Auteur: Bernie Klous

Opleiding:Winc Academy Data Analytics with Python

27 november 2023

Eindopdracht CO_2-emissies

I. Introductie

De Data Analyse met Python opleiding sluit af met een opdracht drie vragen te beantwoorden omtrent de uitstoot van het broeikasgas CO_2. Teneinde deze vragen uit te werken is het gebruik van de juiste online datasets en methoden uit het curriculum vereist.

II. Probleemstelling

De wereld heeft te maken met de gevolgen van CO_2 uitstoot in de atmosfeer en landen doen hun best te voldoen aan de afspraken in het klimaat akkoord van 2015 om de uitstoot terug te dringen. De volgende vragen die te maken hebben met de uitstoot van CO_2 beantwoord ik in dit rapport:

1. Wat is de belangrijkste voorspelfactor voor de CO_2 uitstoot van een land?
2. Welke landen boeken de meeste vooruitgang de uitstoot omlaag te brengen?
3. Welke duurzame energie technologie heeft de laagste prijs in de toekomst?

III. Data en methodiek

In dit onderdeel zal ik per vraag toelichten welke benadering ik heb gekozen voor het probleem, welke data bronnen ik gebruik en op welke manier ik uitvoering heb gegeven aan de opdracht.

Data

Bij vraag 1 is de databron https://data.worldbank.org/indicator/ behalve de dataset Aantal internet gebruikers users deze is van https://ourworldindata.org/, net als de rest van de data die in de opdracht is gebruikt.

Ontbrekende data

Voor een aantal grootheden in de eerste vraag waren minder gegevens beschikbaar. Voor duurzaam energiegebruik en Aantal internetgebruikers was data vanaf 1990 beschikbaar, voor Algemeen

energiegebruik vanaf 1998. Andere variabelen hadden een langere tijdsspanne waarmee ik wel een selectie vanaf 1960 heb gemaakt.

In de gegevens voor de derde vraag ontbrak een waarde die ik heb kunnen aanvullen. Het aantal datapunten in de dataset voor vraag 3. is ook klein. Voor sommige landen maar vijf observaties voor een variabele, en n=11 voor enkel variabelen op wereldniveau. Dat is ongeveer het minimum[1] om een regressie te kunnen doen.

Methoden

In dit gedeelte zal ik de methoden die ik heb toegepast in de uitwerking per vraag toelichten.

1.

XGB-regressie

De XGB-regressie begint met een boom, een voorspelling. De residuen van de voorspelling krijgen een blad. Voor elk blad wordt een precisiescore uitgerekend:

$$Similarity\ score = \frac{\sum (Residuals)^2}{Number\ of\ Residuals + \lambda}$$

Dan wordt het blad gesplitst op een kenmerk in een vertakking, de berekening herhaald en het netto effect van de vertakking gemeten. Wanneer de boom groot genoeg is kan deze worden gesnoeid om te kijken of een vertakking een verbetering in de scores geeft of een optimum heeft bereikt.

Iedere boom vertakt zich dus in kleinere residuen met als doel een minimaal individueel residu, aangeduid met de letter T. Een toekomstige boom is afhankelijk van het gemodelleerde bos en ziet er als volgt uit, waarin ε staat voor de leercurve van het model:

$$Y(t)_{prediction} + \varepsilon T_1 + \varepsilon T_2 + ... + \varepsilon T_i$$

KNeighbors-regressie

KNeighbors-regressie gaat uit van de gedachte dat vergelijkbare input leidt tot vergelijkbare output. Het aantal datapunten dat de invoer het dichtst benadert wordt aangegeven met K. Het gemiddelde daarvan is de voorspelde waarde. De nabijheid wordt bepaald door de Euclidische afstand te nemen. Dit geometrische gereedschap berekent de wortel over de verschillen per eigenschap in het kwadraat bij elkaar opgeteld.

Het resultaat is feitelijk geen voorspelling maar een gemiddelde van waarden gegroepeerd naar gelijkenis.

1 https://online.stat.psu.edu/stat501/lesson/12/12.9

2.

Bij vraag 2 was geen statistiek nodig.

3.

