistetään heikomman menetelmän tuloksia paremman tuloksen saamiseksi. Tässä tapauksessa heikompi menetelmä on päätöspuu. Päätöspuu rakennetaan jakamalla aineiston useaan kertaan ennustemuuttujien suhteen. (Gareth ym. 2015, 319)

Satunnaismetsä käyttää regressiopuita mallin rakentamiseen. Satunnaismetsän rakentaminen alkaa valitsemalla tietty määrä regressiopuita, jotka rakennetaan havaintoaineistolla. Nämä regressiopuut vastaavat heikkoja menetelmiä, joiden tulokset yhdistetään paremman tuloksen aikaansaamiseksi. (Gareth ym. 2015, 319 – 321)

Regressiopuu, kuten myös päätöspuu, on tilastollinen oppimismenetelmä, jossa ennustemuuttujien avaruus jaetaan moneen erilliseen alueeseen, joilla ei saa olla päällekkäisiä osia (ts. leikkaus on tyhjä joukko). Tietyn alueen kaikille havainnoille tehdään sama ennuste, joka on keskiarvo alueen kaikista havainnoista. Esimerkiksi, jos ennustemuuttuja-avaruus jaetaan kahteen erilliseen osaan R_1 ja R_2, joille vastemuuttujan keskiarvoennuste havainnoista alueella R_1 on 10 ja keskiarvoennuste havainnoista alueella R_2 on 20, niin kaikille niille havainnoille jotka sijaitsevat alueella R_1 annamme ennusteeksi 10 ja jos havainnot sijaitsevat alueella R_2, niin annamme niille ennusteen arvoksi 20. Päätöspuu eroaa regressiopuusta siten, että siinä keskiarvoistamisen sijaan alueen arvo muodostetaan äänestämällä. Alueet voivat olla minkä muotoisia tahansa, mutta tavoitteena on löytää alueet, jotka minimoivat aluejakoon perustuvan jäännösneliösumman (Gareth ym. 2015, 306):

$$\sum_{j=1}^{J}\sum_{t=1}^{W}(y_{t+h} - \bar{y}_{R_{j,t}})^2, \tag{21}$$

missä W on liukuva ennusteikkuna, y_{t+h} on havaittu arvo ajalla $t+h$, J on alueiden määrä ja $\bar{y}_{R_{j,t}}$ on vastemuuttujan keskiarvo opetushavainnoista alueella $R_{j,t}$. Käytännössä kaikkia mahdollisia aluejakoja ei voida kokeilla ennusteikkunassa. Siksi usein käytetään ahnetta (engl. ”greedy”) ylhäältä alaspäin etenevää binääristä ennustemuuttuja-avaruutta jakavaa lähestymistapaa. Jakotapaa kutsutaan ahneeksi, koska jakoprosessissa otetaan huomioon vain tämänhetkisen jaon optimitulos, eikä tulevien jakojen optimia ennakoida. Ennustemuuttujien avaruuden jakaminen aloitetaan regressiopuun latvasta, siis ennustemuuttujien koko avaruudesta. Jokaisessa jaossa avaruus ajetaan kahteen osaan, eli päätöspuuhun muodostetaan kaksi uutta oksaa. (Gareth ym. 2015, 303 – 307)

Tätä prosessia jatketaan jakamalla ennustemuuttujien avaruus yhä uudelleen ja valitsemalla ennustemuuttujat ja leikkauspisteet, jotka jakavat aikaisemmissa jaoissa muodostuneet alueet tavoitefunktion suhteen optimaalisiin alueisiin. (Gareth ym. 2015, 303 – 307)

Yksittäinen regressiopuu rakennetaan siis solmu solmulta valitsemalla satunnaisesti M:stä ennustemuuttujasta m muuttujaa ($m < M$) puun solmuun. Se ennustemuuttuja valitusta m muuttujasta, joka tuottaa parhaimman näytejoukon jaon jollakin kohdefunktiolla, valitaan binäärijaon perustaksi puun seuraavalle asteelle. Uusi otos ennustemuuttujista otetaan jokaisessa puun jaossa. Näin jatketaan rekursiivisesti, kunnes kaikki jaot tuottavat saman tuloksen käytetyllä kohdefunktiolla. Gareth ym. 2015, 303-304, 314)

Satunnaismetsäalgoritmi ei missään vaiheessa käytä kaikkia ennustemuuttujia yhdessä solmussa. Syy tähän on se, että jos aineistossa on yksi vahva ennustemuuttuja, niin suurin osa puista käyttäisi tätä ennustemuuttujaa jo ensimmäisessä jaossa. Tämä johtaisi siihen, että kaikki regressiopuut näyttäisivät samanlaisilta ja tämän tuloksena puiden ennusteet olisivat erittäin korreloituneita. Suurin ero bootstrap-aggregoitujen regressiopuiden ja satunnaismetsän välillä on käytettävien ennustemuuttujien osajoukon koko. Jos ennustemuuttujien osajoukon koko on sama kuin ennustemuuttujien lukumäärä ($p = P$), niin satunnaismetsä tuottaa saman tuloksen kuin bootstrap-aggregaatio. (Gareth ym. 2015, 319 – 320)

Formaalisti satunnaismetsämenetelmän toiminta voidaan esittää seuraavasti. Ensin muodostetaan regressiopuu siten, että valitaan ennustemuuttuja $x_{j,t}$ F_t:n joukosta ja näytejoukon jakopiste *s*, joka jakaa näytejoukon kahteen osaan $N_1[j,s] = \{x_{j,t} \leq s\}$ ja $N_2[j,s] = \{x_{j,t} \geq s\}$. Haetaan seuraava ennustemuuttuja $x_{j,t}$ ja jakopiste *s* toteuttaa yhtälön:

$$min_{j,s}\left[min_{c_1}\sum_{x_{j,t}\in N_1[j,s]}(y_t - c_1)^2 + min_{c_2}\sum_{x_{j,t}\in N_2[j,s]}(y_t - c_2)^2\right], \qquad (22)$$

missä $c_1$ja c_2 ovat kyseisen näytejoukon jaon vastemuuttujien y_t keskiarvot. Näin löydettyjen uusien $x_{j,t}$ ja *s* avulla jaetaan näytejoukko kahteen alijoukkoon. Näin jatketaan rekursiivisesti, kunnes saavutetaan etukäteen asetettu pienin näytejoukon koko. (Biau ja D'Elia 2011, 4 – 6)

Seuraavaksi muodostetaan puulle yhdistelmäennustemuuttuja käyttäen puun lehtiä, eli alimman tason solmuja, jotka eivät enää jakaudu:

$$h(\boldsymbol{X}) = \frac{1}{Card\{\frac{t}{x_{j,t}} \in N(\boldsymbol{X})\}}\sum_{\frac{t}{x_{j,t}}\in N(\boldsymbol{X})} y_t, \qquad (23)$$

missä $N(\boldsymbol{X})$ tarkoittaa puun lehtiä, jotka kattavat otososajoukon $\mathbf{X}$ ja $Card\{\frac{t}{x_{j,t}} \in N(\boldsymbol{X})\}$ antaa alimmantason lehtisolujen lukumäärän. (Biau ja D'Elia 2011, 4 – 6)

Satunnaismetsälle, jossa on K-kappaletta regressiopuita, muodostetaan lopullinen ennuste keskiarvoistamalla kaikkien regressiopuiden yhdistelmäennusteet (Biau ja D'Elia 2011, 4 – 6):

$$h(\boldsymbol{X}) = \frac{1}{K}\sum_{k=1}^{K} h_k\,(\boldsymbol{X}). \tag{24}$$

