Norges miljø- og biovitenskapelige universitet

Masteroppgave 2022 30 stp
Fakultet for realfag og teknologi

Predikere antallet brukere av nye sykkelveier i Oslo basert på posisjonsdata fra Strava: en anvendelse av GIS og regresjonsanalyse

Predicting the number of users of new bicycle infrastructure in Oslo based on location data from Strava: an application of GIS and regression analysis

Olav Vikøren Espenes
Geomatikk

Forord

Denne masteroppgaven marker avslutningen på mine fem år ved Norges miljø- og biovitenskapelige universitet (NMBU). Oppgaven har et omfang på 30 studiepoeng og ble skrevet våren 2022.

Takk til veileder Ivar Maalen-Johansen for veiledning, innspill og råd i forbindelse med oppgaven, samt gode samtaler hver gang jeg var innom ditt kontor.

Stor takk til familie og kjæreste som har bidratt med korrekturlesing og forslag til forbedring av oppgaven. Takk for at dere har vist tålmodighet i perioder der oppgaven har krevd mye av min tid.

Vil også takk Rer Roy Laudal, overingeniør i Bymiljøetaten, for hjelp i forbindelse med nedlastning av data fra sykkeltellerportalen til Oslo kommune.

Avslutningsvis vil jeg rette en stor takk til gutta i Eplehagen 25B, gjengen på masterrommet og til andre medstudenter som har bidratt til å gi meg fem uforglemmelige år på Ås.

Olav Vikøren Espenes
Ås, 15. juni 2022

Sammendrag

Denne masteroppgaven ser på muligheten for å benytte posisjonsdata fra treningsapplikasjonen Strava til å undersøke hvorvidt oppgradering av sykkelinfrastruktur i Oslo har en innvirkning på antallet syklister på strekningen. Dette er gjort ved å sammenligne tall fra Strava med tall fra sykkeltellere for så å studere likheten mellom disse. På denne måten kan bruken og nytten av nye sykkelveier evalueres gjennom modeller som benytter data fra nettdugnad som input.

Data fra mai måned i årene 2018-2021 for 37 ulike sykkeltellere i Oslo med tilhørende data fra Strava blir benyttet som datasett i en regresjonsanalyse. Datasettet består av verdier fra daglig registreringer av syklister og aggregeres på et månedlig nivå. I første omgang studeres korrelasjonen mellom data fra sykkeltellerne og Strava for de fire årene. Deretter estimeres to modeller fra regresjonsanalysen; en ordinær minste kvadraters modell og en kvadratrot-transformert modell. Begge modellene blir brukt til å predikere antallet syklister i et område før og etter nye sykkelfelt ble innført.

Resultater fra korrelasjonsberegninger viser at det er varierende, men generell høy korrelasjon mellom tall fra Strava og sykkeltellere ved månedlig aggregering. Oppgavens to modeller er til en viss grad undertilpasset. Av den grunn predikerer ikke modellene tilstrekkelig høye eller lave nok verdier til at det gjenspeiler den sanne variasjonen i det totale antall syklister mellom ulike måneder. Modellene greier i enkelte tilfeller likevel å gjenspeile den samme utviklingen som tallene fra sykkeltellerne viser. Med utviklingen menes trenden i om volumet av syklister øker eller minker i et område.

Abstract

This master's thesis looks at the possibility of using loaction data from the training application Strava to examine whether upgrading the cycling infrastructure in Oslo has an impact on the number of cyclists. This is done by comparing data from Strava with data from bicycle counters and then studying the similarities between these. In this way, the use and usefulness of new cycle paths can be evaluated through models that use data from crowdsoursing as input.

Data from May in the years 2018-2021 for 37 different bicycle counters in Oslo with associated data from Strava are used as dataset in a regression analysis. The dataset consists of values from daily registrations of cyclists and are aggregated on a monthly level. In the first half, the correlation between data from the bicycle counters and Strava for the four years is studied. Then, two models from the regression analysis are estimated; an ordinary least squares model and a square root-transformed model. Both models are used to predict the number of cyclists in an area before and after new bicyclelanes were introduced.

Results from correlation calculations show that there is a varying, but generally high correlation between data from Strava and bicycle counters for monthly aggregation. The two models in the thesis are to a certain extent underfitted. For this reason, the models do not predict sufficiently high or low enough values to reflect the true variation in the total number of cyclists between different months. In some cases, however, the models manage to reflect the same development as the data from the bicycle counters show. By development is meant the trend in whether the volume of cyclists increases or decreases in an area.

INNHOLD

FIGURER

TABELLER

KAPITTEL

1

INTRODUKSJON

I kapittelet *Introduksjon* presenteres motivasjonen bak oppgaven. Videre presenteres forskningsspørsmålene som ligger til grunn, før tidligere studier om tematikken omtales.

