TALLINN UNIVERSITY OF TECHNOLOGY

Faculty of Information Technology

Department of Software Science

Karl Taivo Kama, 176487 IAPM

ANALYSIS AND DEVELOPMENT OF NORDPOOL ELECTRICITY PRICE PREDICTION MODEL

Master's thesis

Supervisor: Juri Belikov, PhD

Tallinn 2020

TALLINNA TEHNIKAÜLIKOOL

Infotehnoloogia teaduskond

Karl Taivo Kama, 176487 IAPM

NORDPOOLI ELEKTRIHINNA ANALÜÜS JA ENNUSTUSMUDELI ARENDAMINE

Magistritöö

Juhendaja: Juri Belikov, PhD

Tallinn 2020

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Karl Taivo Kama

19.04.2020

Analysis and development of Nordpool electricity price prediction model

Abstract

Current work studies electricity price dependency on various factors. The primary goal is to analyse Nordpool electricity market and create a model that can predict Estonian electricity prices. The prices have rather unpredictable behaviour. Knowledge and better prediction of this behaviour can save many household's and consumer's funds.

First, I investigated the parameters that have impact on the prices and selected the most significant ones. This paper also analyses models that were produced by various researchers from different contries, which exhibited many prediction techniques. Study also displays the data origins and how it's being processed. Models and their implementation used in this study are also described. Finally, the paper displays the effect of the COVID-19 and oil prices crisis on the created price prediction model.

Main outcome of this study is an Estonian electricity price prediction model, which has *mean absolute error* of 4.815 and *Mean absolute percentage error* of 11.730 before the crisis. Crisis had a notable effect on the models accuracy. The best model's accuracy dropped due to the crisis around 2.6 times.

Present thesis is written in Estonian and is presented on 68 pages of study. Thesis is devided to 3 main chapters, which include 7 tables and 28 figures.

Nordpooli elektrihinna analüüs ja ennustusmudeli arendamine

Annotatsioon

Antud töö uurimisobjektiks võtsin elektrihindade sõltumise erinevatest faktoritest. Põhieesmärgiks on uurida ning analüüsida Nordpooli elektriturgu ning analüüside põhjal koostada mudelid, millega oleks võimalik Eesti elektrihindasid ennustada. Elektrihinnad on tihtipeale väga ettearvamatud ja nende hindade täpsem ennustamine võib paljudel majapidamiste ning muudel tarbijatel hulga kulusid kokku hoida.

Töö käigus võtsin vaatluse alla elektrihindasid mõjutavad faktorid, uurisin millised nendest on paremad, et võimalikult täpne mudel koostada. Samuti uurisin olemasolevaid erinevate riikide andmete põhjal loodud mudeleid ning seda milliseid meetodeid on nende puhul rakendatud. Lisaks kirjeldan töös kasutatavaid andmeid, nende töötlust. Samuti mudeleid ning nende implementeerimist. Töö lõpus uurin samuti millist effekti avaldavad mudeli tööle töö kirjutamise ajale sattunud naftahinna ning koroona kriis.

Põhitulemuseks on mudel, mis võimaldab ennustada kriisvälisel ajal *Mean absolute error*-iga 4.815 ning *Mean absolute percentage error*-iga 11.730. Kriisi mõju on tuntav ka mudeli ennustuses. Selgus, et parima mudeli ennustustäpsus väheneb kriisi ajal ligi 2.6 korda.

Töö on kirjutatud eesti keeles ning sisaldab teksti 68 leheküljel. Töö on jagatud kolmeks suureks peatükiks ning sisaldab 7 tabelit ja 28 joonist.

Sisukord

Jooniste nimekiri

Tabelite nimekiri

Lühendite sõnastik

ANFIS Adaptive neuro fuzzy inference system.

ARIMA Autoregressive integrated moving average.

HPA Hybrid PSO–ANFIS.

LSTM Long short term memory.

MAE Mean absolute error.

MAPE Mean absolute percentage error.

MCP Market clearing price.

MSE Mean squared error.

NN Neural network.

PSO Particle swarm optimization.

RMSE Root mean square error.

RNN Recurrent neural network.

SARIMA Seasonal autoregressive integrated moving average.

SVM Support vector machine.

Sissejuhatus

Infotehnoloogiat rakendatakse tänapäeval järjest keerulisemate probleemide lahendamisel või lahti murdmistel. Isegi ettenägematut tulevikku aitavad arvutid meil paremini mõista ja tendentse märgata. Antud magistritöö uurimisobjektiks valisin elektrihindade sõltumise erinevatest faktoritest. Energiat rohkelt tarbivatele, ostvatele ja müüvatele ettevõtetele on väga oluline teada, milline on turu seis tulevikus. Igapäevaselt tehakse energiatööstuses palju strateegilisi otsuseid kui palju elektrit osta ja kui palju müüa. Sellist teadmist omades võib säästa palju raha ja saada turul suure eelise konkurentide ees. Peale selle kasutakse antud teadmist ära ka paljudes kodudes, eriti nö. targa kodu lahendustes. Näiteks on võimalik pesumasin, veeboiler, ventilatsioonisüsteem või mõni muu rohkelt energiat tarbiv kodumasin tööle panna kõige soodsamal päeval ja ajal.

Teada on, et elektri hinda võivad mõjutada väga palju väliseid faktoreid. Nendel on erinev mõju hinnale, mõni faktor mõjutab seda drastiliselt, mõni väga minimaalselt. Probleemiks on, kuidas koostada taolist mudelit, mis suudaks hinda võimalikult täpselt ennustada. Antud töö käigus tõstatan ja üritan leida vastust küsimustele:

- Millised faktorid mõjutavad kõige enam elektri hinda?
- Milliste faktorite andmeid on võimalik kõige paremini kasutada, et ennustada elektri hinda?
- Milline töös kasutatavatest ennustamismeetoditest on antud ülesande lahendamiseks kõige sobilikum?
- Kui täpselt on võimalik selliste mudelitega elektri hinda ennustada?
- Milline vea protsent on antud ennustuse puhul aktsepteeritav?
- Millist mõju on mudeli täpsusele avaldanud töö käigus tekkinud koroona vi-

iruse ja naftahinna kriis?

On olemas mitmeid masinõppe algoritme, mille abil saaksime antud probleemile lahenduse leida. Töö mahtu silmas pidades otsustasin analüüsida kahte masinõppe algoritmi, mille abil antud mudelit koostama hakkan. Antud töö käigus soovin uurida kõige rohkem lineaarregressioonist ja *Recurrent neural network*-i erijuhust *Long short term memory*-st saadavaid tulemusi ja tuua välja nende erinevused. Valisin antud meetodid uurimisobjektist lähtuvalt - lineaarregressiooni puhul on tegemist laiadaselt kasutatud meetodiga, millega saab hästi esialgseid järeldusi teha ning mida saab võrdlemisi lihtsalt rakendada. LSTM sobib täpsema mudeli loomiseks, millega on võimalik teha detailsemat analüüsi ja ennustusi.

Töö esimeses osas soovin uurida ja otsida tegureid, mis mõjutavad elektri hinda. Erinevatel parameetritel võib olla elektri hinna väljakujunemisel teatud roll. Oluline on eristada suure mõjuga tegureid nendest, mille mõju hinnale on pigem minimaalne. Need parameetrid on vaja selgelt välja tuua ning nendele fokusseeruda mudelite koostamisel. Samuti soovin uurida ja paika panna vea protsendi, mis on mudeli puhul aktsepteeritav. Protsent annab hinnangu mudeli tööle ning on oluline tulemuste analüüsi osas. Elektri hinna mõjutavate parameetrite ja aktsepteeritava vea protsendi piiri väljaselgitamiseks toetun teadusartiklitele.

Töö teises osas toon välja andmete saamise koha ning kirjeldan töös kasutatavaid andmeid. Samuti soovin teostada lineaar regressiooni ja LSTMi rakendamist ning tuua välja nende eripärad.

Töö kolmandas osas võrdlen ning analüüsin tulemusi. Selle käigus selguvad vastused eelnevalt püstitatud küsimustele: millised tegurid/parameetrid on kõige olulisemad elektrihinna väljakujunemises, milline algoritm annab kõige parema tulemuse. Samuti toimub antud osas tulemuste valideerimine. Saame vastused püstitatud küsimustele ja mudeli täpsusele. Mudeli hindamisel ja valideerimisel kasutan nii esimeses osas leitud aktsepteeritava vea protsendi piiri kui ka võrdlen tegelikku hinda mudeli poolt ennustatava hinnaga.

Elekter ja elektriturg

1.1 Elektrituru omapärad

Viimastelt kümnenditel on toimunud erinevates sektorites märgatavad muutused, et parandada konkurentsi tingimusi, sektori läbipaistvust, jälgitavust ning protsesse on palju efektiivsemaks tehtud. Taolised muutused on läbinud ka energeetikaturg, kus toimus deregulatsioon ja vabanemine valitsuste ning monopolide kontrolli all olevate energia tootmise korporatsioonide alt. Selle tulemusena on elekter saanud oluliseks ostu ja müügi objektiks.

Elektrit võib pidada üheks olulisemaks varaks, mille tootmine ja kasutamine erineb oluliselt teistest varadest. Erinevalt kullast, hõbedast ja teistest loodusvaradest ei ole võimalik elektrit suurtes kogustes talletada, mis tähendab, et elektrit tarbitakse koheselt peale selle tootmist. See omadus teeb elektrituru teistest turgudest palju ettearvamatuks. Seetõttu esineb hindades nii aastaaja kohast kui ka nädalast ja päevast trendi. Lisaks on hind ebapüsiv, mis tähendab, et toimuda võib palju hüppeid nii hinnas kui ka tarbimises.

Peale tarbijate nõudmiste mõjutavad elektri hinda ka välised faktorid nagu ilmastikuolud näiteks temperatuur, päike, tuul ja sademed. Need omakorda mõjutavad päikesepaneelide, tuulegeneraatorite või hüdroelektrijaamade tööd. Samuti mõjutavad hinda ka voolukatkestused, olgu nende allikaks alajaamad, elektriliinid või muu. Olenevalt regioonist on elektrijaamade esmasteks elektriallikateks on üldjuhul kivisüsi, põlevkivi või mõni muu kütus, millest on võimalik kiirelt energiat saada. Hind sõltub samuti sellest, kui palju kütust jaamas alles on.

Eelnevalt toodud näited on vaid väike tükike elektrihinna kujundamisel. Eelnevat arvesse võttes võib väita, et energiahinna ennustamine väga keeruline ja nõuab

suurt andmete kogust ja detailset analüüsi. Kui muidu oleme harjunud ostma kaupa tänase hinna eest ning ei tea, mida homne päev võib tuua, siis elektriga on teisiti. Elekter on omapärasel turul, kus on teada järgmise päeva hind. Seega on oluline ennustada ülehomse päeva erinevate tundide hind, et osata teada, millal oleks kõige mõistlikum elektrit osta või müüa [2, 3].

Turg kujuneb energiat tootvate ja tarbivate firmade pakkumistest järgmisele 24 tunnile. Pakkumiste puhul proovitakse ennustada, milline on tarbimine järgmisel päeval. Riikliku sekkumise vähendamine tegi hinna palju ettearvamatuks. Elektrit tootavad ettevõtted maandavad oma riske ennustustööriistade abil. Samuti luuakse pakkumise strateegiaid, et suurendada oma kasu võrreldes konkurentide omaga [4].

1.2 Elektri hinda mõjutavad faktorid

Eelnevast võib teha järeldusi, et elektri hind on keeruline, mis moodustub justkui pusletükkidena väga paljudest muutujatest kokku. Mõni nendest osadest on olulisemad ja suurema tähtsusega kui teised. Sellest hoolimata tekib nendest üks tervik.

On teada, et tarbimise suurus on otseselt seotud hinnaga. Tarbimine nii ärides, kodudes, tööstustes sõltub vastavalt majanduse seisust. Samuti on elekter otseselt seotud ka välistemperatuuriga ja päevavalguse tundidega, mis omakorda on seotud inimeste käitumisharjumustega. Mingil määral aitab selle teadmine märgata paremini elektrihinna muutust ajas, sest just seetõttu on elektri tarbimine teataval määral etteaimatav. Nimelt on selge, et esineb selgeid mustreid nii päevases, nädalases, kuises kui ka aastases elektri tarbimises. Inimestel on kindlad igapäevased harjumused – suurem osa meist ärkab hommikul kell 8, siis kasvab tarbimine. Minnakse tööle – firmad ja tehased alustavad tööd. Õhtul tullakse koju – valmistatakse sööki, vaadatakse televiisorit. Öösel magatakse ja energia tarbimine on väga madal. Samuti võib taolisi näiteid tuua ka aastaajaliselt, kus tarbimine suvel on madal. Sügise saabudes on järjest enam vaja majapidamisi kütta. Talvel jõuab kütmine haripunkti, mis taas kevade saabudes hakkab langema.

Elektrituru eesmärk on pakkumist ja nõudlust kokku viia, mis loob *Market clearing price (MCP)*. Elektri hinda mõjutavad faktoreid võib klassifitseerida neljaks osaks:

1. Fundamentaalsed muutujad
2. Operatiivsed muutujad
3. Strateegilised muutujad
4. Ajaloolised muutujad

Fundamentaalseteks muutujateks peetakse kõige olulisemaid faktoreid, milleks on:

- Kütuse hinnad maailmaturul (kivisüsi, gaas, nafta)
- Ilmastikuolud (sademed, tuul, päike)

- Välistemperatuur
- Ajahetk (kellaeg, nädalapäev, kuu, aastaaeg, tarbijate harjumused)
- Elektri tootmise kulud

Operatiivseteks muutujateks peetakse:

- Elektri tootmise suurus (ületootmine/alatootmine)
- Elektri koormus
- Võrgu ülekoormus
- Võrgu haldus
- Võrgu operatiivkulud

Strateegilisteks muutujateks peetakse:

- Turu ülesehitus
- Pakkumiste strateegia
- Elektri ostu ja müügi lepingud
- Kahepoolsed lepingud turuosaliste vahel
- Elektri edasimüük

Ajaloolisteks muutujateks peetakse:

- Elektri hinnad ajas
- Nõudlus ajas

Kui elektri nõudlus on kõrge, siis tuleb energiat juurde toota kasutades sageli kallimaid energiaallikaid, milleks on gaas või nafta. Kui nõudlus normaliseerub, siis peatub vajadus lisaenergiale. Ebastabiilsust lisandub kui tekib probleeme energia edastamisega või elektrivõrguga. See omakorda võib korporatsioonidele väga kulukaks osutuda ning seetõttu on head ennustusmudelid hinnas. Kui ei suudeta osta ja müüa õigete hindadega, siis võib ees oodata suur rahaline kaotus [3].

1.3 Tarbijate harjumused

Selleks et saada aru, millal ja kuidas elektrit kasutatakse, tuleks uurida tarbijate tavasid ja harjumusi. Samuti on oluline teada, kuidas harjumused on ajas muutunud. See vähendaks ja väldiks energia ületootmist.

Kodust energiatarbimist kaardistavaid faktoreid on seinast-seina. Mõnede uuringute kohaselt võib olla 13 sotsiaalmajanduslikku, 12 ehitise spetsiifilist ja 37 majapidamistarvikute põhist faktorit [5].

Kaardistades koduse majapidamise energia tarbimise olulisemaid punkte leiame, et üheks oluliseks parameetriks on inimeste arv majapidamises. Samuti näitavad uuringud, et majapidamise rahaliste sissetulekute summa on tugevas seoses energia tarbimisega. Näiteks 1% kõige suurema sissetulekuga majapidamistest tarbivad keskmiselt 4 korda rohkem kui keskmine elektri kasutaja.

Peale selle on leitud seos majapidamises olevate inimeste vanuse ja energia tarbimise vahel. Kui majapidamises vastutav isik on üle 55 aastat vana või vahemikus 19-50 aastat vana, siis tarbitakse energiat tõenäoliselt vähem kui teised vanuse grupid. Perekonnas olevate laste ja nende vanuste mõju on riigiti erinev. Näiteks Taanis ja Belgias tarbimine väheneb, Portugalis tõuseb.

Üheks olulisemaks seoseks tuuakse elamispinna suurus, mis määrab kui suurt pindala oleks vaja kütta või jahutada. Tihtipeale on ka majapidamistarvete suurus korrelatsioonis elamispinna suurusega. Lisaks on leitud seos ka majapidamise vanusega. Nimelt, olenemata sellest, et 2000. aastast tänase päevani on majad ehitatud 30% suuremad, on tegelik energia tarbimine nendes majapidamistes vaid 2% suurem vanematest majadest. Uute tehnoloogiate kasutuselevõtt on parandanud nii insuleerimist ja erinevate tarvikute nagu valgustite ja konditsioneeride efektiivsust. Lisaks peetakse kortermaju energiasäästlikutemaks eramajadest.

Majapidamistarvikute kasutusest sõltub samuti suur osa tarbimisest. Energia tarbimine sõltub kas ja kui palju erinevaid tarvikuid majapidamises leidub. Olgu selleks kodumasinad, meelelahutustarbed või isegi elektrisõidukid. Lisaks peab arvestama, kui kaua ja kui tihti neid kasutatakse. Peale selle peab arvesse võtma masinate efektiivsust ja energia kasutust. Teatavasti uued kodumasinad on valmistatud vastavalt uutele normidele ning seetõttu on oodata järjest enam energia

säästlikumaid esemeid.