Koska satunnaismetsä käyttää vain osajoukkoa ennustemuuttujista jokaisessa jaossa, niin dominoivin ennustemuuttuja ei voi tulla mukaan jokaiseen jakoon. Tällä tehdään regressiopuu korreloimattomaksi (engl. de-correlating), mikä johtaa siihen, että puiden keskiarvoennustus on vähemmän vaihteleva ja siten luotettavampi. Mikäli ennustemuuttujia on paljon ja ne korreloivat voimakkaasti, kannattaa käyttää pientä määrää satunnaisesti valittuja ennustemuuttujia. Koska satunnaismetsässä on monta regressiopuuta, joille on satunnaisesti valittu ennustemuuttujien alijoukko, niin menetelmässä on aina puita, jotka jättävät pois otosjoukossa olevaa kohinaa. (Gareth ym. 2015, 320 – 321)

Hyvä satunnaismetsämalli sovittautuu hyvin uuteen dataan. Sen sisältämät regressiopuut ovat tarkkoja ja ennustemuuttujien kattavuus on hyvä. Satunnaismetsämallia voidaan käyttää myös ennustemuuttujien tärkeysjärjestyksen arvioimiseen regressio-ongelmissa. Tämä tapahtuu seuraavasti. Tehdään satunnaismetsämalli datalle. Mallin muodostamisen aikana mitataan ns. “out-of-bag”-virhe, jossa lasketaan keskimääräinen ennustevirhe havaintoaineistossa niillä päätöspuilla, joiden muodostamisen yhteydessä ei ole käytetty tätä nimenomaista havaintoaineistoa. Näin laskettujen virheiden keskiarvo lasketaan yli koko satunnaismetsän kaikkien ennustepuiden. Mallin havaintoaineiston jakson jälkeen ennustemuuttujan arvot havaintoaineistossa permutoidaan ja ”out-of-bag”-virhe lasketaan uudelleen permutoidulla datalla. Ennustemuuttujan tärkeysarvo tai tärkeysjärjestyssija saadaan laskemalla keskiarvo ”out-of-bag”-virheiden erotuksesta ennen ja jälkeen datan permutoinnin kaikissa päätöspuissa. Tämä tulos normalisoidaan ja standardisoidaan näiden eron keskihajonnalla. Ennustemuuttujat, jotka tuottavat suuren arvon tälle tulokselle ovat tärkeämpiä muuttujia, kuin ne, jotka saivat pienemmän arvon. (Breiman 2001)

Satunnaismetsä-menetelmä ei vaadi samanlaista hienosäätöä kuin useat muut yhdistelmämenetelmät. Menetelmässä on vain kolme parametria, joiden avulla sen suorituskykyyn voidaan vaikuttaa:

- satunnaisesti valittavien ennustemuuttujien määrä (m),
- päätöspuiden lukumäärä satunnaismetsässä,
- päätöspuun koko sen lehtien lukumäärä.

Kun satunnaismetsää käytetään regressioon, käytetään tyypillisesti $m = M/3$ ennustemuuttujaa päätöspuun solmuissa. (Gareth ym. 2015, 329)

5 EMPIIRINEN ANALYYSI

Tässä luvussa sovellan edellä kuvattuja malleja, joiden teoreettista taustaa kävin läpi aikaisemmassa luvussa, makrotaloudelliseen aineistoon, joka on kerätty Macrobond-nimisestä taloustieteellisestä tietokannasta. Macrobond-tietokantaan on koottu historiallista talous- ja finanssidataa 115 maasta päivä-, viikko-, kuukausi- ja neljännesvuositasolla. Tietokannan datat on kerätty virallisista kansallisista tilastoista, keskuspankkien tilastoista, yritysten tilinpäätöksistä ja pörsseistä. (Macrobond Economics Database)

Taulukossa 1 on ennustemuuttujat, jotka ovat samoja kuin mitä Stock ja Watson (1999) ovat käyttäneet. Heidän työnsä sisälsi myös muita muuttujia.

Taulukko 1 Ennustemuuttujat ja niiden kuvaukset Macrobond-tietokannasta

Ennustemuuttujien nimet	**Ennustemuuttujien kuvaukset**
Unemployment.CPS.16.Years.Over	Unemployment of 16 years and over
Producer.Price.Index..All.Commodities	Producer price index including all commodities
Civilian.Employment.SA	Civilian employment seasonally adjusted
Construction.Started.Residential.Total.SA	total residential construction started seasonally adjusted
Consumer.Price.Index.All.Items.Less.Shelter.SA	Consumer price index including all items less shelter seasonally adjusted
Consumer.Price.Index.Durables.SA	Consumer price index durable goods seasonally adjusted
Consumer.Price.Index.Commodities	Consumer price index including commodities
Consumer.Price.Index.Services.SA	Consumer price index service goods seasonally adjusted
Consumer.Price.Index.Transport.SA	Consumer price index transportation seasonally adjusted
Employees.on.Nonagricultural.Payrolls.SA	Employees on non-agricultural payrolls seasonally adjusted
Producer.Price.Index.Finished.Goods.SA	Producer price index finished goods seasonally adjusted
Monetary.Aggregates.M1.Total	Monetary aggregates M1
Monetary.Aggregates.M2.Total.SA	Monetary aggregates M2 seasonally adjusted
NYSE.Composite.Index.Average.of.Period	New York Stock Exchange composite index average
Durable.Goods.Price.Index	Durable goods price index

Non.Durable.Goods.Price.Index	Non-durable goods price index
Services.Price.Index	Services price index
Producer.Price.Index.Crude.Materials.Total	Producer price index crude materials total
Unemployment.Less.than.5.Weeks.SA	Unemployment less than 5 weeks seasonally adjusted
Unemployment.5.to.14.Weeks	Unemployment 5 to 14 weeks seasonally adjusted
Unemployment.15.Weeks.Over.Total.SA	Unemployment 15 weeks over total seasonally adjusted
Unemployment.5.to.16.Weeks	Unemployment 15 to 26 weeks seasonally adjusted
Unemployment.27.Weeks.Over.SA	Unemployment 27 weeks over seasonally adjusted
Unemployment.Average.Mean.SA	Unemployment average mean seasonally adjusted
Non.Industrial.Supplies.Final.Equipment.Total.SA.Index	Non-industrial supplies final equipment total seasonally adjusted index
Non.Industrial.Supplies.Final.Consumer.Goods.Total.SA.Index	Non-industrial supplies final consumer goods total seasonally adjusted index
Manufacturing.Total.SA.Index	Manufacturing total seasonally adjusted index
Materials.Total.SA.Index	Materials total seasonally adjusted index
Industrial.Production.Total.SA.Index	Industrial production total seasonally adjusted index
Non.Industrial.Supplies.Final.Consumer.Goods.Durable.Total.SA	Non-industrial supplies final consumer goods durable total seasonally adjusted
Non.Industrial.Supplies.Final.Consumer.Goods.Non.Durable.SA	Non-industrial supplies final consumer goods non-durable seasonally adjusted
Total..Personal.Income.Excluding.Current.Transfer.Receipts.Constant.Prices.SA	Total personal income excluding current transfer receipts constant prices seasonally adjusted
Manufacturing.Purchasing.Managers.Index	Manufacturing purchasing managers index
Total..Disposable.Personal.Income.Total.Constant.Prices.SA.Chained	Total disposable personal income constant prices seasonally adjusted
Average.Weekly.Initial.Claims..Unemployment.Insurance..SA	Chained ja average weekly initial claims of unemployment insurance seasonally adjusted

Tämän lisäksi aion käyttää yllämainituista ennustemuuttujista 12 periodin viiveitä mukana monimuuttujametodeissa.