1.1 Bakgrunn og motivasjon

Noe av det som kjennetegner en god by er hvor enkelt det er for innbyggere og tilreisende å forflytte seg dit de vil. I løpet av de siste 100 årene med byutvikling har synet på hva som er den beste transportløsningen variert mye. Fram til 50-tallet var gåing, sykling og offentlig transport det vanligste fremkomstmiddelet i byene. Når de første bilene begynte å dukke opp måtte disse tilpasse seg den allerede etablerte bystrukturen. Etter hvert som bilen ble allemannseie skulle det legges til rette for at den raskeste måten å transportere seg rundt i byene på var ved bruk av bil. Dette førte til en mindre effektiv fremkommelighet for syklister og korrektivtransport. I dagens byplanlegging er derimot gåing, sykling og kollektivtransport igjen den prioriterte løsningen når mennesker skal forflytte seg. Dette skyldes at fokuset i dag er rettet mot å motivere til bruk av bærekraftige transportløsninger i størst mulig grad.

Oslo er en by hvor det blir satset stort på å legge til rette for sykling som fremkomstmiddel. I 2015 var den totale lengden på sykkelveinettet for alle bydelene i Oslo sett under ett på omtrent 180 km. I "Plan for sykkelveinettet i Oslo" [1] er ett av målene at sykkelveinettet skal utgjøre totalt 280 km innen 2025. Dersom dette målet oppnås vil 64 % av Oslos befolkning være bosatt innenfor en radius på 200 meter fra nærmeste del av sykkelveinettet.

Utbyggingen av sykkelveier i Oslo skjer ikke uten bekostning av annet areal. I mange tilfeller er det kollektivfelt og parkeringsplasser som må vike ved oppføring av nye sykkelfelt. Dette har skapt trafikale utfordringer flere steder i byen og konfliktnivået er som følge av dette høyt enkelte steder. To omdiskuterte sykkelveier fra de seneste årene er sykkelfeltene i Gyldenløvs gate og Bygdøy som oppslagene i [2] omtaler. I en holdningsundersøkelse Opinion gjennomført i 2020 svarte 74 % av de spurte som syklet ofte eller alltid at de syklet uavhengig av standarden på sykkelveinettet. Det bevilges årlig store summer til oppgradering av infrastruktur for syklister. For 2022 har byrådet i Oslo kommune lagt fra et budsjettforslag om å bruke 673 millioner til dette formålet [3].

NY SYKKELVEI: Her legges det ny asfalt for den nye sykkelveien i Gyldenløves gate. Foto: Mathilde Walberg Hofseth

Koker i kommentarfelt etter ny sykkelvei i Gyldenløves gate

KØ: Bygdøy allé er en av de verste gatene i Oslo når det gjelder fremkommelighet, ifølge Ruter. FOTO: Ida Laingen

Sykkelfeltene i Bygdøy allé skaper kø og kaos

Figur 1.1: Avisartikler fra Norgesnytt [4] [2] som omtaler utfordringer knyttet til oppføring av sykkelfelt i Gyldenløvs gate og Bygdøy allé.

For å forsvare beslaget nye sykkelveier legger på eksisterende areal, samt de store økonomiske investeringene må den positive effekten av satsingen dokumenteres. Som en sentral del i å måle effekten av utbyggingen av sykkelveier, ble det i perioden 2013 til 2017 installert 39 nye sykkeltellere. Sykkelregnskapet viser at det gikk fra 7 tellere i 2013 til 46 i 2017 [5]. Oslo kommune har i perioden mellom 2017 og 2022 hatt ansvaret for disse 46 sykkeltellerne som har registrert passeringer i hele eller deler av perioden. Statens vegvesen har i samme periode hatt 11 operative sykkeltellere i Oslo.

Sykkeltellere leverer god statistikk på volumet av syklister der den er installert, men gir liten eller ingen informasjon om sykkeltrafikken utover dette. Den lave geografiske representasjonen gjør at det må installeres flere sykkeltellere i et område for å måle endringer i rutevalg blant syklister. I et økonomisk perspektiv er oppføring og drift av sykkeltellere også kostbart.

I en rapport fra NINA[1] [6] ble frivillig innsamlede geografiske data i treningsapplikasjoner presentert som en metode for å måle fysisk aktivitet i befolkningen. Her ble Strava presentert som en av aktørene med størst database bestående av slik type data. Ideen til denne masteroppgaven har sitt utspring i denne rapporten og fra behovet etter en bedre måte å evaluere bruken av nye sykkelveier.

[1] Norsk institutt for naturforskning

1.2 Forskningsspørsmål

Målet med denne masteroppgaven er å undersøke muligheten for å benytte posisjonsdata samlet inn fra brukere av Strava til å predikere den totale mengden syklister i ulike områder. På den måten kan en modell i teorien predikere de verdiene som en sykkelteller måler. Mest interessant er det å studere innvirkningen nyoppførte sykkelveier har på tilkomsten av flere syklister. I motsetning til en stasjonær sykkelteller kan geografisk data fra Strava brukes til å analysere mer enn bare antall syklister for en gitt del av en vei. Det er likevel relasjonen mellom data fra sykkeltellerne og data fra Strava det er viktig å få målsatt før beregning og testing av modeller kan begynne. Som en naturlig del i prosessen mot målet for oppgaven er følgende forskningsspørsmål konkretisert:

1. Er det samvariasjon mellom antall registrerte sykkelturer i Strava og antall syklister totalt for et område?

2. Er det mulig å anvende data fra Strava til å predikere det totale antall syklister som tar i bruk nye sykkelveier i Oslo?