Aastatega on muutunud inimesed järjest teadlikumaks energia tarbimise vähendamisega ning rohelise energia eelistamisega. Energia tarbimist on proovitud samuti mõjutada erinevate sotsiaalsete meetoditega, mis võib tulevikus järjest enam tavaks olla. Üks nendest meetoditest on naabrite keskmise elektritarbimise näitamine, mis vähendab uuringute kohaselt tarbimist 1.5-3.5%. Kuna energia tarbimine on inimesele nähtamatu, siis on kasutatud lahendusi, mis visualiseeriks tarbimist ja antakse tarbijale teada keskkonnapõhiseid andmeid, mis vähendas tarbimist kuni 8% [6].

Elektri tarbimise seoste loomiseks on samuti kasutatud närvivõrke. Närvivõrgud kinnitavad ka eelnevalt leidnud seoseid, et leibkonna sissetulekust, maja tüübist, omanikust, kodu suurusest, täiskasvanute ja laste arvust kodudes on energia tarbimine tugevas seoses. Lisaks kasvas energia kasutus tehnika vananemisega [7].

Ennustusmudel

Antud töö käigus on plaanis luua ennustusmudelid, mis võtavad lisaks hinna kõikumisele arvesse ka välised tegurid, mis otseselt on seotud hinna väljakujunemisel. Need tegurid on välja toodud töö esimeses osas ning olulisemad nendest võetakse kasutusele mudelites endas. Töö teises osas toon välja andmed, nende allikad, kuidas andmeid töödeldi, erinevad ennustamise metoodikad ning kuidas töös kasutatavad mudelid toimima hakkavad.

2.1 Ennustamise liigid

Probleemide lahendamisel on erinevates valdkondades kasutusele võetud mitmekülgseid lahendusi, sama kehtib ka elektrihinna ennustamisel. Üldises pildis jagatakse meetodid kuueks: [3]

1. tootmiskulude mudelid
2. multi-agent/tasakaalu/mängu teooria mudelid
3. fundamentaal-/struktuuri mudelid
4. kvantitatiivsed/ ökonomeetrilised/stohhastilised mudelid
5. statistilised mudelid
6. tehisintellektil põhinevad mudelid.

Multi-agent/tasakaalu/mängu teooria mudelid loovad agendid, mis simuleerivad süsteemi ja selle toimimist. Nad suhtlevad omavahel ning sellega loovad pakkumise ja nõudluse. Seeläbi luuakse turg, mis simuleerib tuleviku olukorda. Igale agendile antakse kindlad reeglid, millega langetatakse strateegilisi otsuseid. Samuti võetakse arvesse konkurentide eelmiseid pakkumisi. Taoliste mudelite eeliseks on nende paindlikus, et kasutada erinevaid strateegiaid, samas peavad strateegiates kasutatavad eeldused olema õigustatud [8].

Fundamentaal- ja struktuuri mudelite puhul proovitakse arvesse võtta füüsiliste ja majanduslike parameetrite omavahelisi seoseid, mis on energia müümisel ja tootmisel olulised. Neid seoseid (näiteks ilma andmed, tarbimine ja muu) modelleeritakse ja ennustatakse iseseisvalt tihtipeale statistiliste, tehisintellekti või muude meetoditega. Sageli kutsutakse taoliseid lähenemisi ka hübriidideks, sest paljud regressiooni ja närvivõrkude mudelid võtavad taolisi fundamentaalseid andmeid sisenditeks.

Kvantitatiivsete- ja stohhastiliste mudelite põhiline eesmärk ei ole pakkuda kõige täpsemat tunni-põhist elektrihinda, vaid jäljendada hinna kujunemist ja korrelatsioone hinna vahel. Kui ei valita õiget hinna kujunemise protsessi, siis tõenäoliselt ei leita õigeid elektrihinna kujunemise faktoreid ning seega võib olla antud mudeli anda ebausaldusväärseid tulemusi.

Statistiliste mudelite põhjal on tegemist matemaatiliste mudelitega, mis

kasutavad eelnevaid hindasid ning faktoreid, mis võivad olla otseselt seotud hinna kujunemisel. Olgu selleks ilmastiku, tootmise või muud andmed, mis on hinnaga seoses. Üldises pildis jaotuvad statistilised mudelid kaheks. Multiplikatiivne mudel (*multiplicative model*) ja aditiivne mudel (*additive model*). Antud mudelid on tihedalt seotud ning on võimalik teisendada ühest mudelist teise. Statistiliste mudelite eeliseks teiste mudelite ees on, et nad on võrdlemisi arusaadavad ning kerge kasutada. Kriitika kohtadeks on nendel mudelitel tihti mainitud võrdlemisi piiratud võimalused modelleerida mittelineaarset elektrihinna ja seotud fundamentaalsete faktorite kujunemist [2].

Tehisintellektil põhinevaid mudeleid on väga palju erinevaid ning neid kasutatakse tavaliselt olukordades, kus tihtipeale võivad traditsioonilised statistilised mudelid hätta jääda. Tehisintellektil põhinevad mudelid kasutavad õppimise, evolutsiooni ja hägususe elemente, mis võimaldavad neil lahendada ka väga keerulisi probleeme. Antud gruppi kuuluvad paljude teiste hulgas näiteks tehisnärvivõrgud, hägusad süsteemid, *Support vector machine* ehk SVM, evolutsioonilised süsteemid ja kollektiivne intelligents näiteks *random forest*-i puhul.

Tehisintellektil põhinevad mudelid on hinnatud raskete ja tihtipeale mittelineaarsete ülesannete lahendamisel. Antud mudelid on väga head modelleerimaks elektrihinda mõjutavaid faktoreid, kuid nendel esinevaid ka nõrkusi. Mittelineaarsele hüplikule käitumisele kohandumine ei tähenda alati paremaid ennustamistulemusi. Lisaks on tehisintellektil põhinevaid mudeleid väga palju ning nende seast parima valimine on võrdlemisi keeruline. Sellest hoolimata on sama mudeli põhjal võimalik leida parim võrreldes sisendparameetreid ja kalibreerimist väljundiga [2].

2.2 Mudelite täpsuse hindamine

Mudeleid võib hinnata mitmete parameetrite alusel. Tegemist on parameetritega, mis hindavad mingil määral mudeli ennustuse ja tegeliku hinna erinevust üksteisest.

Töös kasutame mudelite omavahelises võrdluses 6 erinevat vea arvestust:

Mean absolute error (MAE) - Keskmine vea suurus üle terve ennustushulga. Ei võeta arvesse vea suunda ning arvutatakse valemiga [9]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \mu_i|, \tag{2.1}$$

kus y_i puhul on tegemist tegeliku väärutusega, μ_i on ennustatud väärtus ning n on andmehulga suurus.

Mean squared error (MSE) - Kasutatakse suurte andmehulkade puhul adekvaatse veasuuruse hindamisel [10]. Arvutatakse valemiga:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_i)^2. \tag{2.2}$$

Root mean square error (RMSE) - Toob paremini välja suurte vigade esinemist ennustusmudelis. Arvutatakse valemiga:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_i)^2}. \tag{2.3}$$

Mean absolute percentage error (MAPE) - Annab mudelile protsendilise vea hinnangu. Protsendiline hinnang on tihtipeale arusaadavam kui ei ole ennustatavate andmetega kursis. Arvutatakse valemiga [11]:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \mu_i}{y_i}\right|. \tag{2.4}$$

Maximum Error - Maksimaalne viga, kui palju kõige enam mudel eksis.

Minimum Error - Minimaalne viga, kui täpselt tegelikkusele mudel ennustas.

2.3 Olemasolevad mudelid ja nende tulemused

Ennustuse mudeleid on loodud väga paljudel erinevatel eesmärkidel ja moodustel. Ennustamise metoodikad on välja toodud antud töö ennustamise liikide osas. Nende kasutust võib leida nii käekirja, kõne ja teksti tuvastamisel kui ka statitiliste andmete, ilma, aktsiate, hindade ja paljude teiste andmete ennustamisel.

Elektrihinna ennustusi on tehtud paljudes riikides ja erinevate meetoditega. Näiteks allikas [12] uuriti ning prooviti ennustada mandri Hispaania ja Kalifornia elektrihindu kasutades kolmekihilist *feedforward* närvivõrku, mis on treenitud *Levenberg-Marquardt* algoritmi abil. Antud töös andis mudel Kalifornia puhul tulemuseks keskmiselt *Mean absolute percentage error (MAPE)* 3% ning Hispaania puhul keskmiselt ligi 9%, sügisel 13%.

Teise töö [13] puhul on võetud uurimisobjektiks taaskord Hispaania turg, mille puhul on loodud mitmeid mudeleid alustates ARIMA-st lõpetades erinevate närvivõrkudega. Tulemuseks loodi mudelid, mille MAPE veaprotsent varieerus 5-10%. Parimaks mudeliks selles uuringus osutus *Hybrid PSO-ANFIS* (HPA).

Kolmanda töö [14] puhul on uuritud Taani tuule ja Nordpooli andmeid, kus proovitakse ennustada tunniseid elektrihindu kasutades ARIMA ja SARIMA meetodeid. Mudeli MAPE vea protsent varieerus 8-11% vahel.

Elektrihinna puhul on oluline uurida kus piirkonnas hinda ennustatakse. Kui antud asukohas on stabiilne kliima, siis on tõenäoliselt hinnaennustus palju täpsem kui mõnes teises piirkonnas. Samuti milliste andmete põhjal ennustus on tehtud ja kui täpsed ning kvaliteetsed andmed on.

Mudeleid on loodud väga palju ja väga erinevatel alustel. Järelduste tegemiseks võtan aluseks põhiliselt kolm eelnevalt mainitud tööd. Eelneva puhul võib teha järelduse, et head mudelid võivad täpsusega varieeruda vastavalt andmetele, asukohale ja mudelile, mida kasutatakse. Olenevalt olukorrast varieerub mudelite täpsus keskmiselt 5% ja 12% vahel. Muidugi on mudel parem, mida väiksem on veaprotsent, sellest hoolimata annavad 12%-ga mudelid vägagi arvestatava ennustuse.

2.4 Töös kasutatavad mudelid

Eelmistes peatükkides välja toodud ennustamise meetodid on vägagi erinevad. Antud lõputöö eesmärgiks on välja töötada mudel kasutades statistiliste mudelite ja tehisintellekti võimalusi. Töös võtan käsile lineaarregressiooni ja *Recurrent neural network*-i erilahendi *Long short term memory*. Antud peatükis tutvustan kasutatavaid mudeleid ning kuidas nad töötavad.

2.4.1 Lineaarregressioon

Lineaarregressiooni kasutatakse lineaarsete sõltuvuste leidmiseks sõltuva muutuja Y ja iseseisvate muutujate X vahel.

Sõltuv muutuja Y peab olema pidev, seevastu iseseisvad muutujad võivad olla pidevad, binaarsed või näiteks kategoorilised. Lineaarregressiooni kasutamine on mõistlik vaid, siis kui muutujate seosed on tõepoolest lineaarsed. Selleks tehakse *scatter plot*-e, mis näitavad, kas seosed on lineaarsed või mittelineaarsed.

Ühemuutujaline lineaarregressioon uurib seost kahe muutuja vahel, milleks on sõltuv muutuja Y ja iseseisev muutuja X. Lineaarregressiooni mudel kirjeldab antud seost järgnevalt:

$$Y = a + bX, \tag{2.5}$$

kus a on Y lõikejoon ja b on kalle. Algselt arvutatakse a ja b väärtused vastavalt Y ja X väärtustele kasutades statistilisi meetodeid. Tekkinud regressiooni joone abil on meil võimalik ennustada sõltuvat väärtust Y iseseisva väärtuse X abil. Seosed on kujutatud näitena Joonisel 2.1.

Teades reaalset elu, mõjutab tihtipeale üht muutujat rohkem kui üks tegur. Mitmemuutujalist lineaarregressiooni kasutatakse just taoliste olukordade lahendamiseks kui soovitakse mudelis kasutada rohkem kui ühte iseseisvat muutujat. Mitmemuutujalise lineaarregressiooni seost kirjeldatakse järgnevalt:

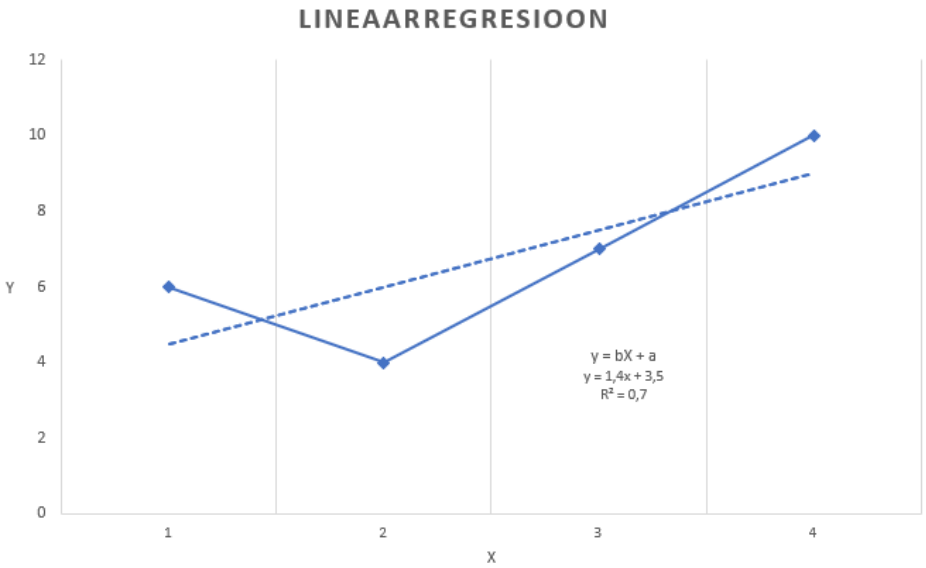
$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n. \tag{2.6}$$

Mudel võimaldab arvutada regressiooni koefitsienti b_i iga iseseisva muutuja X_i puhul. Koefitsient kirjeldab seost sõltuva muutuja Y ja iseseisva muutuja X_i vahel. Oluline on, et mitmed iseseisvate muutujate koefitsiendid ei läheks omavahel segamini.

Koefitsienti r^2 arvutamisel kasutatakse valemit: [15]

$$r^2 = \frac{\sum_{i=1}^{n}(\sigma_i - \mu)^2}{\sum_{i=1}^{n}(y_i - \mu)^2}. \tag{2.7}$$

- n on andmete hulk
- σ_i on eeldatud sõltuva muutuja väärtus, mis arvutati regressiooni valemis
- y_i on vaadeldud i-nda sõltuva muutuja väärtus
- μ on kõikide sõltuvate muutujate keskmine väärtus
- r^2 on murdosa dispersioonist. Mida lähemal on regressiooni mudeli ennustatavad väärtused σ_i, seda lähemal on koefitsent ühele ja seda täpsem mudel on.



Joonis 2.1: Lineaarregresioon - eeldatakse, et vaatluste andmetel (märgitud siniste punktidega) esinevad kõrvalekanded üldisest sõltuvusest (märgitud sinise kriipsjoonega) sõltuva muutuja (Y) ja iseseisva muutuja (X) vahel.

Hoolimata paljudest alternatiividest on lineaarregressioon üks kasutatavamaid elektrihinna ennustuse meetodeid. Tihtipeale kasutatakse seda koos keerulisemate meetoditega [2].

2.4.2 *Recurring Neural Network* - LSTM

Ennustamisel on väga populaarseteks vahenditeks kujunenud tehisnärvivõrgud. Kokkuvõtlikult võib närvivõrke kujutada ette kui neuronid, mis on omavahel kihtidena seotud, igal neuronil on oma kaal, mille järgi otsuseid tehakse. Närvivõrku treenides optimeeritakse antud kaalud ning seega oskab võrk tihtipeale väga täpseid ennustusi teha. Närvivõrke on väga erinevaid, näiteks on olemas *feedforward neural network*, mis ei kasuta eelnevatel õppimise iteratsioonidel paika pandud nodeide kaale. Samuti on olemas *recurring neural network*, millel on taolised tagasiside seosed. See tähendab, et mudel võtab arvesse eelmiste õppimise iteratsioonide kaale.

LSTM ehk *Long short term memory* on RNNi erijuht, kuhu on lisatud mitmed komponendid, mis parandavad RNNi nõrkusi. LSTM on osutunud väga kasulikuks mudeliks, mis käsitleb üksteisele kindlas jadas järgnevaid andmeid. Seetõttu teda kasutatakse paljudes keele ja käekirja tuvastuste ning ka aktsiate hindade ennustuste mudelites. LSTMi oluliseks osaks on tema mälu, mis töötab justkui olekute kogujana. Mälusse sisenetakse, kirjutatakse ja kustutatakse vastavalt kindlate "väravate" abiga. Iga kord kui tuleb ühte neuronisse uus sisend, kogutakse see kokku juhul kui sisend värav i_t on aktiveeritud. Kui aktiveeritakse unustamise värav, siis unustatakse eelmise neuroni olek c_{t-1}. See, kas üldse lastakse väärtus väljundisse on väljundi värava poolt kontrollida. [16, 17, 18]

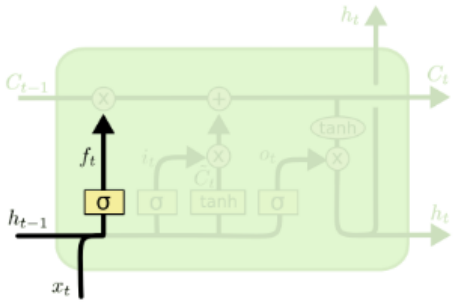
LSTM on hea seepoolest, et informatsiooni hulka kontrollitakse mälu ja väravate abil ning olek hoitakse kindlas neuronis. RNNi suureks probleemiks haihtuva gradiendi või oleku nähtus, kus ei hoita olulisi olekuid, mis kaugemas minevikus oli. See on oluline näiteks aktsiate liikumisel või keeletöötlusel, kus esimeses sõnas võis olla väga oluline seos, mida viimase sõna ennustamisel võib vaja minna. LSTMi puhul ei haihtu gradient nii kiiresti ning on abiks taoliste seoste säilitamiseks.