Stationariteit

Om tijdreeksen te kunnen voorspellen is de voorwaarde dat data stationair is. Dat houdt dat het gemiddelde en variantie constant zijn en de data geen cyclisch of ander patroon kent. Óf een verschuiving in de tijd heeft geen invloed op de verdeling rond het gemiddelde van de tijdreeks dan is differentiatie, het toevoegen van vertragingen, niet nodig. Voor de gegevens in vraag 3 ben ik met de Augmented Dickey-Fuller (ADF) test nagegaan of aan deze voorwaarde is voldaan.

Auto-Regressie Model

De regressie in vraag 3 heb ik gedaan met het Auto-Regressie (AR) model. Dit model gaat er dus vanuit dat een huidige waarde Y(t) wel afhankelijk is van eerdere waarnemingen Y(t-1), Y(t-2) etc. Vanwege deze aanname kunnen we een regressie doen. De formule voor het AR model, dat haar coëfficiënten berekend met de kleinste kwadraten methode, is:

$$Y(t) = a + b\,Y(t-1) + \epsilon(t)$$

Tijdreeksen kunnen niet alleen van zichzelf afhankelijk zijn maar ook van andere tijdsreeksen. Zo bestaat de dataset in vraag 3 uit verschillende energieprijzen waarvan twee sets (zon- en wind) prijzen onderling gecorreleerd zijn. In het geval van reeksen met meer wisselwerking is het beter een Vector Auto-Regressie (VAR) model te kiezen.

Vervolgens moet worden aangegeven hoeveel tijdssprongen beschreven in het model de werkelijkheid het dichtst benadert. Dat bepalen we met een autocorrelatiefunctie.

Autocorrelatie

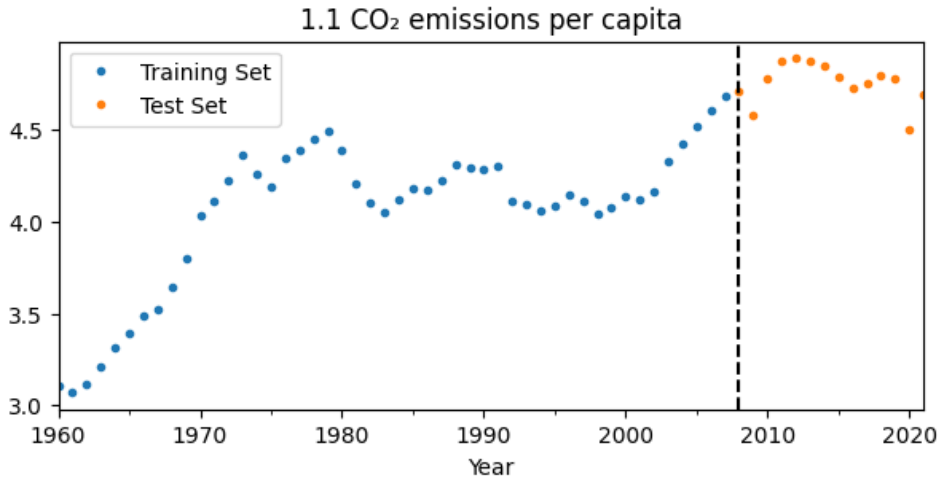
De autocorrelatiefunctie (ACF) kwantificeert de correlatie tussen een tijdreeks en versies van zichzelf in het verleden. In een ACF-plot kunnen we per periode in het verleden de correlatie met het heden zien en vergelijken. Zo bepalen we het aantal antecedente variabelen in het model.

De partiële autocorrelatiefunctie (PACF) berekent dezelfde correlatie maar dan beperkt tot unieke correlaties, het directe effect, zonder inmenging van tussenliggende correlaties. Om het aantal tijdverschillen vast te stellen in een AR model, maken we ook een diagram met daarin de uitkomst van de PACF.