Ennusteongelmana on ennustaa inflaation vauhtia Yhdysvalloissa vuosina 1970–2015 erilaisilla ennustemuuttujilla ja vertailla eri menetelmillä saatujen ennusteiden tarkkuutta. Ennustemuuttujat kuuluvat seuraaviin kategorioihin: tuotannon määrä ja tulot, työllisyys, vähittäismyynti, valmistus ja myynti, kulutus, kiinteistöt, varastot ja tilaukset, osakkeiden hintaindeksit, raha-aggregaatit, hintaindeksit ja palkat.

Tutkimuksen kohteena on 3, 6 ja 12 kuukauden kasvun ennustaminen ja seuraan Stock ja Watson (2004) tapaa esittää vastemuuttuja:

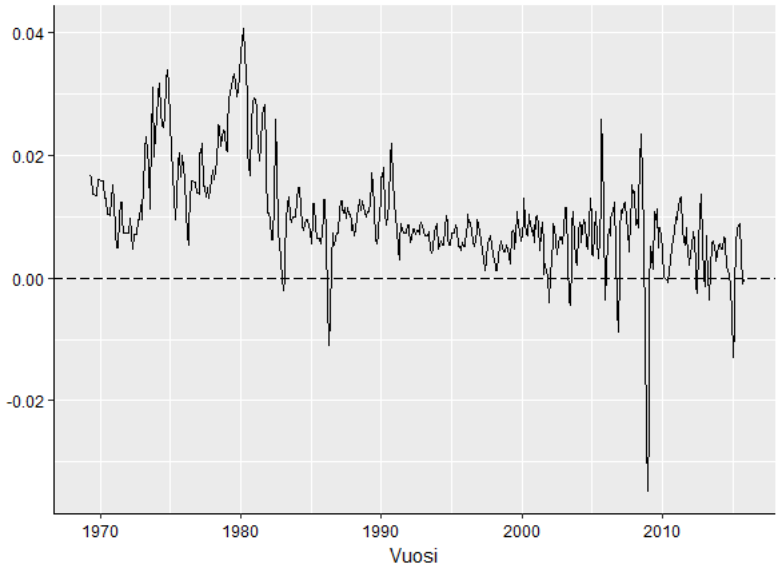
$$y_{t+h} = \frac{1200[\ln(Y_{t+h}) - \ln(Y_t)]}{h}, \tag{25}$$

missä Y_t ja Y_{t+h} ovat prosessoimattomia aikasarjoja ja h on tutkimuksen kohteena oleva horisontin pituus. Kaikille ennustemuuttujille on tehty edellä mainittu transformaatio kuten Stock ja Watson (1999, 2004) ovat kuvanneet. Näistä transformoiduista aikasarjoista on otettu 12 periodin viiveet mukaan ennustemuuttujiksi.

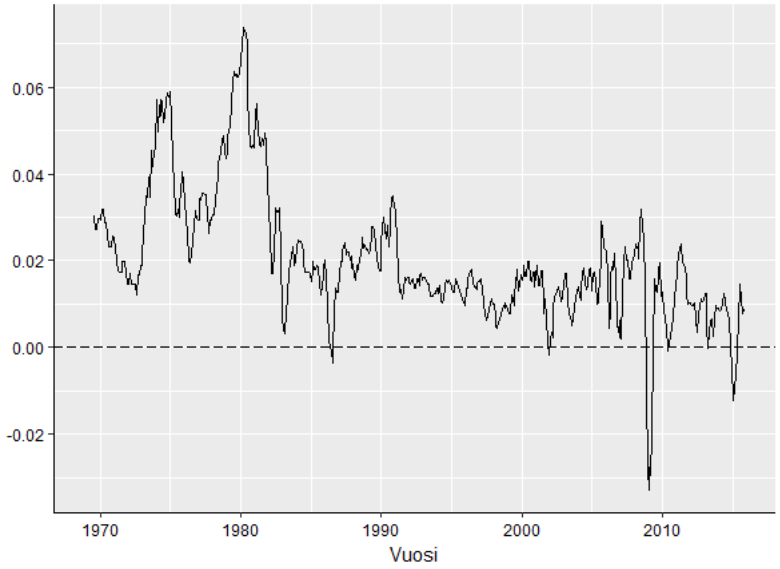
Jokaiselle metodille estimoidaan malli liukuvassa ikkunassa ja tällä mallilla tehdään ennuste $\hat{y}_{t+h}$. Tämän jälkeen ikkuna liukuu yhden havainnon eteenpäin ja malli estimoidaan uudelleen ja sillä tehdään uusi ennuste seuraavalle havainnolle. Kriteeri, jolla mittaan ennusteiden tarkkuutta on ennusteen keskineliövirheen neliöjuuri. Ennusteen keskineliövirheen neliöjuuri on pieni, kun ennustetut arvot ovat mahdollisimman lähellä toteutuneita arvoja (Stock 2001, 566; Gareth ym. 2015, 29 – 30). Keskineliövirheen neliöjuuren empiirinen estimaatti voidaan kirjoittaa seuraavaan muotoon:

$$RMSE = \sqrt{\sum_{t=W-h+1}^{T-h} \frac{(y_{t+h} - \hat{y}_{t+h})^2}{T - W}}, \tag{26}$$

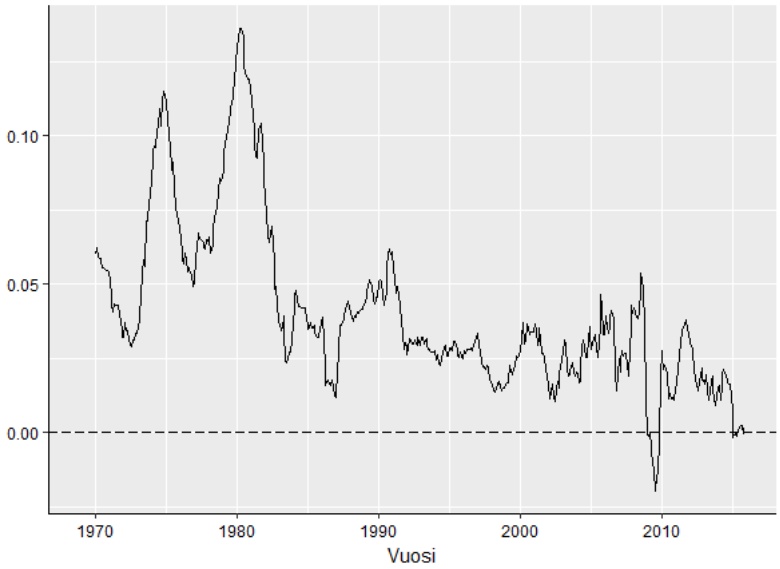
missä T on otoskoko, h on ennusteen horisontti, W on ennusteikkuna, y_{t+h} on havaittu arvo ajalla $t + h$ ja $\hat{y}_{t+h}$ on sitä vastaavaa ennuste ajalle $t + h$. (Jin-Lung ja Tsay 2005, 8) Ennustemuuttujia on yhteensä 481 ja havaintoja on 525 kun $h = 12$, havaintoja on 532 kun $h = 6$ ja havaintoja on 535 kun $h = 3$. Kuvat 4, 5 ja 6 esittävät toteutuneet transformoidut kuluttajahintaindeksit kullekin aikahorisontille.