LSTMi esimeseks osaks on unustamise värav (Joonis 2.2), kus otsustatakse mis osa tasub ära visata. Otsus võetakse vastu vastavalt valemile:

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f), \tag{2.8}$$

kus

- h_{t-1} - väljund eelmisest LSTM blokkist
- x_t - sisend praegusesse LSTM blokki
- b_f - *bias* vektor



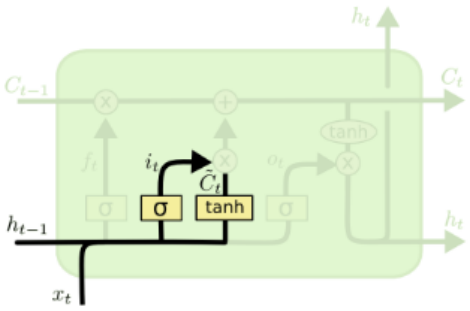
Joonis 2.2: LSTM - *forget gate* [1].

Vastus tuleb 0 ja 1 vahel, kus 0 tähendab, et tasub ära visata ning 1 soovitab alles jätta.

Järgmises osas (Joonis 2.3) vaatab LSTM kas tasub informatsiooni talletada. See protsess on kahes osas, millest esimene on sisend värav i_t ning teises tanh kiht tekitab vektori uute kandidaat väärtustega K_t, mis võidakse lisada olekusse. Valemina on seos kujutatud järgnevalt:

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i), \tag{2.9}$$

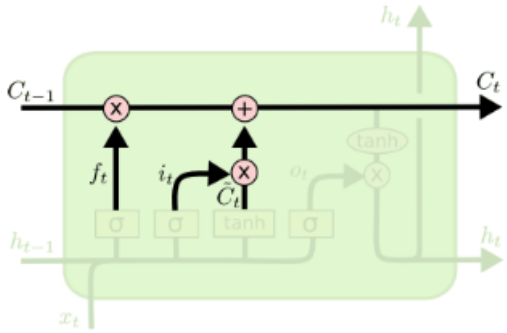
$$K_t = \tanh(W_C \times [h_{t-1}, x_t] + b_C). \tag{2.10}$$



Joonis 2.3: LSTM - *input gate* [1].

Järgmisena (Joonis 2.4) uuendatakse vana olek C_{t-1} uueks olekuks C_t. Selleks võetakse vana olek ning korrutatakse eelnevalt arvutatud f_t-ga ning liidetakse uuendatud kandidaat väärtustega $i_t \times K_t$. Seega uuendamise valemiks saame:

$$C_t = f_t \times C_{t-1} + i_t \times K_t. \tag{2.11}$$

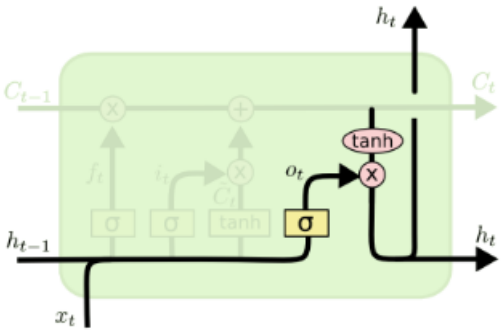


Joonis 2.4: LSTM - *update state* [1].

LSTMi viimaseks väravaks on väljund (Joonis 2.5). Väljund põhineb neuroni olekul, kuid on selle filtreeritud versioon. Algselt läbitakse sigmoid kiht, mis otsustab mis olekus osa saadetakse väljundisse ning seejärel läbib olek tanh fuktsiooni, mis liigutab väärtused -1 ja 1 vahele. Viimaks korrutatakse väljundvärava väärtustega. Valemina on seos kujutatud järgnevalt: [19, 1]

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o), \tag{2.12}$$

$$h_t = o_t \times \tanh(C_t). \tag{2.13}$$



Joonis 2.5: LSTM - *output gate* [1].

2.5 Lähteandmed

Esimeses osas tõin välja hinda mõjutavad erinevad parameetrid. Nendest kõige enam efekti ja kaalu omavad fundamentaalsed parameetrid ja ajaloolised parameetrid. Samuti on just fundamentaalsed ja hinna ajaloolised andmed antud töö kontekstis kõige kättesaadavamad.

Teadvustan, et mudeli tulemused on tihedas sõltuvuses kasutatavatest andmetest, mis on tihtipeale loodud samuti mudeleid kasutades ning ei ole 100% täpsed. Heaks näiteks on ilmaennustuse andmed. Seetõttu võib tekkida anomaalia isegi, siis kui elektriennustusmudel ennustab täpselt. Operatsioonilised ja strateegilised andmed on elektritootjate ja edastajate salajased andmed, mis ei ole tihtipeale kättesaadavad.

Fundamentaalsetest andmetest võtan antud töös uurimise alla ilmastiku andmed Eesti, Läti, Leedu, Soome, Rootsi ja Norra lennujaamadest ajavahemikus 1. jaanuar 2018 kuni 31. märts 2020. Ilmastiku andmetest kõige olulisemateks võib esialgsel hinnangul pidada temperatuuri (C), tuule kiirust(m/s), pilvisust(%) ning sademete hulka (mm). Ilmastiku andmed pärinevad https://rp5.ru/ andmebaasist.

Lisaks võtsin vaatluse alla nafta hinna ning selle seose elektrihinna kujunemisega Eestis. Kasutatavad andmed pärinevad https://www.macrotrends.net/ andmebaasist ning on ajavahemikus 1.jaanuar 2018 kuni 31. märts 2020.

Elektrihinna ajaloolised andmed pärinevad https://www.nordpoolgroup.com/ andmebaasist. Vaadeldavateks andmeteks valisin päevase elektrihinna (EUR/MWh), elektritootmise (MWh), tarbimise (MWh), ostu/müügi (MWh) ning tuule energia (MWh) suurust Eestis, Lätis, Leedus, Soomes, Rootsis ja Norras. Vaadeldavaks ajaperioodiks valisin 1.jaanuar 2018 kuni 31.märts 2020 Nordpoolist saadavate andmete piirangute tõttu.

2.5.1 Lähteandmete töötlemine

Nordpoolist kättesaadavad andmed olid üldjoones väga heas seisus. Andmed olid täielikud ning ei olnud vaja neid palju töödelda.

Samas olid ilmastiku andmed riigiti erinevad. Üldjoones oli tegemist tunniste ning mõnel juhul ka poole tunniste andmetega, mis tuli teisendada päeva keskmiseks andmeteks. Lisaks ei olnud kõik andmed koheselt kasutataval kujul. Heaks näiteks oli mõne ilmajaama pilvisuse andmed, kus oli ühte tabeliritta järjestikku seatud nii pilvede umbkaudne kõrgus, nende tüüp ning pilvisuse protsendi vahemik. See eeldas viimase 27 kuu tuhandete tunniste andmete sobivale kujule teisendamist. Pilvisuse osas oli oluline saada päeva keskmine protsent. Sademete puhul oli samuti segamini nii tekst kui ka arvväärtused, mis pidi teisendama töödeldavale kujule.

Samuti ei olnud kõikide riikide andmed täielikud. Olenevalt puudu olevate andmete suurusest ei võtnud autor neid kasutusse või täitis väljad keskmiste väärtustega. Probleemid esinesid põhiliselt Läti ilmastiku andmetes.

2.6 Korrelatsiooni leidmine

Selleks, et aru saada kas kasutatavatel andmetel on tõepoolest korrelatsioon Eesti elektrihinna väljakujunemisega on vaja kasutada metoodikat, mis annaks meile paremat ülevaadet. Antud töös kasutan korrelatsiooni maatriksi, mis näitab kui tihedas seoses on andmed. Maatriks on skaalal 1 kuni -1, mis tähendab, et väga tugevas seoses andmed on väga lähedal numbrile 1 ja väga tugevas vastupidises korrelatsioonis andmed on lähedal -1-le. Andmed, mis on lähedal 0-le on väga väheses seoses üksteisele.

2.6.1 Nordpooli andmete korrelatsioon

Algselt võtame käsile Nordpooli spetsiifilised andmed lisades andmetele juurde ka nafta hinna. Vaatleme Eesti (EE), Läti (LV), Leedu (LT), Soome (FI), Rootsi (SE), Norra (NO) ja Taani (DK):

- elektri hindu
- tootmise suurust (märgitud tabelis lahtris *"prod"*)
- tarbimist (märgitud tabelis lahtris *"consumpt"*)
- ostu/müüki (märgitud tabelis lahtris *"net exchange"*)
- tuule energiat (märgitud tabelis lahtris *"wind"*).

Rootsi ja Taani elektrituru osas on oluline mainida, et nende riikide turud on samuti jaotatud mitmeteks osadeks. Rootsi koosneb SE1, SE2, SE3 ja SE4, millest kõige enam võtan vaatluse alla SE4, mis korreleerub Eesti elektrihinnaga kõige paremini. Taani koosneb DK1 ja DK2, millest võtan mõlemad vaatluse alla.

Alustame analüüsimist esimesest korrelatsiooni maatriksist (Joonis 2.6). Selgelt joonistuvad välja Eesti elektrihinna tugev seos Soome, Läti ja Leedu elektrihinnaga, mis on tugevas korrelatsioonis (0.95). Samuti on märgata võrdlemisi tugevat seost Soome ja Leedu *net exchange*-iga, mis on mõlemad 0.42.

Lõpetuseks oleks mõistlik ka välja tuua mõningase korrelatsiooniga parameetrid. Nendeks on Eesti tootmine, ost/müük ja tuuleenergia suurus - vastavalt 0.39, -0.39 ja -0.38. Peale nende siseriiklike parameetrite on samuti mõistlik välja tuua Läti

ja Leedu tuuleenergia – vastavalt -0.39 ja -0.38. Teiste parameetrite mõju hinnale on pigem väike või olematu.

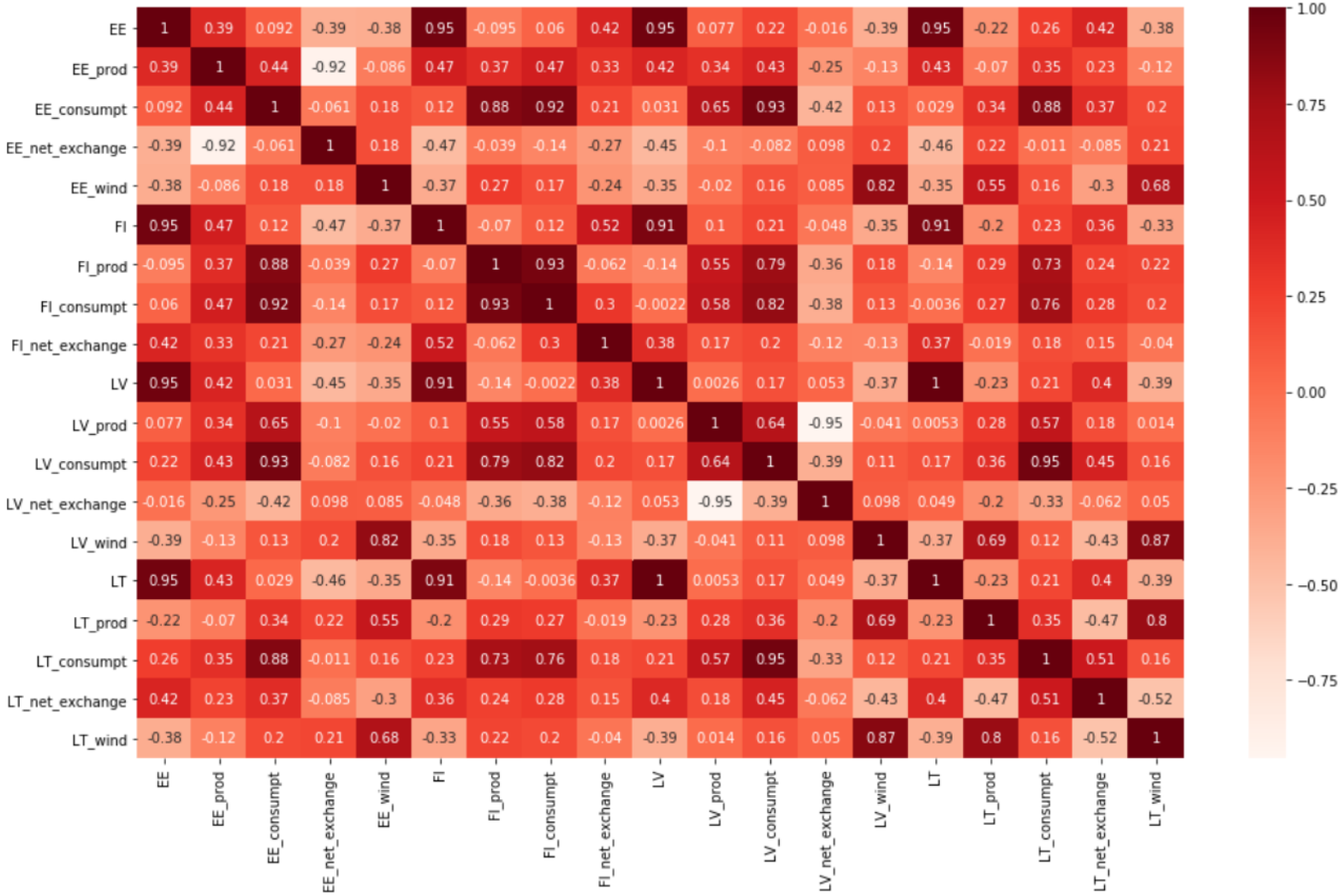
Analüüsides teist maatriksit (Joonis 2.7) leiame kõige rohkem mõjutavamateks parameetrikeks taas hinna. Kõige enam mõjutavad Eesti elektrihinda Roosti (SE4), Taani (DK1 ja DK2) ja Norra elektrihinnad – vastavalt 0.83, 0.77, 0.8 ja 0.73. Lisaks on taas olulisteks parameetriteks nende riikide ost/müük – vastavalt -0.48 Rootsi, -0.36 Taani ja -0.46 Norra.

Viimaseks parameetriks on samuti ka nafta hind, mis huvitaval kombel Eesti elektrihinnaga pigem ei korreleeru. Samas korreleerub võrdlemisi hästi Norra, Rootsi ja Taani hindadega. Järeldan, et maailmaturu nafta hind kajastub Eesti elektrihinnas läbi teiste riikide elektrihindade.

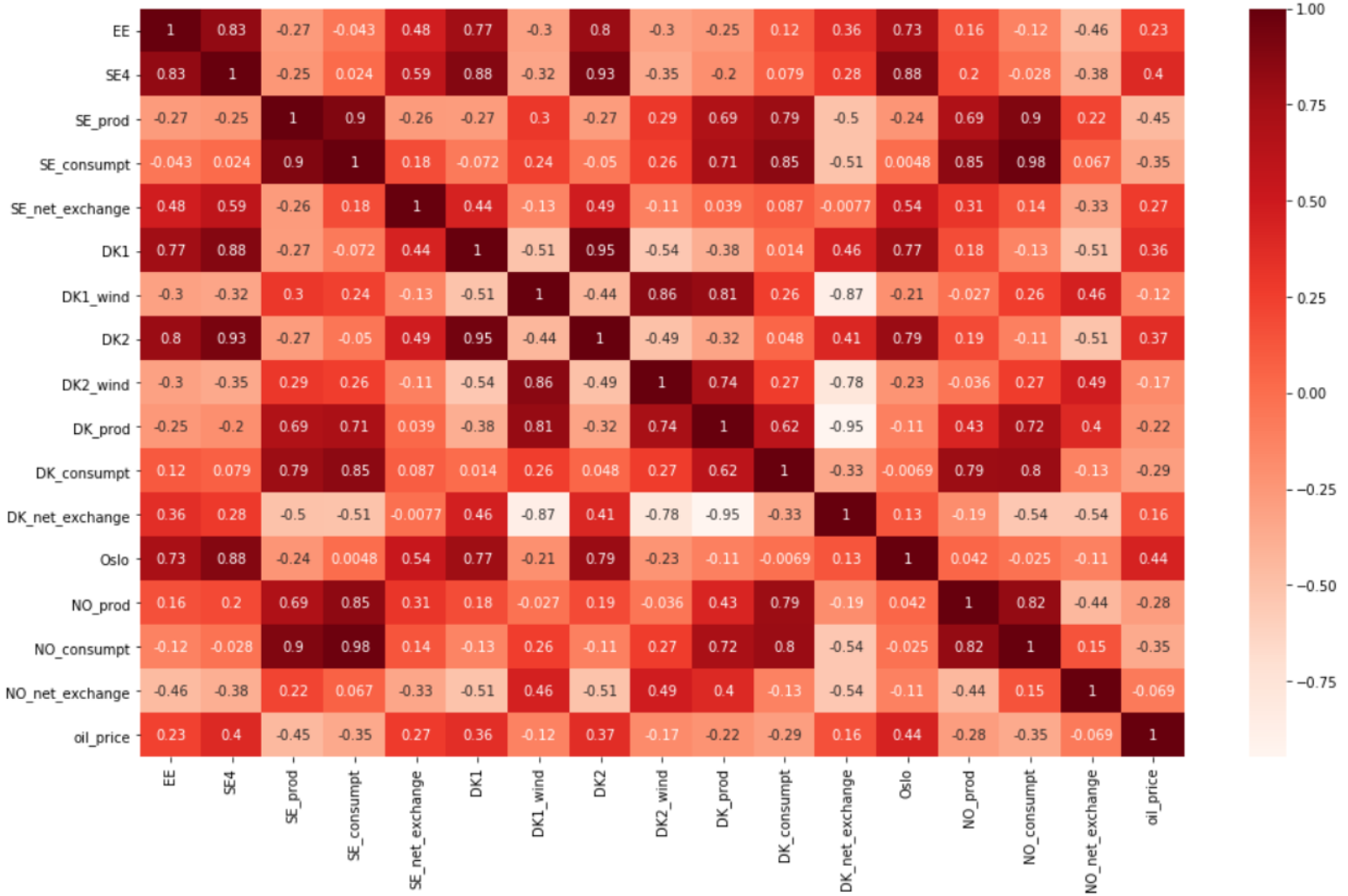
Kui antud töö eesmärgiks oleks ennustada Rootsi elektrihinda, siis oleks parameetrid vägagi sarnased – on näha, et suure korrelatsiooni moodustavad taas hinnad ning ost/müük. Lisafaktoriks on samuti Taani tuuleenergia ning nafta maailmaturu hinnad.