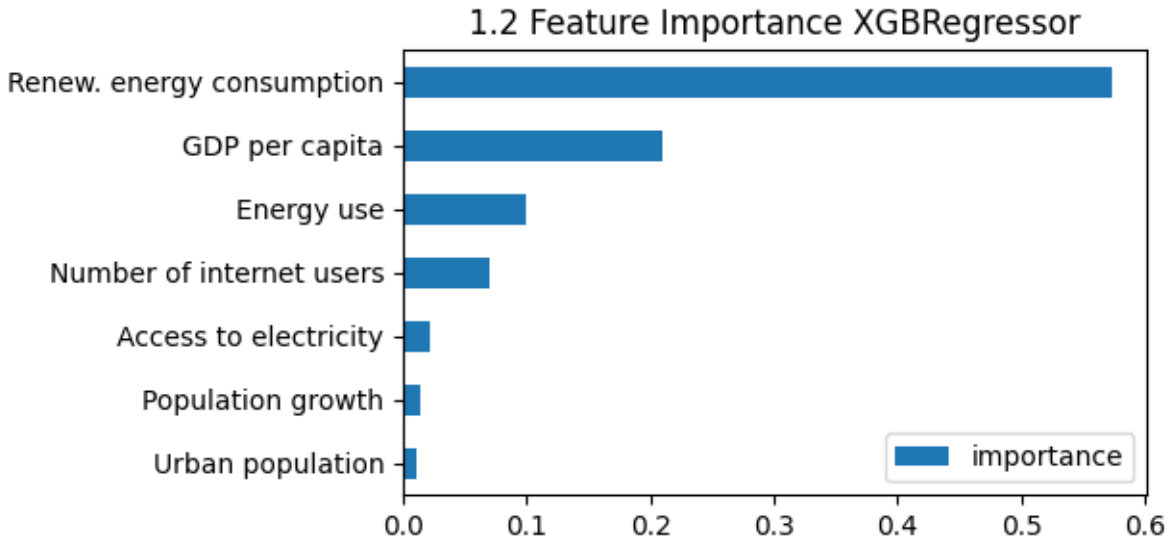
IV. Resultaten

Hier volgen de uitkomsten en bevindingen.

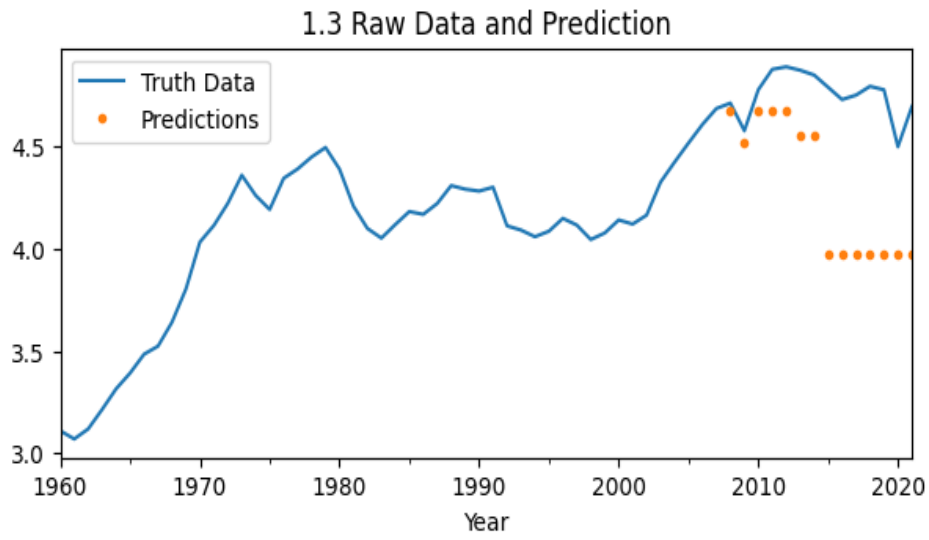
1. Voorspelfactor voor de CO_2 uitstoot van een land



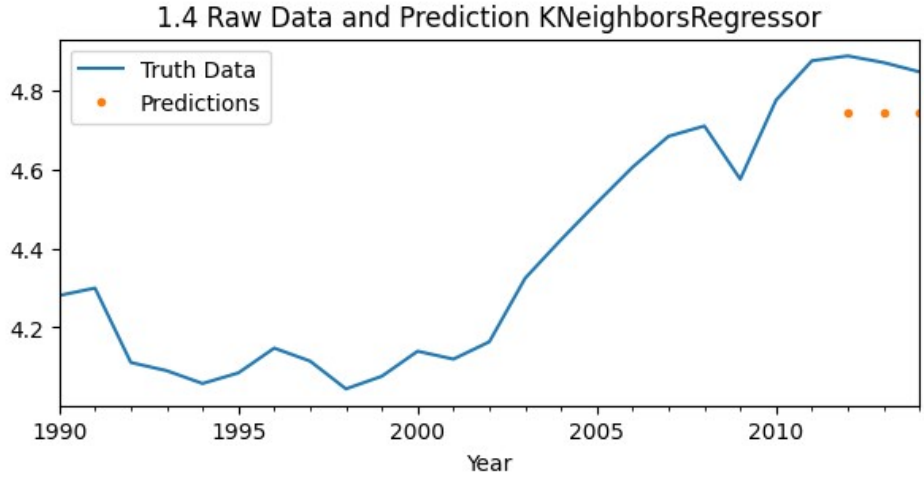
In 1.1 een weergave van het wereldtotaal CO_2 uitstoot per hoofd van de bevolking met de scheidslijn tussen train en test duur, die ik heb gesteld op 2008.