Kuva 4 Kuluttajahintaindeksin kolmen kuukauden kasvuvauhti



Kuva 5 Kuluttajahintaindeksin kuuden kuukauden kasvuvauhti



Kuva 6 Kuluttajahintaindeksin kahdentoista kuukauden kasvuvauhti

Tavoitteena on tutkia erilaisten metodien ennustetarkkuutta, kun ennustetaan inflaatiota mitattuna kuluttajahintaindeksin muutosnopeudella. Taustaoletuksena on, että aineistossa on jokin riippuvuussuhde tapahtuneen havainnon y_{t+h} ja ennustemuuttujien F_t välillä, eli F_t sisältää jotain systemaattista informaatiota vastemuuttujasta y_{t+h}.

Ennusteiden luomisessa käytän rullaavan regression -metodologiaa samalla tavalla kuin Inoue ja Kilian 2006 ovat kuvanneet. Parametrillisten regressiomallien käyttö etenee seuraavasti. Ensin estimoidaan mallin parametreille lähtöarvot oppimisperiodin ensimmäisen liukuvan ikkunan sisältävällä aineistolla. Estimoitujen parametrien avulla muodostetaan seuraavan periodin ennuste, jota verrataan vastaavaan havaintoon laskemalla ennustevirhe ennusteen ja seuraavan periodin toteutuman välillä. Seuraavaksi siirretään mallin opetusperiodi seuraavalle periodille ja laaditaan sen avulla muodostettujen estimaattien avulla ennuste kolmannelle periodille ja lasketaan ennustevirhe ennusteen ja toteutuman välillä. Näin edeten saadaan ennusteet ja niitä vastaavat ennustevirheet jokaiselle testausperiodin ajan hetkelle. Käytettäessä rullaavaa regressiota ikkunan kokoa pidetään koko ajan vakiona sen liikkuessa periodista toiseen. Olen valinnut liukuvan ikkunan pituudeksi 120 kuukautta, kuten Exterkate ym. (2012).

Rullaavan regression vaihtoehtona voitaisiin käyttää rekursiivista regressiota, joka hyödyntää kaikkea ennusteajankohtaa edeltävää dataa muodostaessaan ennusteen seuraavalle periodille. Siinä ikkunan alkuajankohta pysyy samana, mutta ikkunan loppuajankohta kasvaa siirryttäessä periodilta toiselle. Koska rekursiivinen regressio pitää muistissaan koko ennustehistorian, se reagoi hitaammin nopeisiin fundamenttien muutoksiin kuin rullaava regressio. Siksi käytän rullaavaa regressiota analyysissäni.

5.1 Stationaarisuus

Ennen kuin metodeja voidaan soveltaa, on syytä kiinnittää huomiota aikasarjojen stationaarisuuteen. On tärkeää huomioida aikasarjoille tyypillinen epästationaarinen rakenne. Tämän tarkastelun tarkoituksena on selvittää, millaista esiprosessointia vastemuuttuja ja ennustemuuttujat vaativat. On kahdenlaista stationaarisuutta, joista yleisempi on heikko stationaarisuus. Aikasarjan sanotaan olevan heikosti stationaarinen, jos seuraavat ehdot toteutuvat.

1. Havaintojen keskiarvo ei riipu ajanhetkestä t: $E(y_t) = \mu$,
2. Havaintojen varianssi ei riipu ajanhetkestä t: $Var(y_t) = E(y_t - \mu)^2$,
3. Havaintojen kovarianssi riippuu niiden ajallisesta erosta: $cov(y_t, y_{t-k}) = \mathrm{E}\{(y_t - \mu)(y_{t-k} - \mu)\}$, $k = 1, 2, 3, \ldots$

Jos kuluttajahintaindeksiä ennustavan mallin ennustemuuttujana on toinen epästationaarinen aikasarja, ovat ennustemuuttujille estimoidut parametrit virheellisiä, ja mallista saadut ennusteet huonoja. (Verbeek 2004, 258 – 259; Brooks 2014, 181 – 188) Kirjallisuudessa pidetään tunnettuna, että inflaation 12-kuukauden prosentuaalinen muutos on $I(1)$, eli integroitunut asteella 1 (Stock ja Watson 1999, 296). Aikasarjan sanotaan olevan $I(1)$ jos se on epästationaarinen, mutta sen ensimmäinen differenssi on stationaarinen.

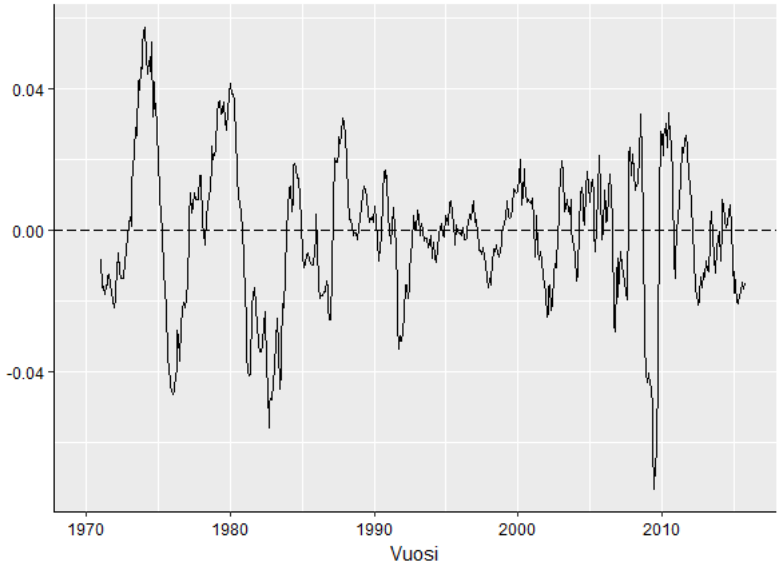
Tässä tutkielmassa tavoitteena on ennustaa transformoituja aikasarjoja, eikä alkuperäisiä aikasarjoja. Syynä tähän on se, että käytetyt metodit olettavat stationaarisuuden. Jos ennusteet laadittaisiin vastemuuttujalle, joka ei ole stationaarinen, niin ongelmaksi tulisi ns. spurious regression, jonka seurauksena mallin tuottamat keskineliövirheet ovat vääristyneitä (Cowpertwait ja Metcalfe 2009, 211 – 214). Kuten Granger ja Newbold (1974) osoittivat, jos kahta epästationaarista ja toisistaan riippumatonta aikasarjaa käytetään regressiossa, niin todennäköisyys päätyä spurious regressioon on suuri.