Kui aga vaadata Läti ja Leedu hinda mõjutavaid faktoreid, siis lisanduvad teiste riikide hinna ja ostu/müügi parameetrite kõrvale ka Eesti elektri tootmine ning nii Eesti, Läti kui ka Leedu tuulenergia tootmise suurus.

Kolmandas Nordpooli korrelatsiooni maatriksis (Joonis 2.8) võrdlesin kõikide Nordpooli piirkondade elektrihindade korrelatsiooni üksteise suhtes.

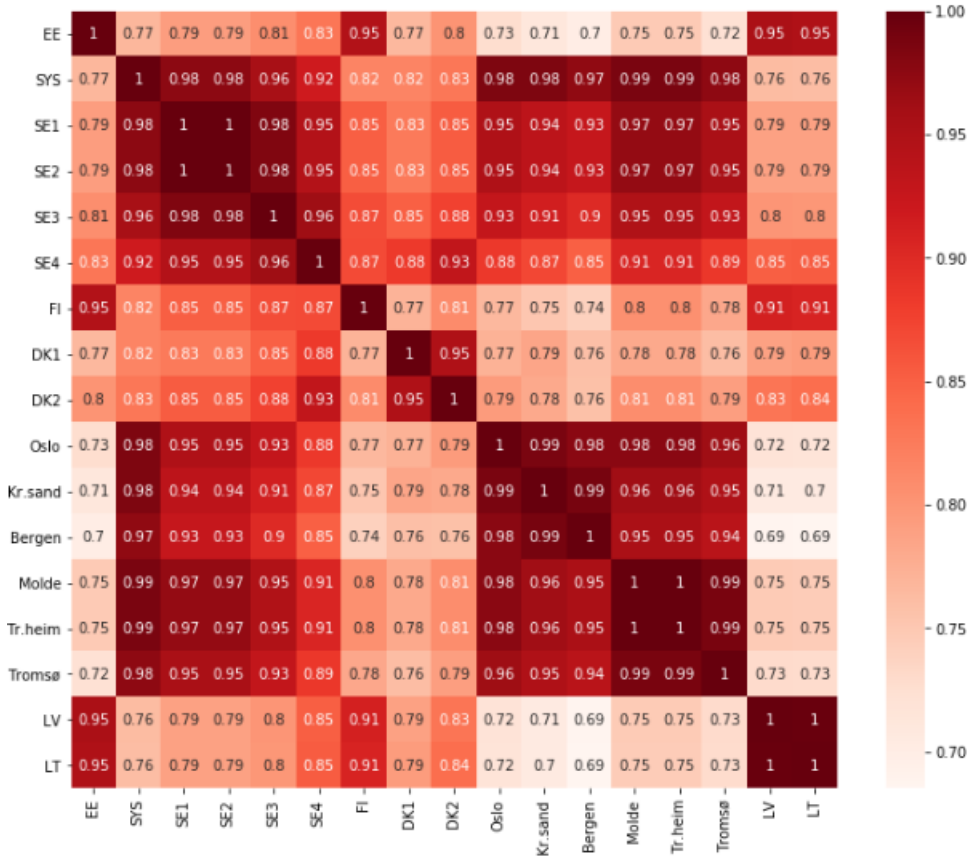


Joonis 2.6: Nordpooli Baltikum ja Soome korrelatsioonimaatriks.



Joonis 2.7: Nordpooli Skandinaavia korrelatsioonimaatriks.

Maatriksilt on näha, et Eesti elektrihinnaga korreleeruvad kõige paremini Soome, Läti ja Leedu elektrihinnad (0.95). Samas korreleerub ka väga hästi Rootsi SE4 ning Taani DK2, mis on vastavalt 0.83 ja 0.80.

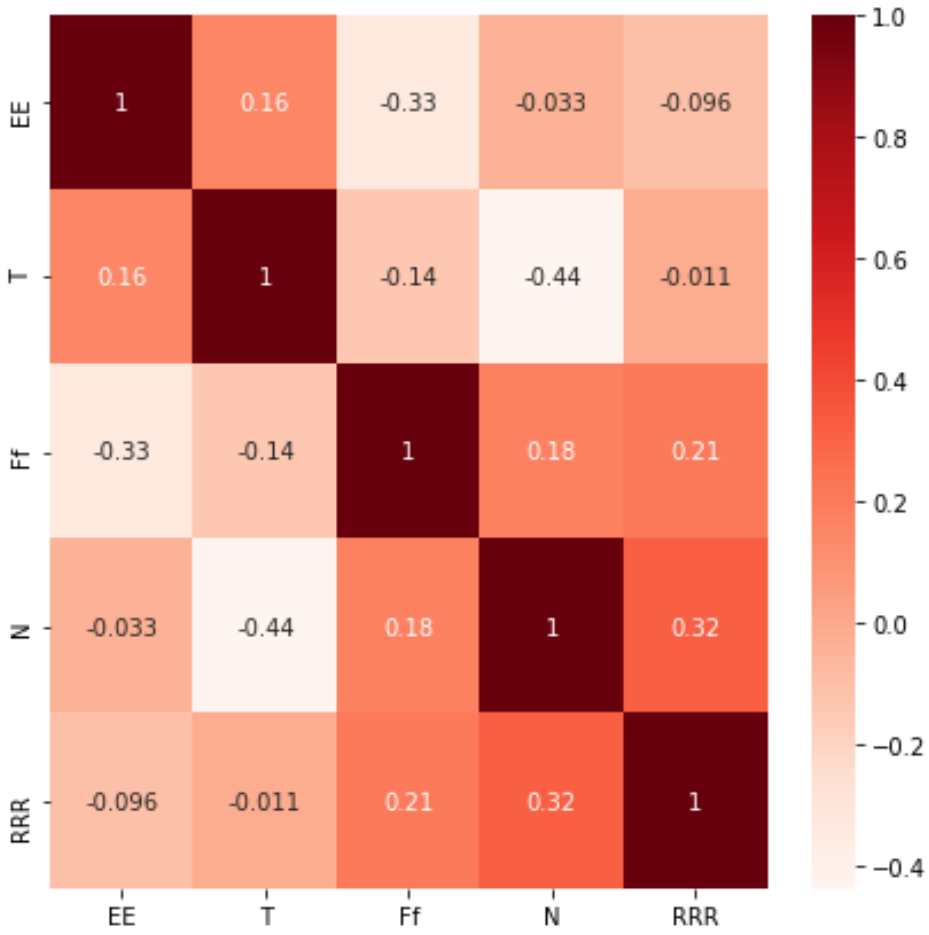


Joonis 2.8: Nordpooli hindade korrelatsioonimaatriks.

2.6.2 Ilmastiku andmete korrelatsioon

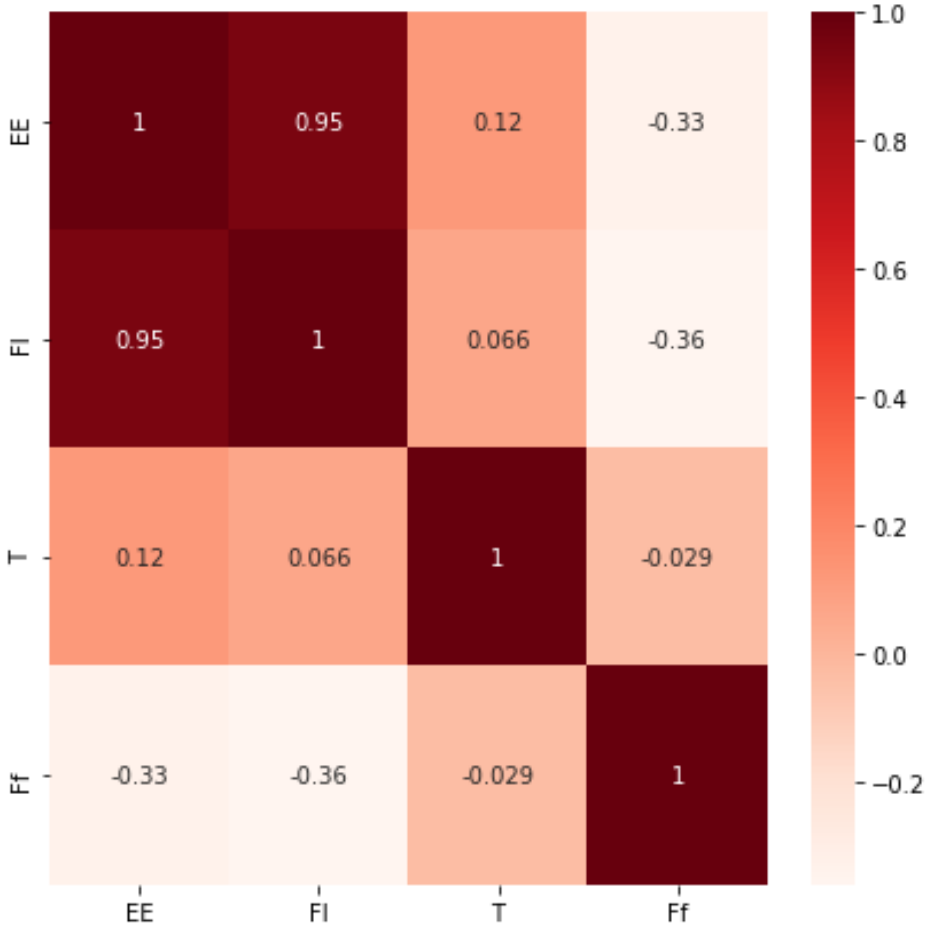
Järgnevalt vaatleme nii Eesti kui ka teiste vaadeldavate riikide ilmastikuandmeid ning uurime, millist mõju avaldavad erinevate riikide õhutemperatuur (T), tuule puhangud (Ff), pilvisus (N) ja sademete hulk (RRR). Lisaks lisasin samasse maatriksi parema võrdluse pärast vastava riigi elektrihinna ning Eesti elektrihinna. Ilmastiku andmete puhul tuleb arvesse võtta andmete osalist puudumist Läti andmete puhul. Kui ülejäänud andmed olid ligi 100% täielikud, siis Läti puhul oli puudu ligi 15% päevaseid andmeid, mille seas oli ligi 80% täielikke tuule, 50% pilvisuse ning 98% temperatuuri andmeid.

Alustades Eesti ilmastikuandmetest (Joonis 2.9) selgub oodatust palju väiksem korrelatsioon õhutemperatuuri ja elektrihinna vahel (0.16). Kõige enam mõjutab hinda tuul (−0.33).



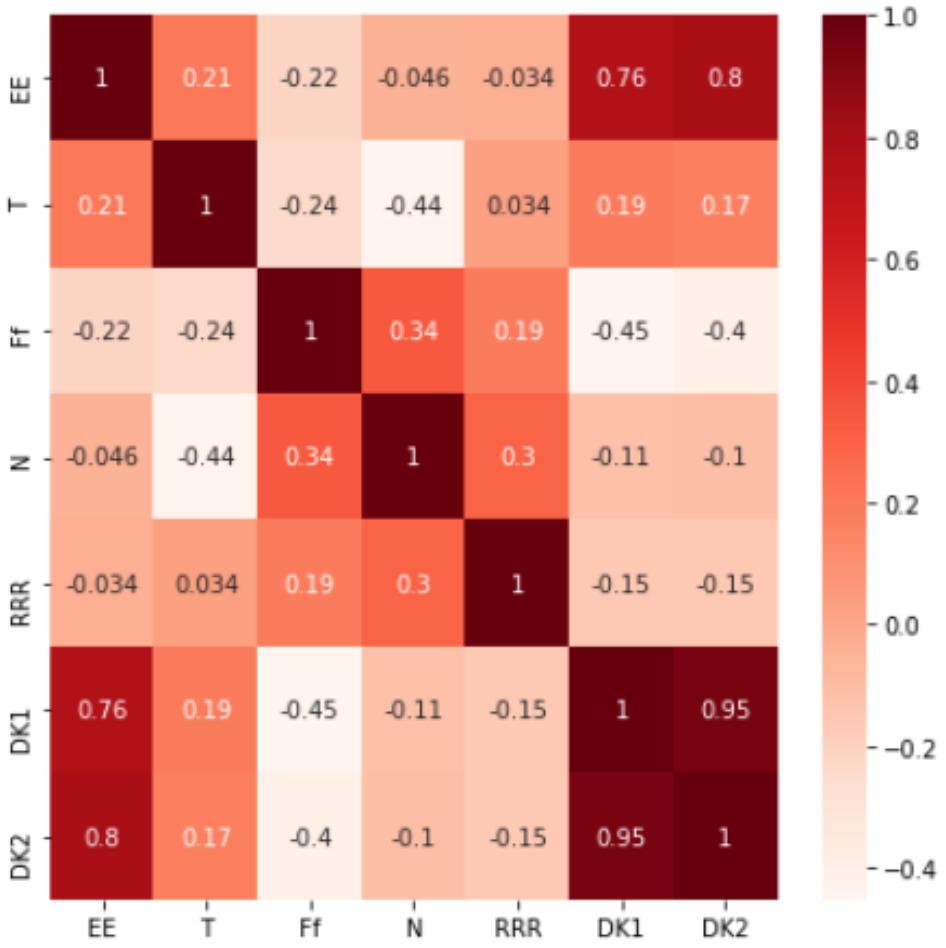
Joonis 2.9: Eesti ilmastiku korrelatsioonimaatriks.

Soome ilmastiku andmetest (Joonis 2.10) oli võimalik võrrelda tuule ning temperatuuri sõltuvust. Jällegi ei ole temperatuur määrav faktor. Kõige enam korreleerub tuul (-0.33).



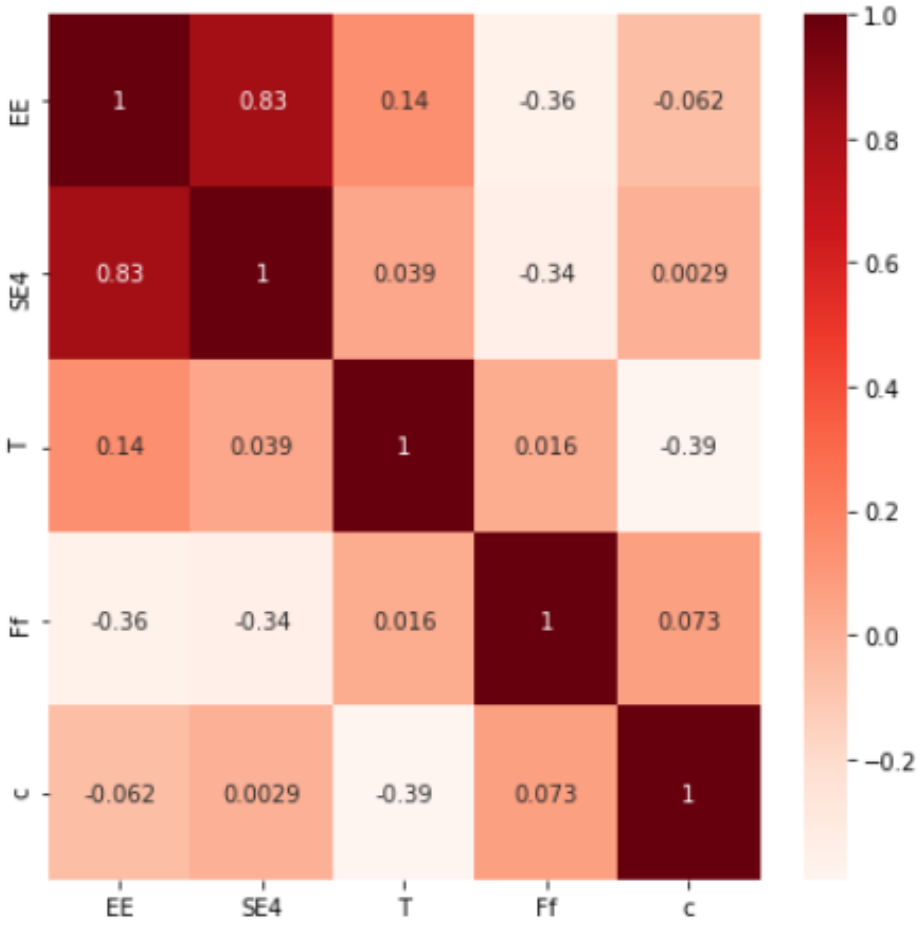
Joonis 2.10: Soome ilmastiku korrelatsioonimaatriks.