In 1.2 zien we per factor in de CO_2 uitstoot het relatieve aandeel na optimalisatie van het XGB-model. Ik vind het opmerkelijk dat populatie er eigenlijk niet toe doet, iets dat ik me niet had gerealiseerd door het beeld dat ik daarover had.

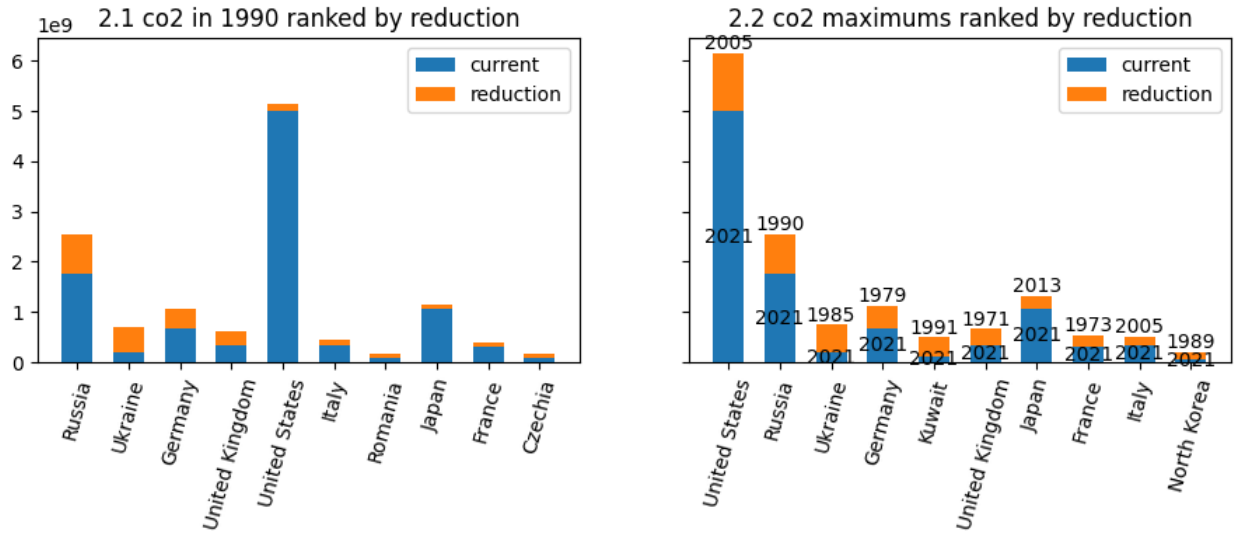


Voor de volledigheid de voorspellingen van de XGB-regressie in 1.3. De laatste 6 zitten er naast, ik denk dat de horizon te ver weg is gekozen, want van het bovenste rijtje voorspellers begint alleen GDP vanaf het jaar 1960.



Dan in 1.4 de voorspelling van de KNN-regressie met de laatste 3 jaar als testvak. Zonder missende waarden en de variabele Toegang to elektriciteit. Het residu of de genormaliseerde afwijking tot 2012 is 0.0511 en de voorspelfout 0.1275 . Een nauwkeuriger resultaat vraagt om een verdere verfijning, sampling zoals cross-validatie kan daarbij helpen. Met cross-validatie bijv. kunnen verschillende maar ook opeenvolgende training sets de uitkomst verbeteren

2. CO_2 reductie



Binnen het kader van het klimaatverdrag van 1992 is in 1997 het Kyoto-protocol overeengekomen en in 2015, het Akkoord van Parijs. Volgens de afspraken in het klimaatverdrag moest de uitstoot van broeikasgassen van Annex I staten in 2000 zijn teruggedrongen tot het niveau van 1990[2].