Tavanomainen käytäntö on käyttää yksikköjuuri testiä selvittämään, onko aikasarja stationaarinen ja tämän perusteella voidaan päättää monenko asteen differentiointi pitää tehdä. Yleisesti käytetty stationaarisuuden testi on laajennettu Dickey-Fuller -testi (ADF-testi). (Stock 2001, 574) ADF-testin nollahypoteesi on, että aikasarja sisältää yksikköjuuren ja on täten epästationaarinen. Vaihtoehtoinen hypoteesi vastaavasti on, että prosessi on stationaarinen. Muita yleisesti käytettyjä testejä ovat Phillips-Perron (PP) testi, joka on johdettu ADF-testistä, ja Kwiatkowski-Phillips-Schmidt-Shin (KPSS) testi, missä nollahypoteesi on, että aikasarja ei sisällä yksikköjuurta ja on täten stationaarinen. Vaihtoehtoinen hypoteesi on, että aikasarja sisältää yksikköjuuren ja on täten epästationaarinen. (Verbeek 2004, 267 – 273)

Olen testannut kuluttajahintaindeksin ja ennustemuuttujien stationaarisuutta yllä mainituilla testeillä, jotta voin tehdä oikeanlaiset muunnokset aikasarjoille. ADF-testiä varten

valitut viiveiden pituudet on valittu käyttämällä Bayesian-informaatiokriteeriä. Kun $h = 12$ vastemuuttujalle kaikki kolme yksikköjuuritestiä hylkäävät stationaarisuuden, ja päädyn siis siihen tulokseen, että aikasarja on epästationaarinen. Kun $h = 3$ ja $h = 6$ vain KPSS-testi hylkää stationaarisuuden. Kun otan näistä aikasarjoista ensimmäisen differenssin, niin testit eivät hylkää stationaarisuutta. Tämän tuloksen perusteella otan myös näistä kahdesta transformoidusta aikasarjasta differenssiä. Kun ennustemuuttujista ja niiden viiveistä otetaan ensimmäisen asteen differenssi, ovat kaikki muuttujat stationaarisia 5% luottamustasolla.

Kuvassa 7 on 12 kuukauden kasvuvauhdin ensimmäinen differenssi. Tämä differenssi on laskettu samalla tavalla kuin Stock ja Watson (1999).



Kuva 7 Kuluttajahintaindeksin kahdentoista kuukauden kasvuvauhdin ensimmäinen differenssi

5.2 Tulokset

AR(p)-prosessi on lineaarinen malli, jossa ennustemuuttujina ovat aikaisemmat havainnot y_{t-1}, ..., y_{t-p}. Verbeek (2004) mainitsee kirjassaan, että Bayesian-informaatiokriteerin käyttö on suositeltavaa, sillä Akaiken-informaatiokriteerillä on taipumus yliparametrisoida malleja. Tämän takia käytän Bayesian-informaatiokriteeriä valitsemaan viiveiden määrän ja käytän tästä merkintää AR(BIC). Tämä viiveiden etsiminen on rajattu $1 \leq p \leq 12$, samalla tavalla kuin Stock (2001).

Kaikissa empiriaosuuden taulukoissa ja kuvissa on keskineliövirheiden neliöjuuret kerrottu sadalla, jotta tuloksia on helpompi vertailla keskenään. Taulukossa 2 on AR(BIC) tuottamat keskineliövirheet listattuna eri h:n arvoille.

Taulukko 2 AR(BIC) tuottamat ennusteiden keskineliövirheiden neliöjuuret

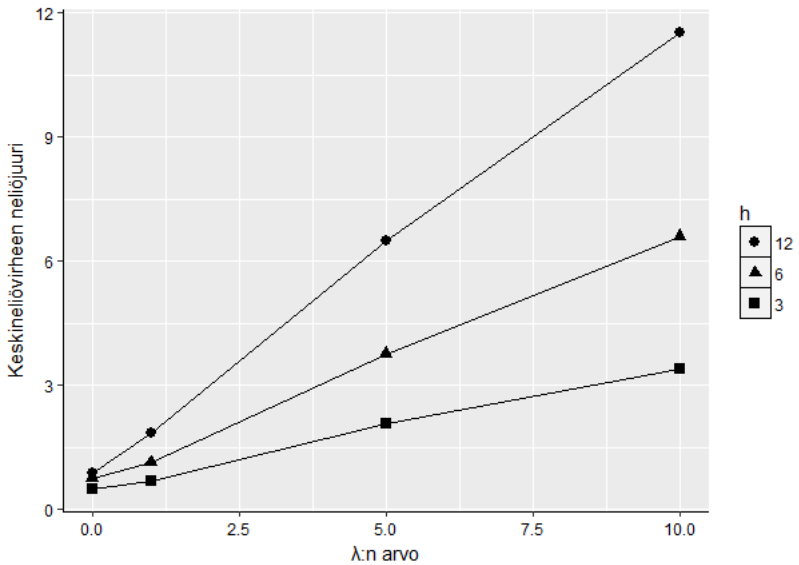
Horisontti		
3	6	12
0.2767	0.3441	0.4402

Harjanneregressio pienentää kertoimia asettamalla rajoituksen kertoimien kokoon ja pienentämällä niitä kohti nollaa. Tässä metodissa käytetään kaikkia alkuperäisiä muuttujia ennustamiseen. Tämän metodin ideana on korjata multikollineaarisuuden tuomia ongelmia. Taulukossa 3 on harjanneregression tuottamat keskineliövirheet eri h:n ja λ:n arvoille.

Taulukko 3 Harjanneregression (RR) tuottamat keskineliövirheiden neliöjuuret

Horisontti			
λ	3	6	12
0	0.4902	0.7431	0.8937
1	0.6968	1.1486	1.8536
5	2.0633	3.7664	6.4761
10	3.3892	6.6018	11.521

Taulukosta näkee, että nämä keskineliövirheet ovat suuremmat kuin AR(BIC) tuottamat keskineliövirheet. Kuvassa 8 on harjanneregression virheet eri λ:n arvolla.



Kuva 8 Harjanneregressiossa (RR) käytetyn rangaistustermin suuruuden vaikutus keskineliövirheiden neliöjuuriin

Kuvasta nähdään, miten virhe kehittyy muuttamalla λ:n arvoa. Tästä kuvasta ja taulukosta 3 näkee, että λ: n arvo 0 tuotti pienimmän keskineliövirheen h:n arvoille $3, 6$ ja 12.

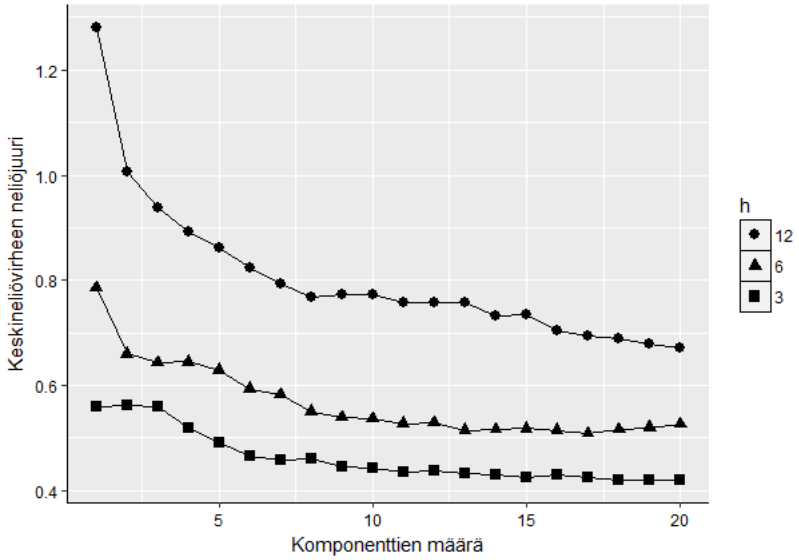
Pääkomponenttiregressiolla supistetaan ennustemuuttujien määrää pienemmäksi, mutta kuitenkin niin, että tämä pienempi supistettu muuttujajoukko sisältää kaiken oleellisen informaation alkuperäisistä ennustemuuttujista. Kun aineistossa on paljon muuttujia, niin pääkomponentit auttavat tiivistämään aineiston pienempään muuttujien lukumäärään. Nämä muodostetut pääkomponentit ovat lineaarinen yhdistelmä alkuperäisistä muuttujista ja niitä käytetään ennustemuuttujina lineaarisessa mallissa. Olen valinnut testattavaksi ensimmäiset 20 pääkomponenttia, kuten Jin-Lung ja Tsay (2005) ovat tehneet.