Taani ilmastik (Joonis 2.11) ei mõjuta samuti Eesti elektrihindu. Lähim korrelatsioon on taas tuul (-0.22). Kui võrrelda Taani enda elektrihinnaga, siis korreleerub tuul kaks korda rohkem. Tuule puhul on tegemist Taani ühe suurima energiaallikaga. Temperatuur on samuti väikese mõjufaktoriga.



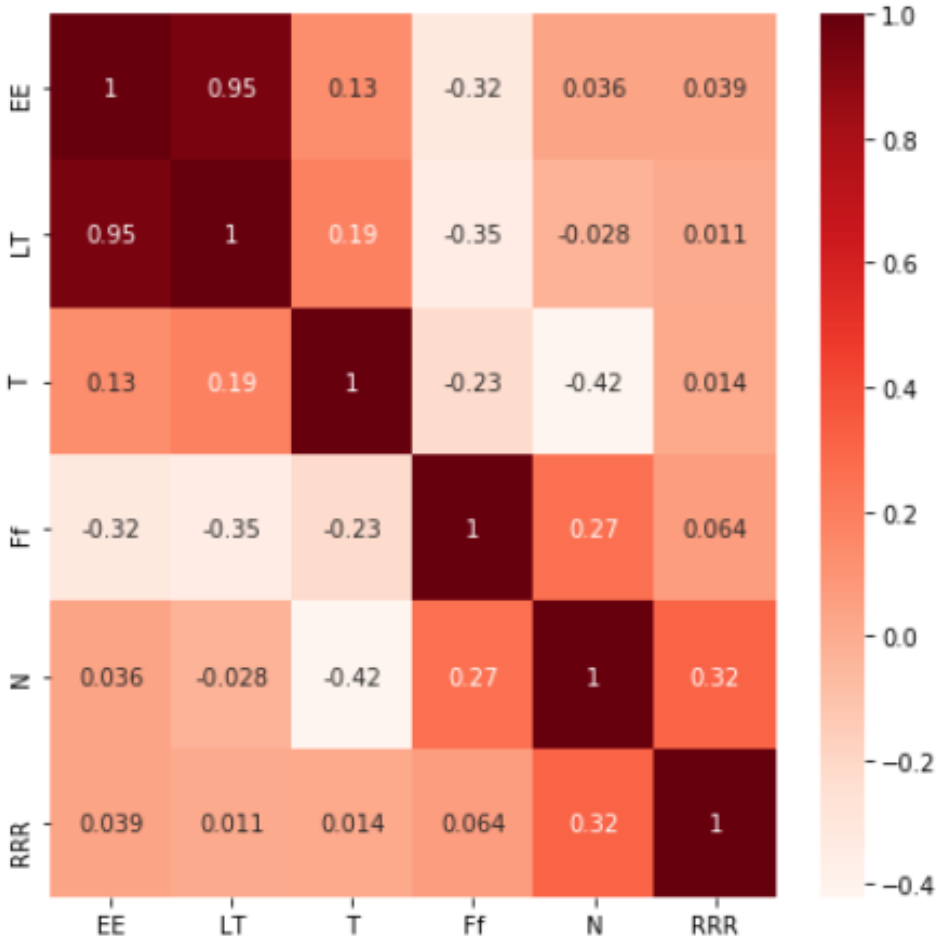
Joonis 2.11: Taani ilmastiku korrelatsioonimaatriks.

Rootsi puhul on põhiliseks korrelatsiooniks tuul (−0.36). Pilvisus ja temperatuur mängib ilma kujunemisel väikest rolli (Joonis 2.12).



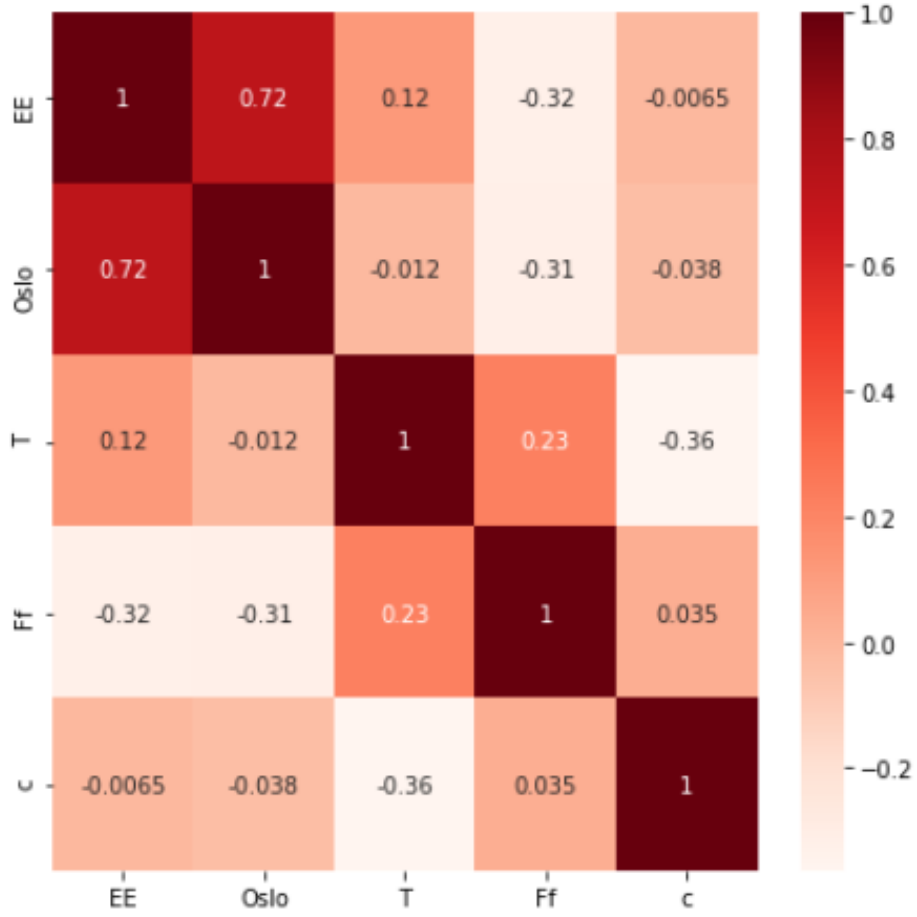
Joonis 2.12: Rootsi ilmastiku korrelatsioonimaatriks.

Leedu ilmastikus (Joonis 2.13) on näha sama trendi kui eelnevates riikides. Kõige enam mõjutab tuul (-0.32). Nii sademed, temperatuur kui ka pilvisus ei korreleeru.



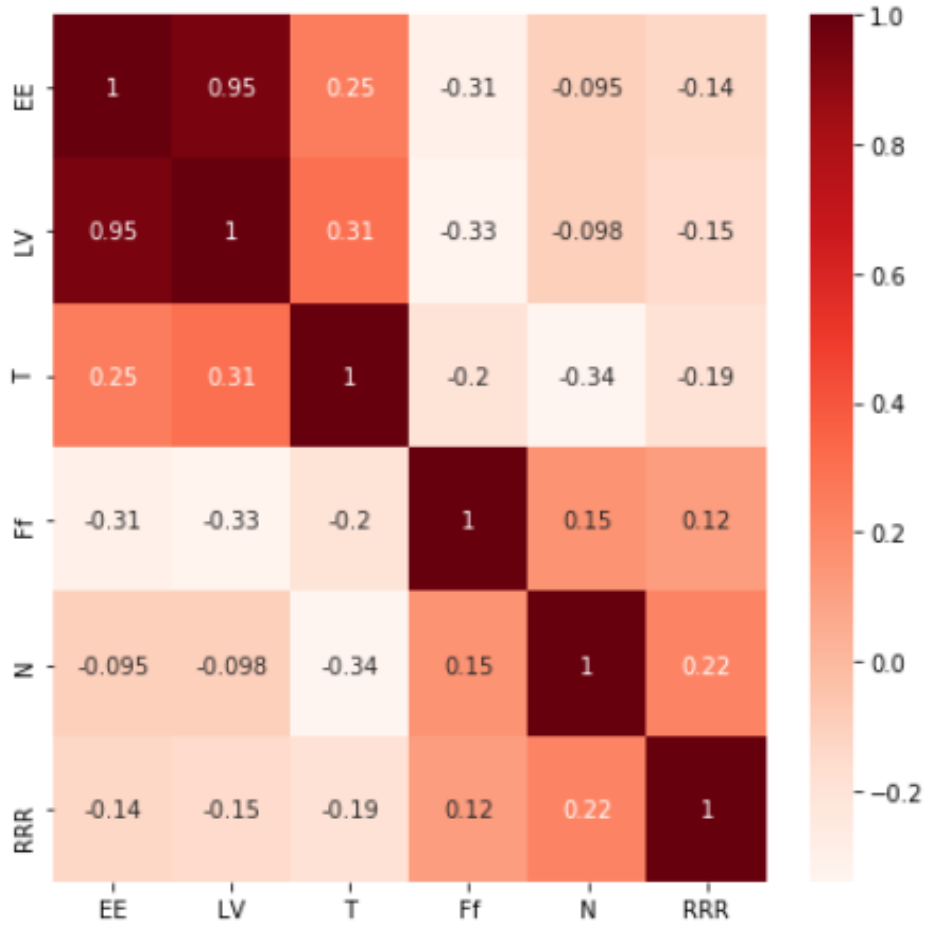
Joonis 2.13: Leedu ilmastiku korrelatsioonimaatriks.

Norra kliimast (Joonis 2.14) korreleerub kõige enam tuul (-0.32). Teised faktorid ei korreleeru.



Joonis 2.14: Norra ilmastiku korrelatsioonimaatriks.

Läti andmete põhjal (Joonis 2.15) korreleerub kõige enam tuul (−0.31). Temperatuur samas korreleerub eelmiste riikide andmetega võrreldes hästi (0.25), kuigi antud korrelatsioon on siiski suhteliselt väike, et kasutada hinna ennustamisel.



Joonis 2.15: Läti ilmastiku korrelatsioonimaatriks.

2.6.3 Mudelites kasutatavad sisendid

Võttes kokku eelnevate korrelatsiooni tulemused saab teha järeldused, milliseid andmeid tasub töö sisenditena kasutada. Ilmastiku andmetest saab kõige paremini kasutada tuule andmeid. Tuule andmed on samuti kajastatud ka Nordpooli Eesti tuuleenergia tootmise suuruses, mis korrelleeruvad isegi paremini. Nordpooli andmetest saab sisendina suurepäraselt kasutada Soome, Läti, Leedu, Rootsi, Taani ja Norra hinna andmed. Lisaks Eesti, Soome, Läti ja Rootsi ostu/müügi andmeid. Samuti lähevad kasutusse Eesti varasemad elektrihinnad ning Eesti elektrienergia tootmise suurus.

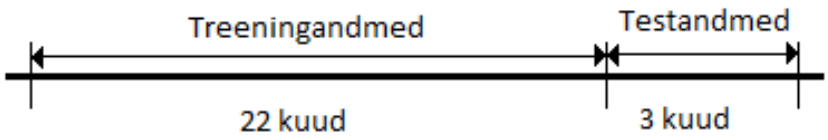
Tabel 2.1: Mudelite sisendid

Sisend	Andmehulk 1	Andmehulk 2
Eesti hind	X	X
Soome hind	X	X
Läti hind	X	X
Leedu hind	X	X
Rootsi hind	X	X
Taani hind	X	X
Norra hind	X	X
Eesti ost/müük	X	-
Soome ost/müük	X	-
Läti ost/müük	X	-
Rootsi ost/müük	X	-
Eesti tuuleenergia	X	-
Eesti elektritootmine	X	-

Lähteandmed jagan kahte suuremasse andmehulka. Esimeses andmehulgas on kajastatud kõik eelnevalt mainitud sisendid. Teises andmehulgas on kajastatud välja valitutest kõige paremini korrelleeruvad andmed ehk eelpool mainitud riikide hindade andmed.

2.7 Mudelite loomine

Töö kirjutamise ajal tuleb mudelite loomisel arvesse võtta maailmas valitsevaid trende ja olukorda. Nafta maailmaturu hindade suur kõikumine ja koroonaviiruse mõju on selgelt mõjutanud elektrihindade mudeleid ja ennustusi. Antud töös uurin lisaks kui palju on viimase paari kuu ennustusi mõjutanud koroonaviiruse ja nafta maailmaturu hinna kukkumine. Selleks kasutan erinevaid andmehulki, et näha ennustuste erinevust. Mudelite koostamisel kasutan 2018.jaanuar-2020.jaanuari kuu andmeid. Nende loomisel jaotan andmed test- ja treeningandmeteks, võttes testandmeteks viimased 3 kuud. Proportsioon on kujutatud Joonisel 2.16. Eriolukorra mõju hindamisel võtan arvesse 2018.jaanuar-2020.märtsi kuu andmeid.



Joonis 2.16: Treening ja test andmete proportsioon.

Mudelite loomisel kasutasin antud töös Pythoni tekstitöötlus tööriista Jupyter, mis töötab Anaconda Navigatori abil. Viimases sai ka määrata keskkonna täpsemaid versioone ning Pythoni mooduleid, mida kasutada. Töös kasutasin Pythoni 3.6.10 versiooni, Jupyter Notebooki 6.0.3 versiooni.

2.7.1 Lineaarregressiooni mudeli loomine

Mudeli loomisel võtsin aluseks Nagesh Singh Chauhan-i poolt loodud koodinäite, mille ma oma vajaduste põhiselt ümber kohandasin [20].

Mudeli töötluseks ja koostamiseks on peamiselt vajalik kasutada *pandas*, *numpy* ning *sklearn* mooduleid. Algselt loeme andmed kasutades *pandas.read_csv* meetodit.

Algselt koostame kaks *dataset*-i, milles ühes (X) on väärtused, mida me kasutame, et ennustada Y väärtust. Väärtuseks võtame kõige parema korrellatsiooniga andmed Eesti elektrihinnale. Lisame samuti Eesti elektrihinna (EE_past), mis on eelneva ajahetke Eesti elektrihind.

```
X = dataset[['FI', 'LV', 'LT', 'SE4', 'DK2', 'Oslo', 'EE_prod', '
    EE_net_exchange', 'EE_wind', 'FI_net_exchange', 'LV_net_exchange
    ', 'SE_net_exchange', 'EE_past']]
Y = dataset['EE']
```

Jagame andmed treening- ning test-andmeteks. Test-andmeteks võtame taas viimased 3 kuud.

```
X_train, X_test, Y_train, Y_test = train_test_split(X.values, Y.
    values, test_size=0.13, shuffle=False)
```

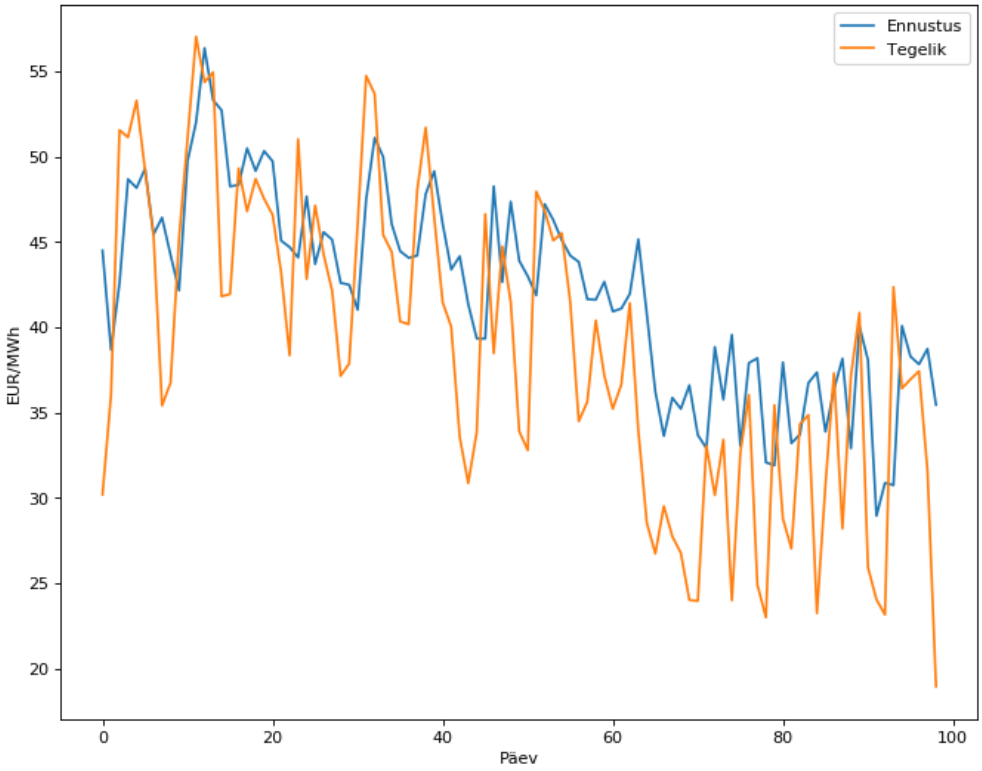
Järgnevalt alustame mudeli treenimist:

```
regressor = LinearRegression()
regressor.fit(X_train, Y_train)
```

Peale mudeli treenimist uurime kas ka mudeli ennustamine on täpne ning testime seda testandmetega.

```
Y_pred = regressor.predict(X_test)
```

Kujutame tulemused graafikule (Joonis 2.17), et oleks visuaalselt paremini jälgitav.