Om inzicht te krijgen in de uitstoot per land heb ik eerst 2 subvragen geformuleerd:

2.1 Welk land heeft t.o.v. 1990 de meeste CO_2 bespaard?

2.2 Welk land heeft t.o.v. het maximum in jaar x de meeste CO_2 bespaard?

Verder gesorteerd op maximale uitstoot en geanalyseerd welke van de 10 landen met de grootste CO_2 uitstoot het meest hebben gedaan in het licht van eerdere afspraak.

Figuur 2.1

In figuur 2.1 staat de besparing van CO_2 in 2021 t.o.v. 1990. Deze landen hebben dus conform het verdrag CO_2 bespaard. Daarnaast heb ik dus gekeken naar de situatie t.o.v. maximale emissie in figuur 2.2.

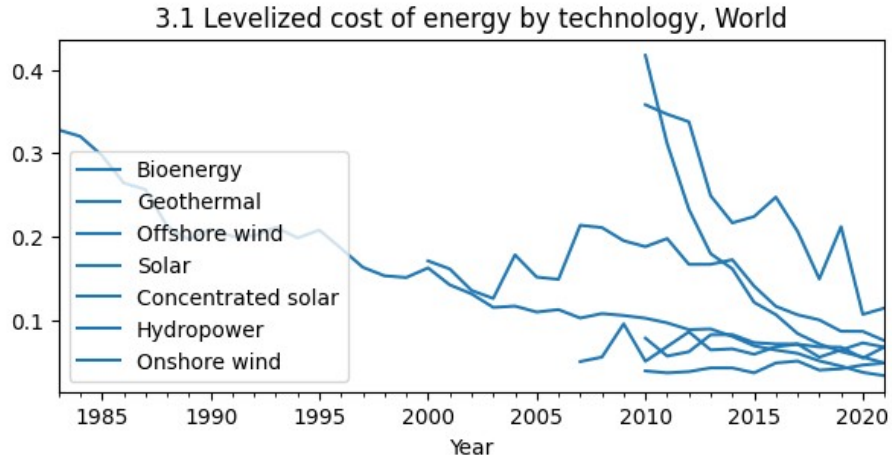
Figuur 2.2

Dan zien we jaartallen staan later dan 1990, wat een slecht teken is, de jaartallen vóór 1990 laten zien dat landen op de goede weg waren gezien de klimaatdoelen. De VS is rijkelijk laat begonnen met besparing met andere woorden dit land heeft in de laatste 20 jaar het probleem verergerd. Italië idem en in Koeweit is hetzelfde gebeurd in de jaren 90 maar dat is veranderd met de Golfoorlog. Alhoewel emissie van de oliebranden is meegenomen in het totaal voor dat jaar. Frankrijk was al op de goede weg, sinds 1973 is de

2 https://nl.wikipedia.org/wiki/Klimaatverdrag

uitstoot met meer dan 232 miljoen ton afgenomen, dat is bijna evenveel als de 255 miljoen ton die het Verenigd Koninkrijk links sinds 1990 de atmosfeer in bracht.. De reden is dat Frankrijk heeft in het verleden ingezet op kernenergie.

3. Prijs van duurzame energie technologie



Om met een regressie gegevens te voorspellen is het noodzakelijk dat data stationair is. Dus dat het gemiddelde niet beweegt en geen sprake is van cyclische patronen. In de lijngrafiek 3.1 is te zien dat de gemiddeldes dalen. Een cyclisch patroon zie ik ook niet en aangezien het om jaarmetingen gaat is het ook niet mogelijk verder in te zoomen op een trend. Of aan deze voorwaarden is voldaan heb ik verder bekeken met de ADF methode.