Taulukoissa 4 on pääkomponenttiregression tuottamat keskineliövirheet pääkomponenteille 1-20 eri h:n arvoille. Taulukoista näkee, että tarkin ennuste saatiin 18, 17, ja 20 pääkomponentilla kun $h = 3, 6$ ja 12.

Taulukko 4 Pääkomponenttiregression (PCR) tuottamat keskineliövirheiden neliöjuuret

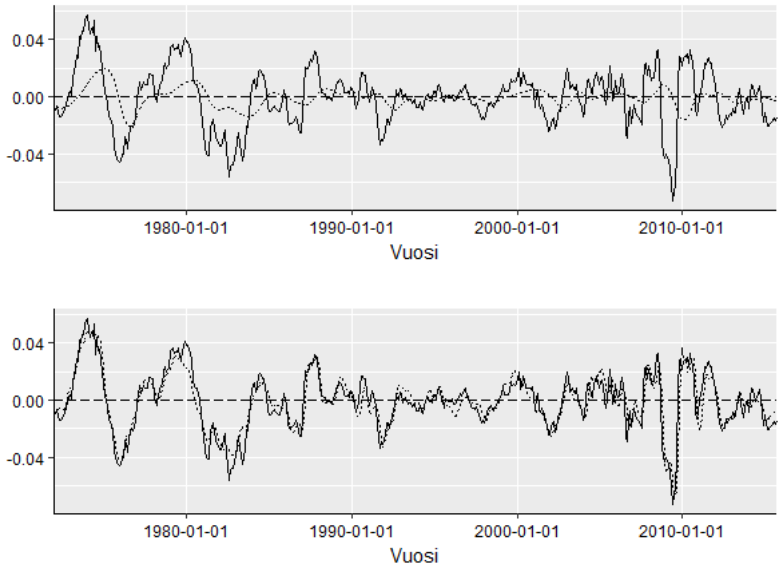
Horisontti			
komponentit	3	6	12
1	0.5590	0.7851	1.2814
2	0.5622	0.6598	1.0064
3	0.5591	0.6433	0.9370
4	0.5180	0.6444	0.8911
5	0.4898	0.6289	0.8609
6	0.4656	0.5928	0.8223
7	0.4579	0.5821	0.7939
8	0.4594	0.5497	0.7664
9	0.4459	0.5390	0.7734
10	0.4421	0.5366	0.7734
11	0.4336	0.5272	0.7580
12	0.4382	0.5284	0.7580
13	0.4320	0.5142	0.7563
14	0.4305	0.5161	0.7311
15	0.4242	0.5173	0.7344
16	0.4298	0.5144	0.7032
17	0.4237	0.5083	0.6950
18	0.4185	0.5154	0.6894
19	0.4192	0.5197	0.6795
20	0.4190	0.5256	0.6705

Kuvassa 9 on pääkomponenttiregression keskineliövirheet eri pääkomponenttien määrillä. Kuvasta näkee, miten virhe kehittyy lisäämällä pääkomponentteja regressioon. Kun $h = 6$ ja $h = 12$ niin näemme, että keskineliövirheen neliöjuuri tasoittuu noin kymmenennen komponentin kohdalla.



Kuva 9 Pääkomponenttiregressiossa (PCR) käytettyjen komponenttien lukumäärien vaikutus keskineliövirheiden neliöjuuriin

Kuvassa 10 on ensimmäisen pääkomponentin avulla tehdyt ennusteet ja ensimmäisen kahdenkymmenen pääkomponentin avulla tehdyt ennusteet kuluttajahintaindeksin 12 kuukauden kasvuvauhdin ensimmäiselle differenssille. Toteutuneet havainnot on merkitty kiinteällä viivalla ja mallien ennusteet on merkitty katkoviivalla. Kuvasta näkee, kuinka suuri ero näiden kahden mallin ennusteilla on.



Kuva 10 Ensimmäisen ja ensimmäisen kahdenkymmenen pääkomponentin ennusteen ero kuluttajahintaindeksin kahdentoista kuukauden kasvuvauhdin ensimmäiselle differenssille

PLS on dimension pienentämisen menetelmä, on myös dimension pienentämisen menetelmä. Olen valinnut testattavaksi ensimmäiset 20 komponenttia.

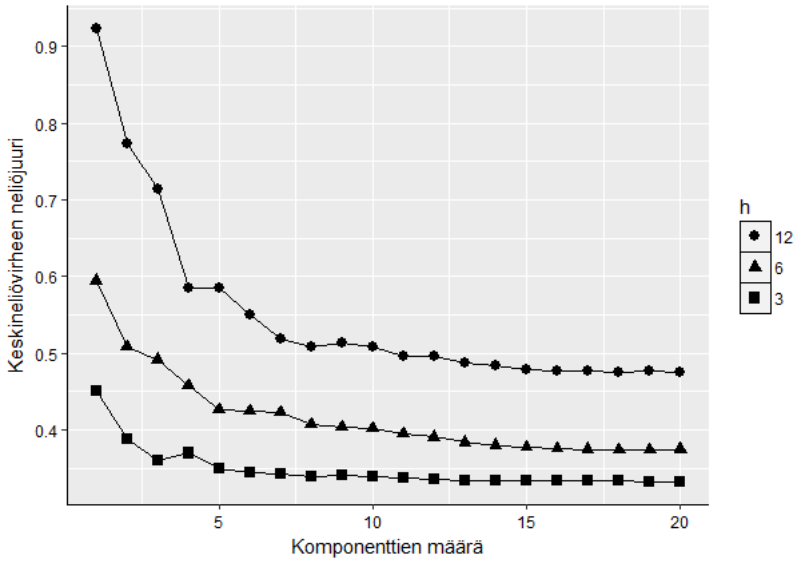
Taulukossa 5 on listattu PLS-metodin tuottamat keskineliövirheet eri määrillä komponentteja kun $h = 3, 6$ ja 12. Taulukoista näkee, että tarkin ennuste saatiin 20, 18, ja 20 komponentilla.

Taulukko 5 Osittaisen pienimmän neliön regression (PLS) komponenttien lukumäärien keskineliövirheiden neliöjuuret

Horisontti			
komponentit	3	6	12
1	0.4505	0.5942	0.9236
2	0.3874	0.5078	0.7741
3	0.3597	0.4917	0.7144
4	0.3697	0.4571	0.5848
5	0.3500	0.4262	0.5848
6	0.3451	0.4241	0.5504
7	0.3426	0.4225	0.5196
8	0.3391	0.4066	0.5086
9	0.3409	0.4036	0.5141
10	0.3399	0.4014	0.5082
11	0.3376	0.3945	0.4959
12	0.3353	0.3906	0.4959
13	0.3337	0.3837	0.4868

14	0.3346	0.3797	0.4832
15	0.3337	0.3779	0.4783
16	0.3338	0.3757	0.4764
17	0.3341	0.3746	0.4765
18	0.3333	0.3738	0.4757
19	0.3327	0.3744	0.4762
20	0.3316	0.3747	0.4746

Kuvassa 11 on osittaisen pienimmän neliösumman regression keskineliövirheet eri komponenttien määrillä.