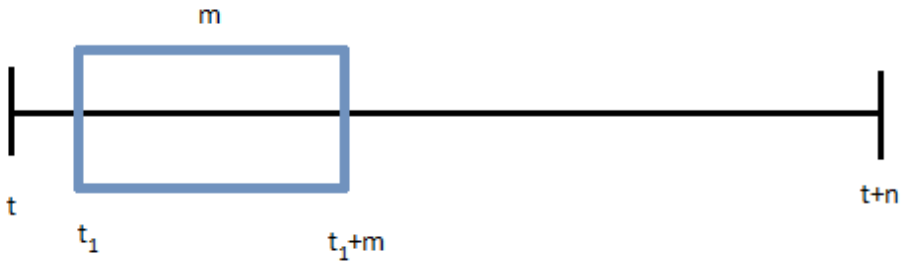
Joonis 2.17: Esialgne lineaarregressiooni graafik.

Arvutame ka *Mean absolute error*-i, *Mean squared error*-i ja *Root mean square error*-i. Sellega näeme terve mudeli vigade suurust.

```
Mean Absolute Error: 5.47066
Mean Squared Error: 46.12075
Root Mean Squared Error: 6.79123
```

2.7.2 Sliding window lineaarregressiooni mudeli loomine

Sliding window ehk liikuva akna meetodit kasutatakse olukordades, kus soovitakse näha kas ja kui suurt mõju avaldab mudelile teatud periood. Vaatluse all olev periood määratakse akna suuruseks. See tähendab, et kui akna suuruseks määratakse 1 kuu, siis kasutatakse järgmise päeva elektri hinna ennustamisel vaid 1 kuu andmeid, mis sellele perioodile eelnes. See omakorda tähendab, et iga järgneva päeva ennustusega liigutatakse akent 1 päeva võrra edasi. Seega tekivad ennustused, mis on tihedamalt seotud perioodiks määratud aja andmetega, mis on tihtipeale lähiminevik.



Joonis 2.18: *Sliding window* implementatsioon.

Eelnevalt toodud Joonisel 2.18 on kirjeldatud liikuva akna meetodit, kus t on andmete esimene ajahetk, n on andmete kogus. Liikuva akna esimest elementi kirjeldab t_1 ning liikuva akna suurust kirjeldab m.

2.7.3 LSTM mudeli loomine

Mudeli loomisel võtsin aluseks Jason Brownlee poolt loodud koodinäite, mille ma oma vajaduste põhiselt ümber kohandasin [21].

Sarnasel eelnevaga loeme väärtused kasutades *pandas* moodulit sisse. Järgnevalt peame kindlad olema oma andmete õigsuses, seega määrame nad *float* formaati.

```
values = values.astype('float32')
```

Selleks, et mudel paremini õpiks ja et andmed ühtlased oleks, tuleb neid normaliseerida. Sellisel juhul jäävad kõik andmete väärtused 0 ja 1 vahele.

```
scaler = MinMaxScaler(feature_range=(0, 1))
scaled = scaler.fit_transform(values)
```

Lisaks on vaja algoritmi paremaks õppimiseks luua uued tabeli veerud, mis kajastavad ajahetke $t + 1$ tulemusi.

```
reframed = series_to_supervised(scaled, 1, 1)


def series_to_supervised(data, n_in=1, n_out=1, dropnan=True):
        n_vars = 1 if type(data) is list else data.shape[1]
        df = DataFrame(data)
        cols, names = list(), list()
        # input sequence (t-n, ... t-1)
        for i in range(n_in, 0, -1):
                cols.append(df.shift(i))
                names += [('var%d(t-%d)' % (j+1, i)) for j in range
                   (n_vars)]
        # forecast sequence (t, t+1, ... t+n)
        for i in range(0, n_out):
                cols.append(df.shift(-i))
                if i == 0:
                        names += [('var%d(t)' % (j+1)) for j in
                           range(n_vars)]
                else:
                        names += [('var%d(t+%d)' % (j+1, i)) for j
                           in range(n_vars)]
        # put it all together
        agg = concat(cols, axis=1)
        agg.columns = names
```

```
        # drop rows with NaN values
        if dropnan:
                agg.dropna(inplace=True)
        return agg
```

Kui on loodud vastavad veerud, siis eemaldame ebavajalikud veerud, mida ennustada ei soovi.

```
reframed.drop(reframed.columns
    [[14,15,16,17,18,19,20,21,22,23,24,25]], axis=1, inplace=True)
```

Seega tekkisid Joonisel 2.19 kujutatud veerud:

	var1(t-1)	var2(t-1)	var3(t-1)	var4(t-1)	var5(t-1)	var6(t-1)	\
1	0.188068	0.202458	0.177710	0.177710	0.218480	0.283134	
2	0.270146	0.280808	0.255486	0.255486	0.299954	0.367566	
3	0.237958	0.250082	0.224853	0.224853	0.267865	0.260199	
4	0.274399	0.284868	0.260917	0.260917	0.305644	0.400751	
5	0.264628	0.275540	0.262437	0.262437	0.307237	0.426527	

	var7(t-1)	var8(t-1)	var9(t-1)	var10(t-1)	var11(t-1)	var12(t-1)	\
1	0.240515	0.456575	0.414292	0.354559	0.530518	0.452665	
2	0.306027	0.697574	0.239656	0.645769	0.756186	0.321181	
3	0.271619	0.659770	0.314182	0.243248	0.663415	0.353561	
4	0.308078	0.692035	0.290248	0.303978	0.670526	0.434303	
5	0.299077	0.674565	0.301613	0.407432	0.580343	0.295140	

	var13(t-1)	var1(t)
1	0.561525	0.270146
2	0.646885	0.237958
3	0.813291	0.274399
4	0.674029	0.264628
5	0.563514	0.264168

Joonis 2.19: Reframed tabel - andmehulgad, mida kasutame ennustamisel.

Seejärel jaotame taas andmed test- ja treeningandmeteks.

```
values = reframed.values
days_recorded = len(reframed.values)
n_train_days = round(days_recorded / 8)

test = values[n_train_days*7:]
train = values[:n_train_days*7]
```

Tulemuseks saame 665 päeva test- ja 95 päeva treeningandmeid. Testandmeteks on viimased 95 päeva ehk sisuliselt 3 kuud. *Sequence data* ehk järjestikuste andmete puhul on oluline andmeid mitte segada. Samuti tuleb võtta ajavahemikud, mitte üksikud päevad terve andmehulga peale. *Sequence data* puhul sõltuvad andmed enda

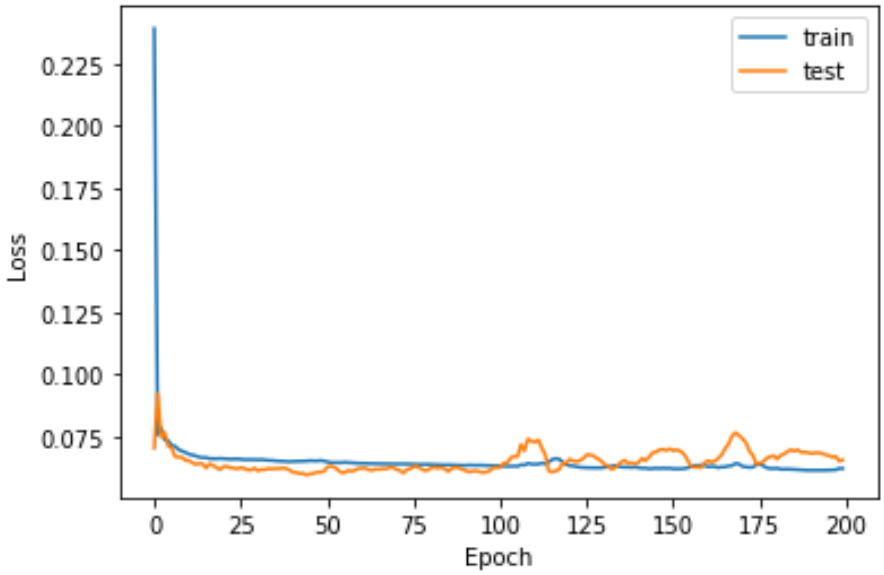
järjekorrast. Seejärel jagan andmed sisenditeks ja väljunditeks. Lõpuks teisendan andmed kujule [samples, timesteps, features]

```
train_X, train_y = train[:, :-1], train[:, -1]
test_X, test_y = test[:, :-1], test[:, -1]

train_X = train_X.reshape((train_X.shape[0], 1, train_X.shape[1]))
test_X = test_X.reshape((test_X.shape[0], 1, test_X.shape[1]))
```

Andmete eeltöötlemine on sellega valmis ning järgnevalt asume mudeli loomise juurde. Üks parimatest NN töövahenditest on *tensorflow*, mis võimaldab väga hästi võrgustikke ehitada ja nendega vajalikke operatsioone teha. Algselt koostame LSTM mudeli, milles on 32 nodei ning lisame 1 väljundkihi, mis annab meile meie väljundi. Epochide arvuks määrame 200 ja *batch size*-iks 32. Seejärel alustame mudeli treenimist.

```
model = Sequential()
model.add(LSTM(32, input_shape=(train_X.shape[1], train_X.shape[2])
    ))
model.add(Dense(1))
model.compile(loss='mae', optimizer='adam')

history = model.fit(train_X, train_y, epochs=200, batch_size=32,
    validation_data=(test_X, test_y), verbose=2, shuffle=False)
```

Kaardistame visuaalselt kuidas mudel treenis (Joonis 2.20).



Joonis 2.20: Mudeli treeningut kujutav graafik.

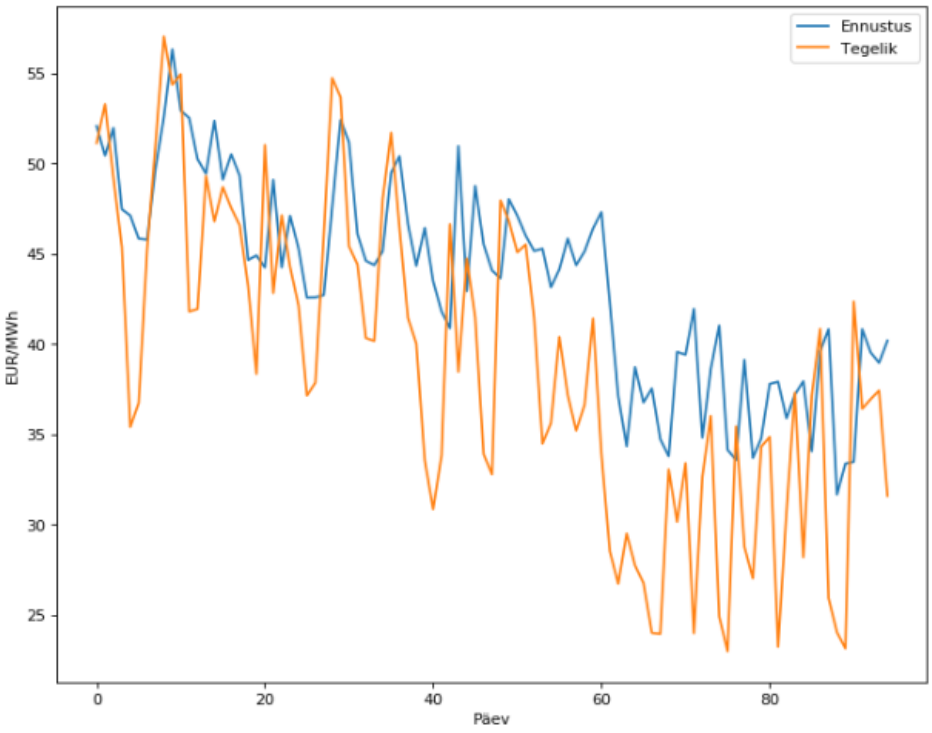
Graafikult on näha kuidas iga *epoch*-iga vähenes vigade arv kuni 100-nda *epoch*ini, mis hetkest hakkas pigem ebastabiilselt kõikuma. Järgnevalt testime mudelit testandmetega ning arvutame kui palju erineb mudeli ennustus tegelike andmetega.

```
yhat = model.predict(test_X)
test_X = test_X.reshape((test_X.shape[0], test_X.shape[2]))
```

Selleks tuleb ennustatud tulemused ja algselt normaliseeritud väärtused tagasi teisendada.

```
inv_yhat = concatenate((yhat, test_X[:, 1:]), axis=1)
inv_yhat = scaler.inverse_transform(inv_yhat)
inv_yhat = inv_yhat[:,0] #hinna ennustuse v22rtused

test_y = test_y.reshape((len(test_y), 1))
inv_y = concatenate((test_y, test_X[:, 1:]), axis=1)
inv_y = scaler.inverse_transform(inv_y)
inv_y = inv_y[:,0] #eesti hinna v22rtused
```

Kanname tulemused graafikule (Joonis 2.21) ning arvutame mudeli hindamiseks vead.



Joonis 2.21: Esialgne LSTM hinnaennustuse graafik.

```
Mean Absolute Error: 6.047977
Mean Squared Error: 55.153454
Root Mean Squared Error: 7.426537
```

2.7.4 LSTM mudeli hüperparameetrite otsimine

Hüperparameetriteks võib pidada tehisnärvivõrgu nuppudeks, mille muutmisel mudeli täpsus paraneb või just vastupidi halveneb. Kui leida õige hüperparameetrite kombinatsioon, siis võib mudel drastiliselt paremaks muutuda. Antud töö raames määrasin muudetavateks parameetriteks LSTM mudeli esimese kihi neuronite arvu, esimese kihi aktiveerimisfunktsiooni, teise kihi olemasolul teise kihi neuronite arvu, teise kihi aktiveerimis funktsiooni. Lisaks *epoch*-ide arv ning *batch size*. Igal parameetril on kindlate väärtuste hulk, mille seast leida kõige õigem.

```
FIRST_LAYER_NODES = [8, 16, 32, 64]
FIRST_LAYER_ACTIVATIONS = ['sigmoid', 'tanh', 'relu']
SECOND_LAYER_STATES = [0, 1]
```

```
SECOND_LAYER_NODES = [1, 2, 4, 8, 16, 32, 64]
SECOND_LAYER_ACTIVATIONS = ['sigmoid', 'tanh', 'relu']
EPOCHS = [20, 40, 60, 100, 200, 350, 500]
BATCH_SIZES = [32, 64, 128]
```

Õige kombinatsiooni leidmiseks on mitmeid erinevaid meetodeid. Manuaalne proovimine ning katsetamine on kõige lihtsam. Samuti on kasutatud tabeli otsingu (*grid search*) meetodit, kus luuakse kõikide parameetrite kombinatsioon ning proovitakse ükshaaval mudeli peal ära. Lisaks eksisteerib *random search*, kus proovitakse suvalisi kombinatsioone, mis pakub häid tulemusi palju kiiremini kui tabeli otsing, kuid tõenäoliselt ei leia parima kombinatsiooni. Peale selle eksisteerib veel *Bayesian optimization algoritm*, mis kasutab heuristikaid.

Antud töö raames otsustasin kasutada *grid search* algoritmi, sest soovisin leida kõige paremat kombinatsiooni ning ei olnud ajaliselt piiratud, et leida vajalik kombinatsioon. Selleks kirjutasin koodi, mis loob erinevad hüperparameetrite kombinatsioonide seadete faili.

Seda faili kasutab LSTM mudeli kood, mis katsetab igat seadete kombinatsiooni mudelis. Antud töö põhjal genereerisin 5544 erinevat hüperparameetrite kombinatsiooni. Iga kombinatsioon ja mudeli vea tulemused kirjutasin omakorda tulemuste faili. Kuna igat seadet prooviti mudelis vaid 1 kord, siis võib tulemuseks olla juhuslik seade, millel võis lihtsalt õnneks minna. Seetõttu otsisin välja 5544 kombinatsiooni seast 100 seadet, millest kõik olid 4% parimate *Mean Absolute Error*, *Mean Squared Error* ja *Root Mean Squared Error* tulemuste seas.