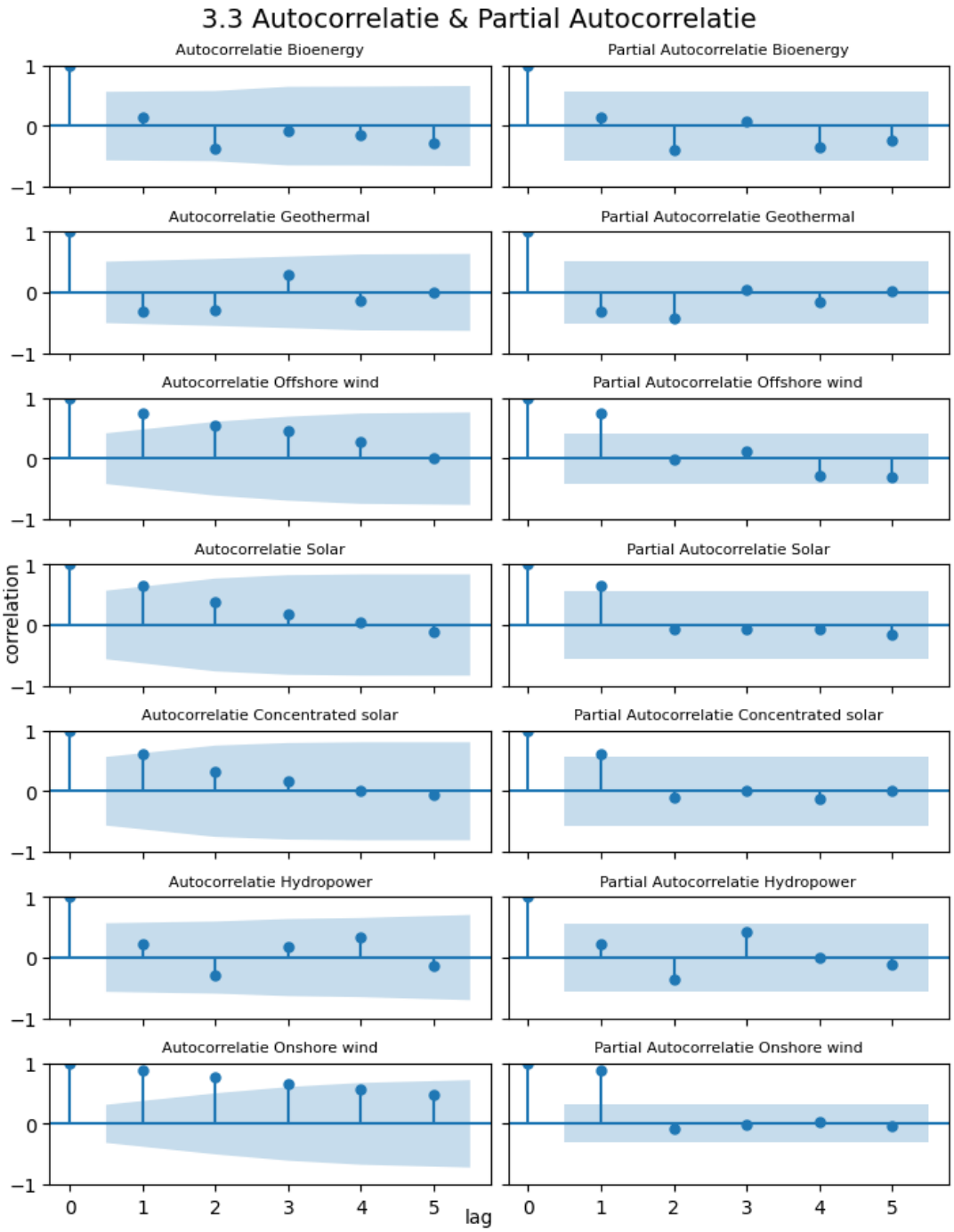
3.2 Stationariteit

Energy	Bioenergy	Geothermal	Offshore wind	Solar	Conc. solar	Hydropower	Onshore wind
P-value	0.863327	0.830764	0.709018	0.863327	0.863327	0.863327	-1.261024