Kuva 11 Osittaisen pienimmän neliön (PLS) regressiossa käytettyjen komponenttien lukumäärien vaikutus keskineliövirheiden neliöjuuriin

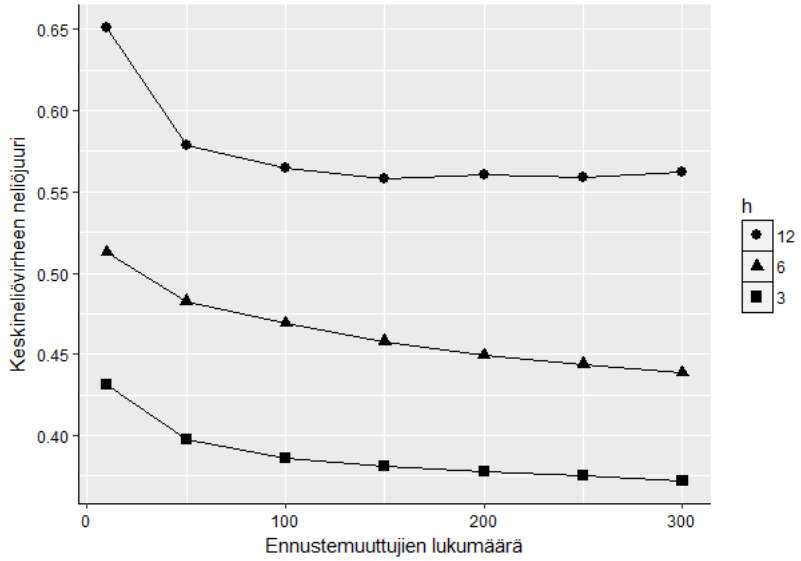
Kuvasta nähdään, miten virhe kehittyy lisäämällä komponentteja regressioon. Kun $h = 6$ ja $h = 12$ niin näemme, että keskineliövirheen neliöjuuri tasoittuu noin kymmenennen komponentin kohdalla kuten pääkomponenttiregressiossa.

Taulukossa 6 on satunnaismetsän tuottamat keskineliövirheet 10, 50, 100, 150, 200, 250 ja 300 satunnaisesti valituilla muuttujilla kun $h = 3, 6$ ja 12. Taulukosta näkee, että kun $h = 3$ ja $h = 6$ niin 300 tuotti pienimmän keskineliövirheen ja kun $h = 12$ niin 150 tuotti pienimmän keskineliövirheen. Kun $h = 12$, niin tärkeimmät muuttujat olivat vastemuuttujan ensimmäinen viive, Consumer.Price.Index.All.Items.Less.Shelter.SA ja Consumer.Price.Index.All.Urban.Consumers.SA. Kun $h = 6$ tärkeimmiksi muuttujiksi valikoitui vastemuuttujan ensimmäiset kaksi viivettä ja Non.Durable.Goods.Price.Index neljäs viive. Kun $h = 3$, niin tärkeimmiksi muuttujiksi valikoitui vastemuuttujan ensimmäinen viive, Consumer.Price.Index.All.Items.Less.Shelter.SA yhdestoista viive ja Consumer.Price.Index.All.Urban.Consumers.SA yhdestoista viive.

Taulukko 6 Satunnaismetsään (RF) valittujen muuttujien lukumäärien keskineliövirheiden neliöjuuret

Horisontti			
Satunnaisesti valittujen muuttujien lukumäärä	3	6	12
10	0.4310	0.5127	0.6512
50	0.3975	0.4827	0.5784
100	0.3864	0.4690	0.5647
150	0.3810	0.4580	0.5582
200	0.3781	0.4493	0.5604
250	0.3756	0.4439	0.5585
300	0.3718	0.4388	0.5623

Kuvassa 12 on satunnaismetsässä käytettyjen ennustemuuttujien lukumäärät ja niiden keskineliövirheiden neliöjuuret. Kuvasta näkee, että ennustevirhe tasaantuu kun $h = 12$ noin 150 muuttujan kohdalla.



Kuva 12 Satunnaismetsään (RF) valittujen ennustemuuttujien lukumäärien vaikutus keskineliövirheiden neliöjuuriin

5.3 Yhteenveto tuloksista

On normaalia käytäntöä valita paras ennustemetodi vertailemalla metodeja niiden otoksien ulkopuolisten virheiden mukaan (Inoue ja Kilian 2006, 274). Taulukko 7 näyttää mallien relatiivisen keskineliövirheen neliöjuuren samaan tapaan kuin Stock (2001).

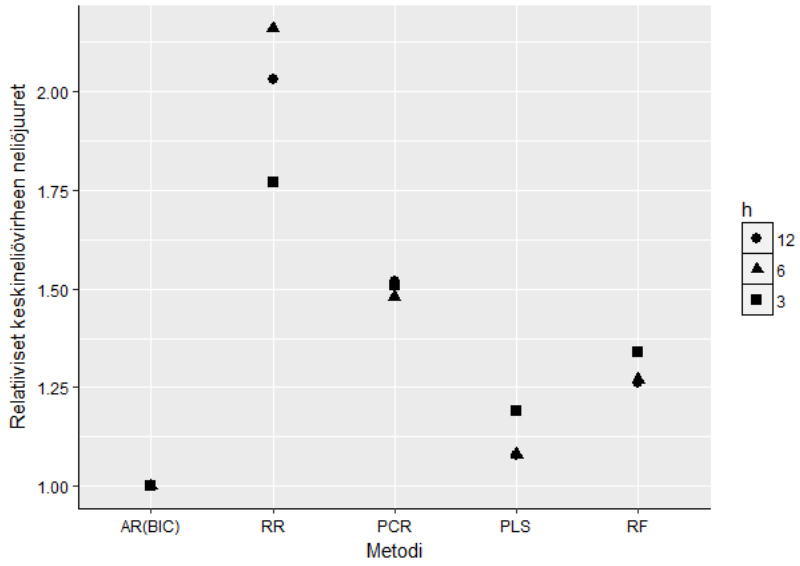
AR(BIC) tuotti pienimmän virheen ja muita metodeja vertaillaan tähän. Pienempi relatiivinen keskineliövirheen neliöjuuri tarkoittaa tarkempaa ennustetta simuloidussa otoksen ulkopuolisessa ennusteessa.

Taulukko 7 Yhteenveto mallien soveltuvuudesta

Mallin nimi	**Relatiivinen keskineliövirheen neliöjuuri $h = 3$**	**Relatiivinen keskineliövirheen neliöjuuri $h = 6$**	**Relatiivinen keskineliövirheen neliöjuuri $h = 12$**
AR(BIC)	1.00	1.00	1.00
RR	1.77	2.16	2.03
PCR	1.51	1.48	1.52
PLS	1.19	1.08	1.08
RF	1.34	1.27	1.26

AR(BIC), eli autoregressiivinen prosessi, jossa viiveiden määrä on valittu minimoimalla Bayesian-informaatiokriteeri, tuotti pienimmän virheen. PCR- ja PLS-metodit eivät tuottaneet hyviä tuloksia, kun komponenttien määrä on pieni. Suuremmalla määrällä komponentteja PLS-metodi oli kuitenkin lähellä AR(BIC)-metodin tuottamaa keskineliövirhettä kaikilla h:n arvoilla. Molemmissa metodeissa keskineliövirheen neliöjuuri tasaantui kaikilla h:n arvoilla, kun malliin lisättiin tarpeeksi monta komponenttia. Oletuksena on, että molemmissa metodeissa keskineliövirheen neliöjuuri nousee, kun malliin lisättäisiin vielä enemmän komponentteja niin, että komponenttien määrä p lähenee alkuperäisten ennustemuuttujien lukumäärää.