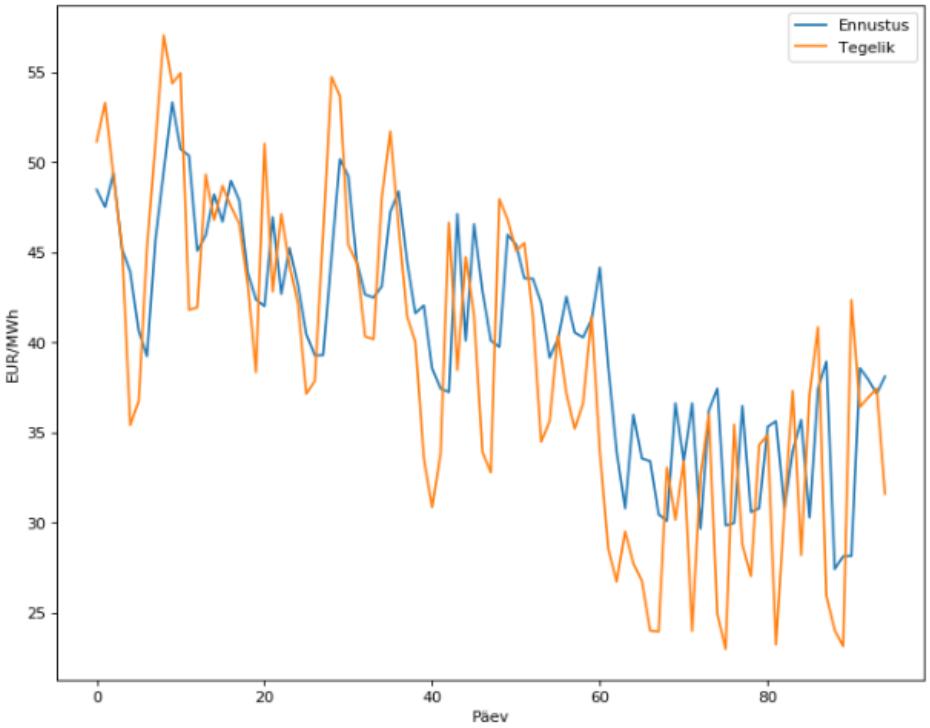
Parimad alles jäänud 100 seadet jooksutasin taas 10 korda LSTM mudelist läbi ning leidsin nendest keskmise. Parimaks osutus kahe kolme kihiline mudel, mille esimene kiht koosnes 64 neuronist ning sigmoid aktiveerimisfunktsioonist. Teine kiht koosnes 1 neuronist ning tanh aktiveerimisfunktsioonist. Mudelit treeniti 500 *epoch*-iga ning *batch size*-iks võeti 32.

Kasutades antud seadeid mudel loomisel saame keskmisteks vea suurusteks järgneva:

```
Mean Absolute Error: 4.68796
Mean Squared Error: 34.13489
Root Mean Squared Error: 5.84094
```

Antud mudeli loomisel kasutasin peatükis 2.2 leitud parimaid elektrihinda mõjutavaid parameetreid, mis minu hinnangul on Soome, Läti, Leedu, Taani, Rootsi

ja Norra elektrihind. Lisaks Eesti tarbimine ja tuuleenergia kogus. Samuti Eesti, Soome, Läti ja Rootsi elektri ost/müük. Kannan ennustus ja tegelikud andmed graafikule (Joonis 2.22). Joonise x telg on kajastatud ajaühikutena, milleks on päevad. Tabeli y telg on kajastatud elektrihinnana. Tegelikud andmed on kujutatud graafikul oranžiga ning ennustatud andmed sinisega. Nagu näha on elektri hind väga hüplik, kuid sellest hoolimata suutis mudel suhteliselt edukalt hinna liikumist ennustada. Mudeli murekohaks on suuremad hüpped, mida on raske prognoosida. Heaks näiteks on ajahetk 40, kus hind langes järsku alla 35 hinna ühiku. Mudel suutis küll trendi väga hästi näha, kuid seda mitte, kui madalale hind reaalselt langeda võib. Näha on ka trendi, kus jaanuarikuu hind muutus palju ebastabiilsemaks, mis võis olla tingitud maailmas alguse saavast nafta hinna sõjast ja selle langusest, samuti sai alguse koroona kriis juba 2019. aasta detsembris. Sellest hoolimata hindan antud mudelit suhteliselt täpseks arvestades ebastabiilset turgu.



Joonis 2.22: Parimate hüperparameetritega LSTM hinnaennustuse graafik.

Tulemuste analüüs

3.1 Mudelite tulemuste võrdlus

Töö kolmandas osas uurin, milline töös loodud mudelitest on kõige täpsem. Samuti võrdlen töös kasutatavaid elektri hinna ennustamise mudeleid. Lisaks hindan millist mõju avaldas koroona ja naftahinna langemisega mudelite täpsus, mis võtavad testandmehulgaks andmed, mis pärinevad 1. veebruarist 2020 kuni 31. märtsini 2020. Lõpetuseks pakun välja mudelite edasiarenduste ja täiustamise ideid, mis antud töö raamidesse ei mahtunud.

3.1.1 Parim *sliding window* mudel

Töö andmete analüüsi osas leidsime, et kõikidest andmetest kõige paremaid korrellatsiooni tulemusi näitasid Soome, Läti, Leedu, Rootsi, Norra ja Taani elektri hinnad ning Soome, Rootsi ja Läti ostu/müügi kogused. Lisaks Eesti hind, elektri toodang, ost/müük ja tuul. Proovides ennetada eriolukorrast tingitud trende mudelisse õppimast, võtame arvesse perioodi 1. jaanuar 2018 - 31. jaanuar 2020.

Parima liikuva akna mudeli väljaselgitamiseks proovisin erinevaid akna suuruseid ning kandsin tulemused tabelisse.

Tabel 3.1: Sliding window mudeli võrdlus 1.01.2018 - 31.01.2020

Window size	30	60	90	120
Mean Absolute Error	8.239	6.987	6.872	8.138
Mean Squared Error	101.671	79.127	69.106	100.913
Root Mean Squared Error	10.083	8.895	8.313	10.046
Mean Absolute Percentage Error	27.726	21.334	19.191	22.030
Maximum Error	24.171	23.946	21.848	24.794
Minimum Error	0.034	0.109	0.099	0.057

3.1.2 Parimate korrellatsioonidega andmete mudelid

Kasutame samasid lähteandmeid nagu eelmises alampeatükis ning võtame taas arvesse perioodi 1. jaanuar 2018 - 31. jaanuar 2020. Võtame võrdlusesse samal ajavahemikul nii lineaarregresiooni tavalist mudelit, *sliding window* mudelit kui ka LSTM mudeli.

Eelnevas alapealtükis selgitasime välja, et parimaid *sliding window* mudeli tulemusi andis 90 päeva pikkuse aknaga mudel. Võrdleme mudelite vigade suurust omavahel.

Tabel 3.2: Kõikide korrelleerivate andmete mudelite võrdlus 1.01.2018 - 31.01.2020

Type of error	Lineaarregressioon	Sliding window	LSTM
Mean Absolute Error	5.471	6.872	4.750
Mean Squared Error	46.121	69.106	34.827
Root Mean Squared Error	6.791	8.313	5.900
Mean Absolute Percentage Error	16.820	19.191	14.062
Maximum Error	16.516	21.848	14.415
Minimum Error	0.062	0.099	0.034

3.1.3 Elektrihinna andmetega mudelid

Korrellatsiooni analüüsis tuli selgelt välja fakt, et kõige enam mõjutavad Eesti elektri hinda Nordpooli teiste riikide hinnad. Seetõttu soovin teada, kas mudelite ennustus paraneb kui ainult hinna andmeid mudelites kasutan. Proovides ennetada eriolukorrast tingitud trende mudelisse õppimast, võtame arvesse perioodi 1. jaanuar 2018 - 31. jaanuar 2020. Mudelite tulemused kannan tabelisse.

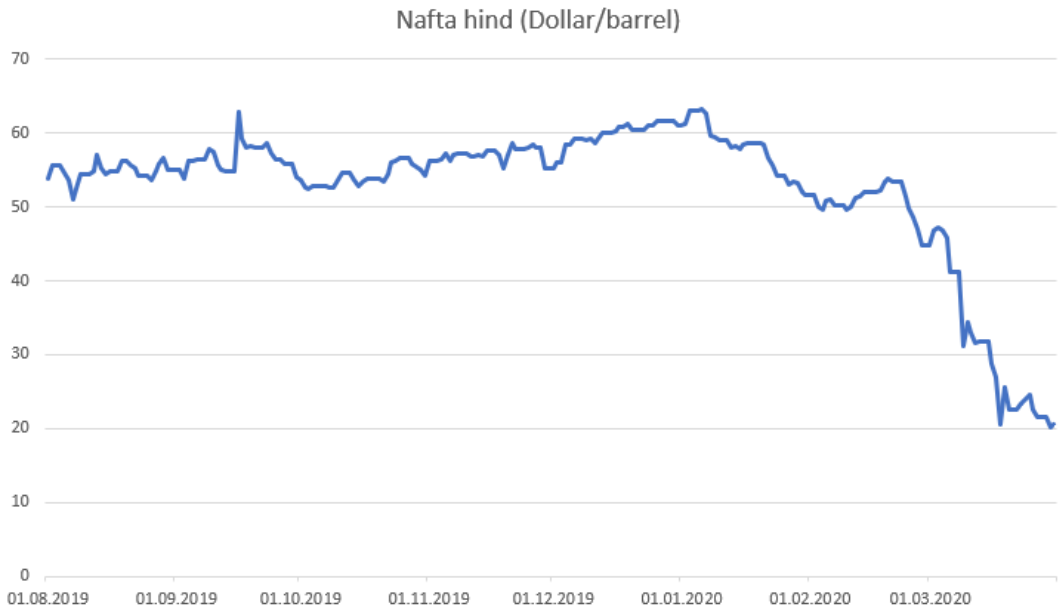
Tabel 3.3: Elektrihindadest loodud mudelite võrdlus 1.01.2018 - 31.01.2020

Type of error	Lineaarregressioon	LSTM
Mean Absolute Error	5.598	4.902
Mean Squared Error	47.867	35.450
Root Mean Squared Error	6.919	5.950
Mean Absolute Percentage Error	17.071	14.510
Maximum Error	15.043	15.538
Minimum Error	0.089	0.148

Tabelist on näha, et mõlema, nii lineaarregressiooni kui ka LSTM mudeli puhul ennustus halvenes vähesel määral kui kasutada vaid hinna andmeid.

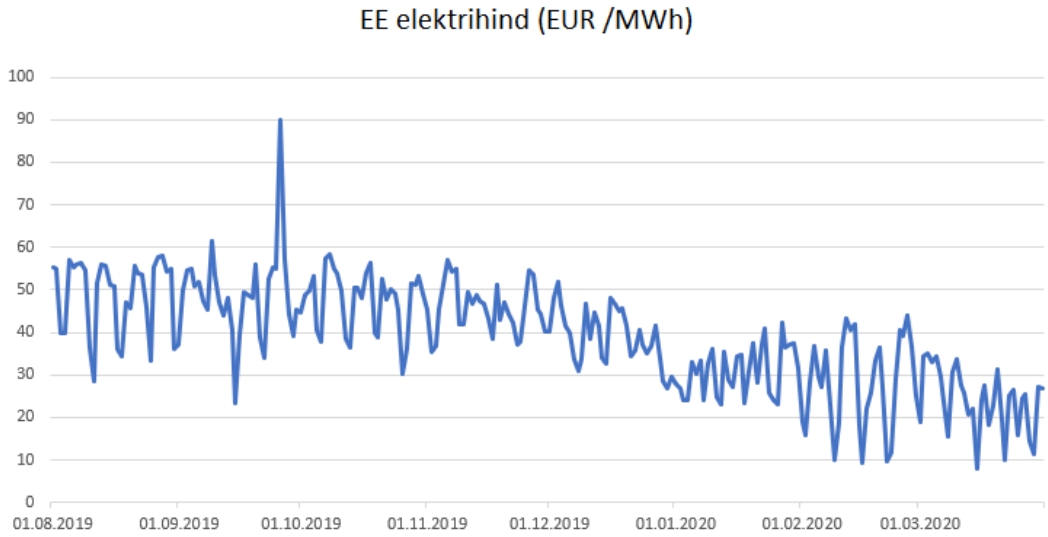
3.2 Maailma sündmuste mõju mudelite ennustusele

2019. aasta lõpus algas Hiinas globaalne viiruse COVID-19 pandeemia. 2020. aasta alguses algas nafta maailmaturu hinna langus, mis oli tingitud nii riikide vahelisest hinnasõjast kui ka viiruse levikust. Tehased suleti, inimesed olid kodus karantiinis, mistõttu nõudlust kütusele enam ei olnud. Antud eriolukord tekitas nafta ja elektrihindade suure languse, mis on näha ka kujutatud Joonisel 3.1. Graafiku x teljeks on aeg ning y teljeks nafta barreli hind dollarites.



Joonis 3.1: Nafta hinna ajalooline graafik.

Lisasin ka Eesti elektrihinna languse trendi, mis on kujutatud Joonisel 3.2. Selgelt on näha uue aasta algusega on hinnad languses. Kui varem ulatusid hinnad ligi 60 euroni, siis 2020. aastast sattusid üksikutel päevadel hinnad üle 40 euro ning pigem kõikusid 30 euro lähistel.



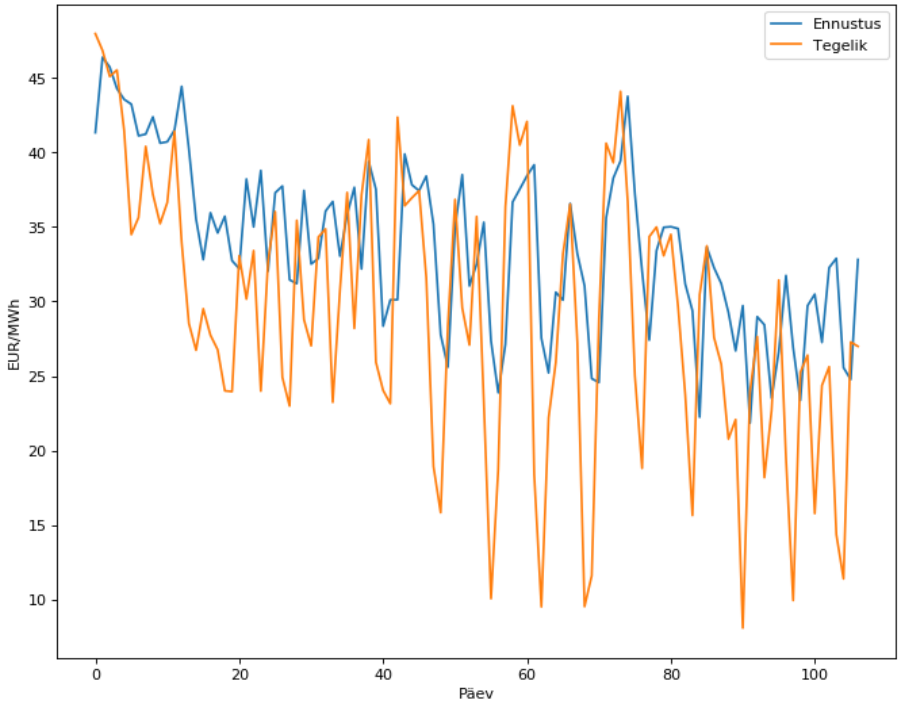
Joonis 3.2: Eesti elektrihinna ajalooline graafik.

Töös eelnevalt loodud mudelid on loodud kriisieelsete andmetega. Võtame kriisiaegsed test andmed ning uurime, kui suurt mõju antud ennustusele kriis avaldas. Võrdleme nii mudeleid, mis on loodud vaid riikide hindadest (tähistatud *) kui ka mudeleid, mis on koostatud kõikidest tähtsamatest parameetritest. Samuti kujutame kriisiaegse lineaarregressiooni ennustuse Joonisel 3.3.

Tabel 3.4: Linaarregressiooni mudeli ennustuse täpsus kriisi ajal

Type of error	Enne kriisi	Kriisi ajal	Enne kriisi*	Kriisi ajal*
MAE	5.471	6.667	5.598	6.806
MSE	46.121	72.625	47.867	73.503
RMSE	6.791	8.522	6.919	8.573
MAPE	16.820	34.039	17.071	33.893
Maximum Error	16.516	21.619	15.043	21.145
Minimum Error	0.062	0.0034	0.089	0.029

* Ainult hinna andmeid kasutav mudel.



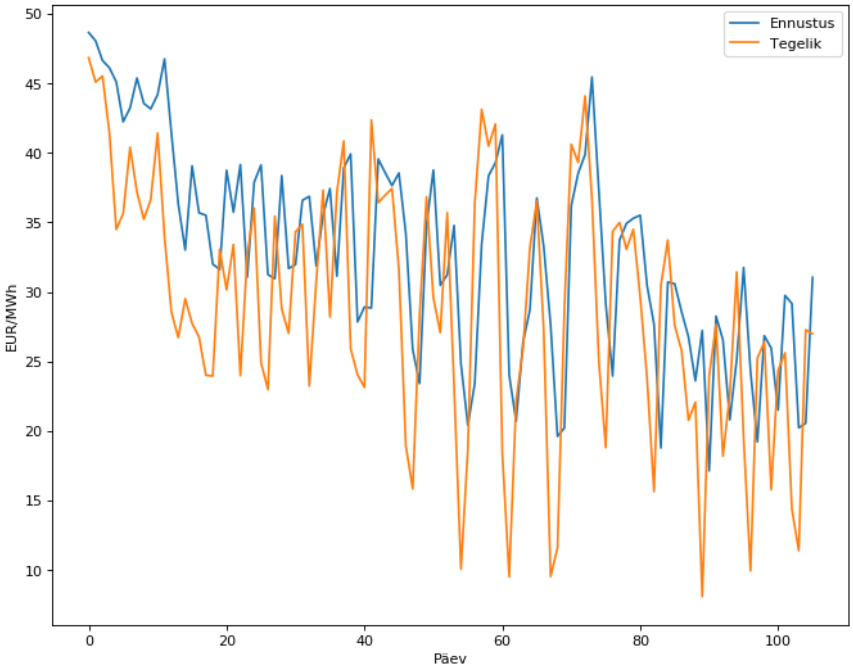
Joonis 3.3: Kriisiaegne lineaarregressiooni ennustuse graafik.