ADF test

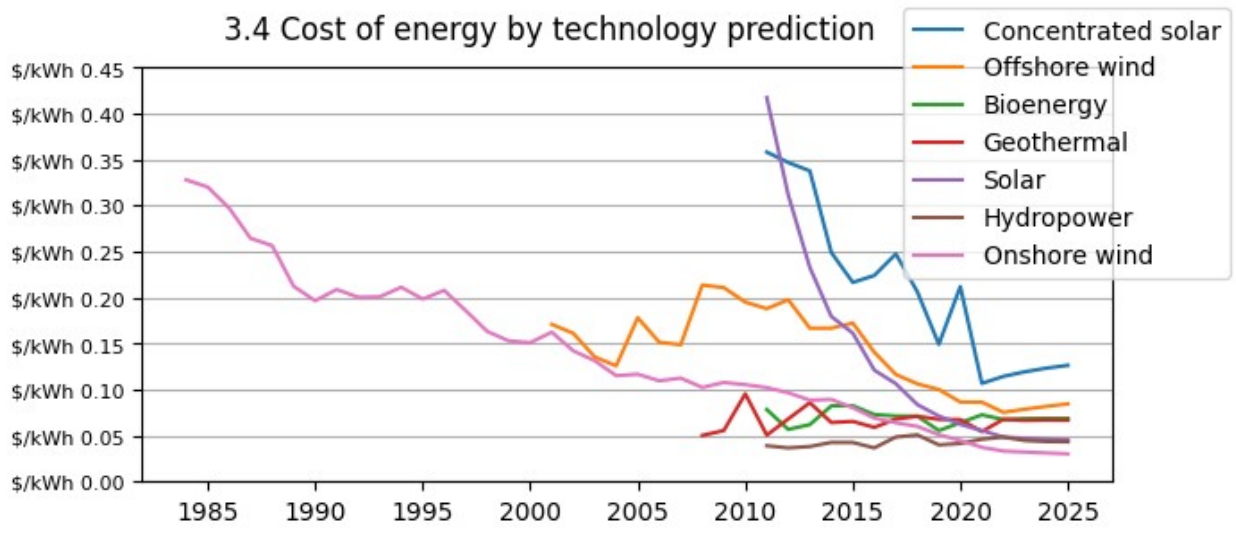
De ADF test heeft controleert of de variantie van de schattingsfouten van de kleinste kwadraten methode constant is, zoniet dan is sprake van heteroscedasticiteit en non-stationariteit. Dat betekent dat de data wel tijdsafhankelijk is.

Als criterium voor de ADF test geldt dat bij een betrouwbaarheidsinterval van 95% de p-waarde lager is dan 0.05 en dat is voor deze reeksen niet het geval, zie tabel 3.2. Met een betrouwbaarheidsinterval kunnen we vaststellen of een meting valt buiten statistische grenzen van meetonzekerheid van de waarnemingen, 95% in dit geval, aangegeven in een standaardspreiding rond het gemiddelde.



Aan de linkerkant zien we de ACF per energievorm met uitkomsten die redelijk verschillen. Hoe meer correlaties significant verschillen van 0 hoe minder willekeurig de tijdreeks en hoe duidelijker een trend en patronen zijn. Rechts is 4 van de 7 keer een tijdsverplaatsing significant gecorreleerd met Y(o). Voor

Y(t-2) zijn 3 correlaties negatief maar niet significant. Dus kies ik in het AR model Y(t-1). Ook wil ik de reeksen niet separaat inregelen en bijstellen want ik moet tenslotte een vergelijking maken.



Grafiek 3.4 toont de voorspellingen voor 2022, 2023 en 2024 en Y(t-1) als verklarende variabele. We zien kleine stijgingen voor 2024 van de bovenste twee lijnen en dat past wel bij het patroon; een dalende trend bij een klein aantal datapunten. Wanneer ik Y(t-2) toevoeg blijft de volgorde in prijs dezelfde, echter de lijnen in de grafiek lopen iets steiler naar beneden, te steil in mijn beleving.

Het gaat er ook om de relevante variabelen te vinden. Teveel features kan leiden tot overfitting, een perfect model, vooral als deze onderling goed correleren, ofwel sprake is van multicollineariteit. Jammer dat het aantal observaties gering is en ongelijk dan zou opvolgende steekproefneming en evaluatie mogelijk zijn.

V. Conclusie

Wat is de belangrijkste voorspelfactor voor de CO_2 uitstoot van een land?

De belangrijkste voorspelfactor voor de CO_2 uitstoot van een land is consumptie van duurzame energie. Dat kan an sich geen verrassing zijn want wie geen fossiele brandstof gebruikt en groene stroom draagt bij aan de oplossing voor het klimaatprobleem veroorzaakt door CO_2. Ik denk tevens dat alternatieve meer specifieke aspecten kunnen zijn informatie over vervoer, de logistieke sector en de, staal- en cementconsumptie van een land.

Welke landen boeken de meeste vooruitgang de uitstoot omlaag te brengen?

Rusland, Oekraïne, Duitsland, VK en de VS hebben de meeste uitstoot bespaard, met de kanttekening dat de laatste pas in 2005 de omslag heeft gemaakt.

Welke duurzame energie technologie heeft de laagste prijs in de toekomst?

Om toekomst te voorspellen moet je het heden kennen en het verleden. De vraag is hoever terug in het verleden terug te kijken. Gehouden aan de grenzen van het bereik van de data en de normen voor interpretatie van de gekozen aanpak meen ik dat Onshore wind de laagste prijs heeft in de toekomst.