Kuvassa 13 on eri metodien relatiiviset keskineliövirheet eri h:n arvoille. Kuvasta on helppo vertailla eri metodien relatiivisia keskineliövirheitä eri horisonteilla.



Kuva 13 Metodien relatiiviset keskineliövirheiden neliöjuuret eri horisonteilla

Sofistikoidummat mallit eivät tuoneet parannusta autoregressiiviseen malliin nähden. Tulokset satunnaismetsän kohdalla ovat yhdenmukaisia Biau ja D'Elia (2010) kanssa, jotka testasivat satunnaismetsää BKT:n ennustamiseen. Samanlaisia tuloksia sai mm. Stock (2001), joka vertaili neuroverkkoja (NN) ja varianssi-autoregressiivista mallia (VAR) AR(BIC) ja AR(4) malleihin. Tuloksena oli, että neuroverkot eivät tuoneet parannusta ennusteeseen ja varianssi-autoregressiivinen malli toi pientä parannusta AR(4):ään verrattuna.

6 JOHTOPÄÄTÖKSET

Ennustemallien muodostusmenetelmien määrä on kasvanut viimeisen vuosikymmenen aikana merkittävästi tietokonepohjaisten menetelmien seurauksena. Ei ole yhtä dominoivaa menetelmää, vaan ennustemallin muodostamismenetelmän valinta riippuu kontekstista, kuten aineiston määrästä ja mallien käytön kokemuksesta. Ennusteet ovat yleensä vain yksi osa informaatiosta, jota käytetään päätösten tekemiseen. Se ei korvaa kontekstin tuntemusta, vaan täydentää sitä. Ennusteita käytetään antamaan painoarvoja eri tapahtumamahdollisuuksille. Puhtaat ”black box”-mallit eivät ole saaneet suurta huomiota ekonomistien keskuudessa ja niitä ei ole käytetty niin laajasti kuin olisi voinut olettaa. (Elliott ja Timmermann 2008, 50 – 51)

Tässä tutkielmassa vertailin eri metodeja Yhdysvaltojen inflaation ennustamiseksi käyttämällä kuukausitason dataa vuosilta 1970–2015. Vertailin autoregressiivistä prosessia, harjanneregressiota, pääkomponenttiregressiota, osittaisen pienimmän neliön regressiota ja satunnaismetsää. Lopputulos oli se, että vastemuuttujan viiveisiin perustuva AR(BIC) tuotti pienemmän keskineliövirheen kuin muut metodit.

Taloustieteellistä tuntemusta tarvitaan ennustusprosessin useissa vaiheissa. Sen avulla voidaan valita mielekäs tavoitefunktio ennustuksen sisältämille optimoinneille. Ennustemuuttujien valinta kannattaa perustaa taloustieteelliseen näkemykseen siitä, mitkä muuttujat ovat relevantteja teoreettisesta näkökulmasta. Ekonometriset menetelmät avustavat myös ennustemallin funktionaalisen muodon valitsemisessa. (Elliott ja Timmermann 2008, 50 – 51)

Taloustieteen ja rahoituksen aikasarjamallit eivät ole yleensä stabiileja eri aikajaksoina, joten näiden mallien ennusteet ovat parhaimmillaan suuntaa-antavia ja ajanhetkeen sidottuja. Ei siis kannata olettaa, että sama malli tuottaa parhaan tuloksen eri aikakausilla. (Elliott ja Timmermann 2008, 50 – 51)

Lisätutkimus voisi käyttää yhdistelmämalleja, jotka käyttävät muiden mallien ennustuksia ennustemuuttujina. Toinen mielenkiintoinen aihe olisi selvittää, mitkä mallit soveltuvat parhaiten eri ajanjaksoille ja miten ajanjaksot jakautuvat eri mallien tarkkuuden mukaan. Alkuperäisen aineiston dimension pienentämisen ja muuttujien valintametodien kaksi muuta potentiaalista kandidaattia ovat Tibshiranin (1996) esittelemä LASSO (Least absolute shrinkage and selection operator) -metodi ja Efron ym. (2004) esittelemä LARS (Least angle regression) -metodi.

7 LÄHTEET

Atkeson, A. ja Ohanian L. E. (2001) *Are Phillips Curves Useful for Forecasting Inflation?* Federal Reserve Bank of Minneapolis Quarterly Review Volume 25, No. 1.

Biau, O. ja D'Elia, A. (2011) *Euro area GDP forecasting using large survey datasets*. 6th Eurostat Colloquium on Modern Tools for Business Cycle Analysis: the lessons from global economic crisis, 26th - 29th September 2010

Breiman, L. (1996) *Bagging Predictors*. Machine Learning. New York: Springer.

Breiman, L. (2001) *Random Forests.* Machine Learning. New York: Springer.

Brooks, C. (2014) *Introductory Econometrics for Finance*. New York: Cambridge University Press.

Burnham, K., Anderson, D. (2002) *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach*. Second Edition, 2002 Springer

Burnham, K., Anderson, D. (2004) *Multimodel inference: understanding AIC and BIC in Model Selection*. Sociological Methods & Research 33.

Cowpertwait, P., Metcalfe, A. (2009) *Introductory Time Series with R*. New York: Springer

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). *Least angle regression*. Annals of Statistics 32, 407–499.

Elliott, G., Timmermann, A. (2008) *Economic Forecasting*. Journal of Economic Literature, 46(1): 3-56.

Exterkate, P., Groenen, P., Heij, C., Dick, D. (2012) *Nonlinear Forecasting With Many Predictors Using Kernel Ridge Regression*. CREATES Research Papers, Department of Economics and Business Economics, Aarhus University.

Fortmann-Roe, S. (2012) *Understanding the Bias-Variance Tradeoff*. <http://scott.fortmann-roe.com/docs/BiasVariance.html> haettu 23.07.2016

Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2015) *An Introduction to Statistical Learning*. New York: Springer.

Geman, S., Bienenstock, E., Doursat, R. (1992) *Neural networks and the bias/variance dilemma.* Neural computation, MIT Press Vol. 4, No. 1.

Granger, C., Newbold P. (1974) *Spurious Regressions in Econometrics*. University of Nottingham. Journal of Econometrics 2 111-120.

Guyon I., Elisseeff A., (2003) *An Introduction to Variable and Feature Selection*. Journal of Machine Learning Research 3 1157-1182.

Hastie, T., Tibshirani, R., Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Inoue, A., Kilian, L. (2006) *On the selection of forecasting models*. Journal of Econometrics 130 273 - 306

Jin-Lung, L., Tsay R.S. (2005) *Comparison of Forecasting Methods with Many Predictors*. Graduate School of Business, University of Chicago

Macrobond Economics Database, <https://www.macrobond.com>

Rudin, C. (2012) *Prediction: Machine Learning and Statistics.* Spring 2012. Massachusetts Institute of Technology: MIT OpenCourseWare, <https://ocw.mit.edu> haettu 12.03.2017

Shumway, R., Stoffer, D. (2011) *Time Series Analysis and Applications*. New York: Springer.

Stock, J.H., Watson, M. (2004) *Forecasting with many predictors*. forthcoming Handbook of Economic Forecasting

Stock, J.H. (2001) *Forecasting economic time series*. Companion in Theoretical Econometrics. Basil Blackwell, Malden, MA.

Stock, J.H., Watson M.W. (1999) *Forecasting inflation*. Journal of Monetary Economics 44.

Tibshirani R. (1996). *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society B, 58, 267-288.

Verbeek, M. (2004) *A Guide to Modern Econometrics, Second Edition*. John Wiley & Sons Ltd.