Vaatleme millist mõju avaldas kriis LSTM mudelile. Võrdleme samuti kriisi aegseid ja kriisile eelnevate ennustuste täpsust. Mudelid, mille loomisel kasutati vaid hinna andmeid on tähistatud tärniga(*). Kujutame kriisi aegse LSTM ennustuse Joonisel 3.4.

Tabel 3.5: LSTM mudeli ennustuse täpsus kriisi ajal

Type of error	Enne kriisi	Kriisi ajal	Enne kriisi*	Kriisi ajal*
MAE	4.750	6.624	4.902	7.677
MSE	34.827	67.286	34.450	91.310
RMSE	5.900	8.198	5.950	9.555
MAPE	14.062	30.888	14.510	36.440
Maximum Error	14.415	21.662	15.538	26.389
Minimum Error	0.034	0.089	0.148	0.127

* Ainult hinna andmeid kasutav mudel.



Joonis 3.4: Kriisiaegne LSTMi ennustuse graafik.

Selgub, et paremaid tulemusi annavad mudelid, milles on lisaks riikide hinna andmetele ka ostu/müügi andmed ning Eesti tootmise ja tuuleenergia andmed. LSTM mudel oli selgelt parem lineaarregressiooni mudelitest ning ajaperioodil 01.01.2018-31.01.2020 tuli LSTM mudeli parimaks MAE väärtuseks 4.750. Antud mudeli MAPE väärtuseks tuli 14.062.

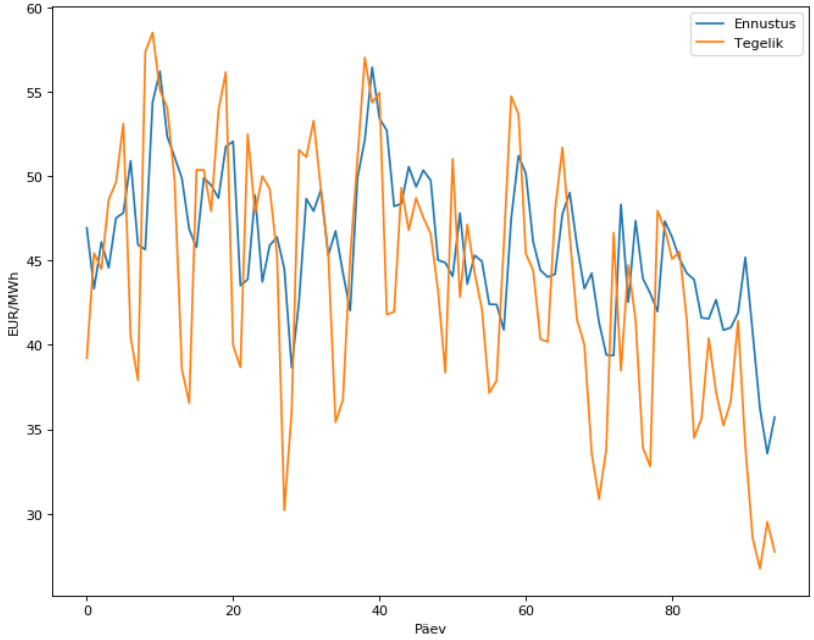
Lisaks ei olnud võimalik saada paremaid lineaarregressiooni tulemusi *sliding window* meetodit kasutades. Järeldusena saab öelda, et elektri hinna trendi on parem ennustada lineaarregressiooniga kasutades kaugema mineviku andmeid ning lähimineviku andmed mängivad ennustamisel vähem rolli.

Vaadates nafta hinna graafi on näha, et hinna langus algas juba jaanuari algusest, seetõttu proovin mudelit ka perioodil 01.01.2018-31.12.2019, et näha millist effekti see mudelile avaldab.

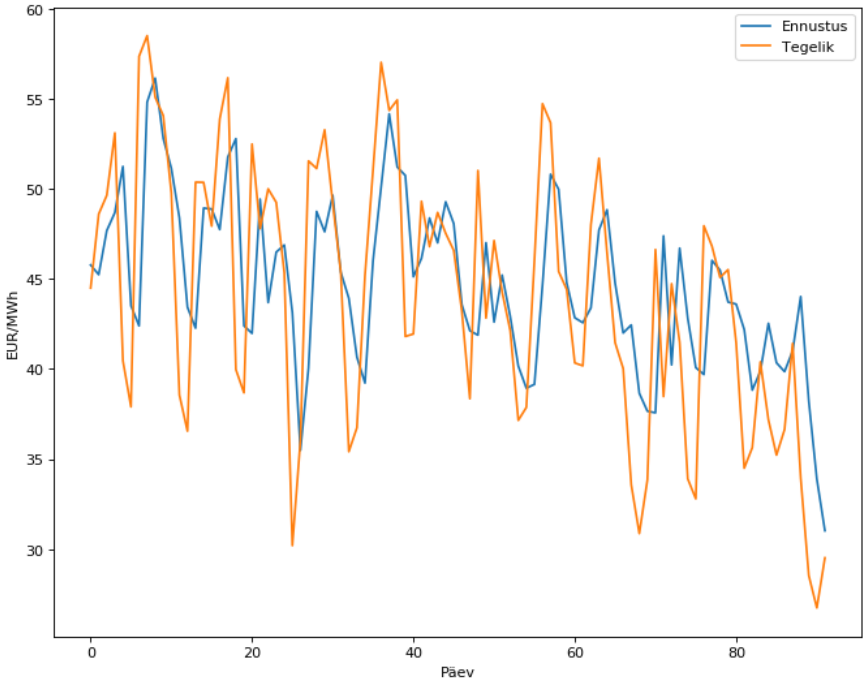
Tabel 3.6: Kõikide korrelleerivate andmete mudelite võrdlus 1.01.2018 - 31.12.2019

Type of error	Lineaarregressioon	LSTM
Mean Absolute Error	5.012	4.815
Mean Squared Error	37.319	36.367
Root Mean Squared Error	6.109	6.018
Mean Absolute Percentage Error	12.612	11.730
Maximum Error	14.320	15.502
Minimum Error	0.032	0.029

Tulemusena näeme, et nii lineaarregressiooni ja LSTM mudeli tulemus paranes taas märgatavalt. Lineaarregressiooni (Joonis 3.5) puhul paranes kõik vaadeldavad vea hulgad. MAPE paranes koguni 4.2 ühiku võrra (12.61). Kuigi LSTMi puhul MAE, MSE ja RMSE läksid vähesel määral suuremaks, vähenes MAPE 2.33 ühikut (11.730). Seega avaldab juba kriisi varasemad staadiumid olulist mõju mudeli ennustusvõimele.



Joonis 3.5: Kriisile eelnev lineaarregressiooni ennustuse graafik va. jaanuari kuu.



Joonis 3.6: Kriisile eelnev LSTM ennustuse graafik va. jaanuari kuu.

Parimaks mudeliks saime LSTM mudeli (Joonis 3.6), mille parimad tulemused avalduvad kasutades kriisieelseid andmeid vahemikus 01.01.2018-31.12.2019. Mudeli MAE-ks sain 4.815, RMSE 6.018, MAPE 11.730. Seega hindan mudeli sooritust heaks ning antud mudeli veaprotsent jääb peatükis "Olemasolevad mudelid ja nende tulemused" välja toodud mudelite veaprotsendi vahemikku.

3.3 Mudeli edasiarendus ja täiustamine

Töös selgitasime välja olulised faktorid, mis mõjutavad elektri hinda ning ehitasime ennustuse mudelid, mis töötavad elektri hinna ebastabiilsel turul võrdlemisi hästi. Sellest hoolimata on võimalik mudeleid paremaks ja kindlamaks ehitada. Olulisteks mudeli täpsuse faktoriteks on andmed. Töö raames oli kasutuses 27 kuu andmed. Kui Nordpool või mõni teine asutus oleks valmis väljastama rohkem andmeid, siis paraneksid ennustused samuti. Lisaks võib elektri hinda mõjutada mõni faktor, mida antud töö raames ei uuritud ning mille kasutusele võtmisel paraneks mudel märgatavalt.

Töö käigus oli selgelt näha, kui suurt mõju võib erakorraline olukord ja kriis hinnale ja selle ennustamisele mõjuda. Seetõttu oleks oluline tulevikus uurida, kuidas oleks võimalik parandada ennustusi nii, et kriisi ajal oleks samuti ennustused võimalikult täpsed. Oluline oleks leida parameetreid, mis kirjeldavad kriisiolukorda hinnaturul kõige paremini.

Lisaks eelnevale on oluline ka proovida ennustamisel teisi meetodeid. Antud töö raames sai uuritud ühte neurovõrkude ja ühte statistilist meetodit. Samas eksisteerib palju erinevaid meetodeid, mille abil on võimalik ennustada. Taolised meetodid tõin välja antud töö ennustamise liikide osas.

Kokkuvõte

Töö käigus võtsime vaatluse alla nii elektrituru, seda mõjutavaid tegureid, elektri hinna väljakujunemise, kui ka kõik vajalikud vahendid, et hinda edukalt ennustada. Töö jagati kolme etappi, mille jooksul prooviti vastuseid saada püstitatud küsimustele. Töö esimeses osas uurisime elektrit ja elektriturgu, selle omapärasid ning komponente. Püstitasin eesmärgiks leida faktorid mis mõjutavad kõige enam elektri hinda. Kõige enam mõjutas Eesti elektri hinna kujunemist Soome, Norra, Rootsi, Läti, Leedu elektri hinnad. Arvestataval määral mõjutasid hinda ka Soome, Rootsi ja Läti elektri ostu/müügi kogused. Lisaks olid oluliseks faktoriteks Eesti enda elektrihinna trend, elektri ost/müük, elektritoodangu suurus ning tuuleenergia kogus. Arvatust väiksemat korrellatsiooni omas temperatuur. Kõige paremini oli võimalik ennustamisel kasutada Nordpooli poolt pakutavaid andmeid, mis olid suuremas osas heas korrelatsioonis ning andmete kvaliteet oli samuti väga hea.

Teises osas võtsin vaatluse alla erinevad ennustamise metoodikad ning kirjeldasin lahti töös kasutatavaid võtteid. Lisaks uurisin olemasolevate tööde põhjal, milliseid võtteid on ennustamisel kasutatud ning kui täpsed ennustused antud mudelid on teinud. Selgus, et ennustused kõiguvad olenevalt riigist keskmiselt 5-12 *Mean absolute percentage error* vahel, mis on elektri ebastabiilse turu kohta vägagi aktsepteeritavad tulemused.

Töö kolmandas osas uurisin täpsemalt mudeleid ning kuidas erinevad mudelid erinevates ajahetkedes toime tulevad. Selgus, et LSTM tehisnärvivõrk töötas ennustamisel nähtavalt paremini võrreldes lineaarregressiooni mudeliga. Samas ei andnud lineaarregressiooni mudel üldse halbu tulemusi. Kriisieelse aja ennustuse *Mean absolute error* oli 5.012 ning *Mean absolute percentage error* oli 12.612. Parimaid tulemusi näitas kriisieelne LSTM mudel, mille *Mean absolute error* oli

4.815 ning *Mean absolute percentage error* oli 11.730. Lisaks uurisin töö käigus maailma mõjutanud naftahinna ja koroona kriisi. Tahtsin teada, kui palju ja mil määral on kriis mõjutanud mudelite ennustamist ning tulemusi. Selgus, et kriisi ajal vähenes LSTM mudeli ennustamise täpsus ligi 2.6 korda ja lineaarregressiooni mudeli täpsus 2.7 korda.

Töö käigus sain vastused püstitatud küsimustele. Töö edasiarendamisel oleks huvitav leida, milliseid tulemusi võib saada teiste närvivõrgu või muude ennustamise mudelitega. Samuti oleks hea leida parameetreid või faktoreid, millega oleks võimalik kriisiolukordades viia ennustuse viga võimalikult väikeseks.

Kasutatud kirjandus

[1] C. Olah. Understanding lstm networks. 08 2015. URL https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[2] R. Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30 (4):1030 – 1081, 2014. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2014.08.008. URL http://www.sciencedirect.com/science/article/pii/S0169207014001083.

[3] G. P. Girish and S. Vijayalakshmi. Determinants of electricity price in competitive power market. *International Journal of Business and Management*, 8: 70–75, 10 2013. doi: 10.5539/ijbm.v8n21p70.

[4] J. Stephenson, B. Barton, G. Carrington, D. Gnoth, R. Lawson, and P. Thorsnes. Energy cultures: A framework for understanding energy behaviours. *Energy Policy*, 38(10):6120 – 6129, 2010. ISSN 0301-4215. doi: https://doi.org/10.1016/j.enpol.2010.05.069. URL http://www.sciencedirect.com/science/article/pii/S0301421510004611. The socio-economic transition towards a hydrogen economy - findings from European research, with regular papers.

[5] R. V. Jones, A. Fuertes, and K. J. Lomas. The socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings. *Renewable and Sustainable Energy Reviews*, 43:901 – 917, 2015. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2014.11.084. URL http://www.sciencedirect.com/science/article/pii/S1364032114010235.

[6] I. S. Bayram and T. S. Ustun. A survey on behind the meter energy

management systems in smart grid. *Renewable and Sustainable Energy Reviews*, 72:1208 – 1232, 2017. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2016.10.034. URL http://www.sciencedirect.com/science/article/pii/S1364032116306852.

[7] P. Shrestha and P. Kulkarni. Identifying factors that affect the energy consumption of residential buildings. pages 1437–1446, 05 2010. ISBN 978-0-7844-1109-4. doi: 10.1061/41109(373)144.

[8] M. Ventosa, A. Baillo, A. Ramos, and M. Rivier. Electricity market modeling trends. *Energy Policy*, 33(7):897–913, 2005. URL https://EconPapers.repec.org/RePEc:eee:enepol:v:33:y:2005:i:7:p:897-913.

[9] JJ. Mae and rmse — which metric is better? 03 2016. URL https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d.

[10] C. Kathuria. Regression — why mean square error? 12 2019. URL https://towardsdatascience.com/https-medium-com-chayankathuria-regression-why-mean-square-error-a8cad2a1c96f.

[11] A. Myttenaere, B. Golden, B. Le Grand, and F. Rossi. Mean absolute percentage error for regression models. *Neurocomputing*, 192:38 – 48, 2016. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2015.12.114. URL http://www.sciencedirect.com/science/article/pii/S0925231216003325. Advances in artificial neural networks, machine learning and computational intelligence.

[12] J.P.S. Catalão, S.J.P.S. Mariano, V.M.F. Mendes, and L.A.F.M. Ferreira. Short-term electricity prices forecasting in a competitive market: A neural network approach. *Electric Power Systems Research*, 77(10):1297 – 1304, 2007. ISSN 0378-7796. doi: https://doi.org/10.1016/j.epsr.2006.09.022. URL http://www.sciencedirect.com/science/article/pii/S0378779606002422.

[13] H.M.I. Pousinho, V.M.F. Mendes, and J.P.S. Catalão. Short-term electricity prices forecasting in a competitive market by a hybrid pso–anfis

approach. *International Journal of Electrical Power and Energy Systems*, 39(1):29 – 35, 2012. ISSN 0142-0615. doi: https://doi.org/10.1016/j.ijepes.2012.01.001. URL http://www.sciencedirect.com/science/article/pii/S0142061512000026.

[14] T. Kristiansen. Forecasting nord pool day-ahead prices with an autoregressive model. *Energy Policy*, 49:328 – 332, 2012. ISSN 0301-4215. doi: https://doi.org/10.1016/j.enpol.2012.06.028. URL http://www.sciencedirect.com/science/article/pii/S0301421512005381. Special Section: Fuel Poverty Comes of Age: Commemorating 21 Years of Research and Policy.

[15] A. Schneider, G. Hommel, and M. Blettner. Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Deutsches Arzteblatt international*, 107(44):776–782, Nov 2010. ISSN 1866-0452. doi: 10.3238/arztebl.2010.0776. URL https://doi.org/10.3238/arztebl.2010.0776.

[16] Yaroslav Shkarupa, Roberts Mencis, and Matthia Sabatelli. Offline handwriting recognition using lstm recurrent neural networks. volume 1, page 88, 11 2016.

[17] A. Graves, N. Jaitly, and A. Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 273–278, 2013.

[18] K. Chen, Y. Zhou, and F. Dai. A lstm-based method for stock returns prediction: A case study of china stock market. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2823–2824, 2015.

[19] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. K. Wong, and W.-C. WOO. Convolutional lstm network: A machine learning approach for precipitation nowcasting. 06 2015.

[20] N. S. Chauhan. A beginner's guide to linear regression in python with scikit-learn. 02 2019. URL https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[21] J. Brownlee. Multivariate time series forecasting with lstms in keras. 08 2019.

URL https://machinelearningmastery.com/\multivariate-time-series-forecasting-lstms-keras/.