1.3 Tidligere studier

Data fra Strava har blitt anvendt og undersøkt i flere tidligere studier. Noen studier ser utelukkende på korrelasjonen mellom telledata og dataene fra Strava. Andre studier estimerer modeller som de så anvender i forbindelse med evaluering av infrastruktur for syklister.

Boss mfl. [7] påviste korrelasjon mellom data fra 11 sykkeltellere og data fra Strava. Sykkeltellerene målte med 15 minutters intervall fra 6 om morgenen til 8 om kvelden. Det ble beregnet korrelasjonsverdier mellom de to datasettene på 0,76 til 0,96.

Strava data fra 2016 ble i Roy mfl. [8] satt opp imot tellerdata fra 44 ulike lokasjoner i Maricopa County, Arizona, USA. Datasettet besto av åtte sammenhengende to-ukers perioder samlet inn i månedene april, mai, oktober og november. Høyeste R^2-verdi ble oppnådd ved bruk av en enkel lineær regresjonsmodell og var på 0,76.

Hong, McArthur og Livingston [9] ville i sin studie undersøke om de kunne benytte data fra Strava datert årene 2013 til 2016 for å undersøke ny infrastruktur for syklister hadde ført til økt volum på aktuelle rutene. Resultater viste at tre av fire ruter fikk en positivt økning i antall syklister etter oppgraderingen sammenlignet med før. Økningen var på 12 til 18 prosent. Studien poengterte samtidig at data fra Strava måtte brukes med forsiktighet og at resultater fra estimert modeller må brukes med omhu.

Heesch og Langdon [10] benyttet posisjonsdata fra smarttelefoner til å evaluere betydningen av ny infrastruktur for sykkel. Varmekart og graderte volumkart ble gjennomgått, og posisjonsdata ble sammenlignet med telledata.

Da forskningsspørsmålene for denne masteroppgaven ble utarbeidet var det ingen studier som hadde undersøkt relasjonen mellom data fra Strava og sykkeltellere i Oslo.

1.4 Oppgavens struktur

Oppgaven er strukturert på følgende vis:

- Kapittel 1: Introduksjon
- Kapittel 2: Bakgrunn
- Kapittel 3: Metode
- Kapittel 4: Resultater
- Kapittel 5: Diskusjon
- Kapittel 6: Konklusjon

KAPITTEL

2

BAKGRUNN

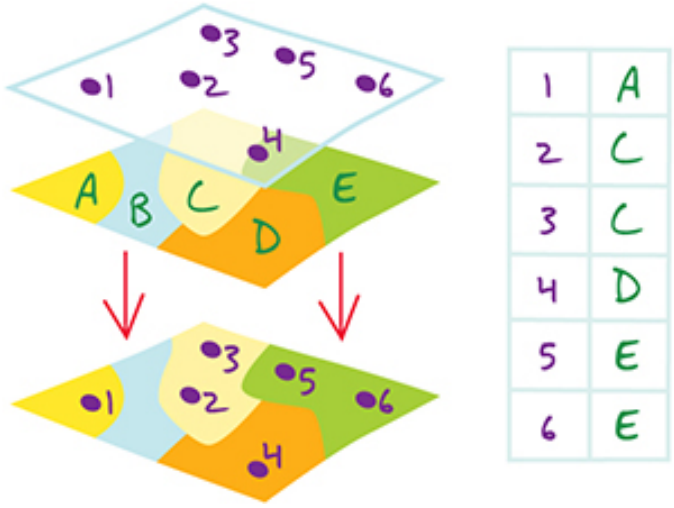
Kapittelet *Bakgrunn* inneholder teori som beskriver verktøy og analyser som er benyttet i kapittelet *Metode*. Her er også prosessen med innhenting fra datakildene beskrevet, samt nøyaktighetsmål og begrensninger tilknyttet disse. I teoridelen er hovedfokuset på GIS og statistikk.

2.1 Teori

2.1.1 GIS

2.1.1.1 Romlig kobling

Romlig kobling, også kjent som *spatial join* eller *overlay*, er en mye brukt geografisk analysemetode i GIS. Metoden går ut på å koble ulike romlige datasett for å hente ut og undersøke felles informasjon. På den måten kan objekter i et datasett tilføres ytterligere geografisk informasjon. I denne studien er det benyttet punkt-i-polygon overlay hvor et datasett med punkter er koblet med et datasett med polygoner basert på felles geografisk tilknytning.



Figur 2.1: Punkt-i-polygon overlay. Illustrasjon fra Esri [11].

2.1.1.2 Klassifisering med naturlige inndelinger (Jenks)

Naturlig inndelinger (engelsk: natural breaks) er en univariat[1] klassifiseringsmetode som deler inn klassene etter naturlige grupperinger i datasettet. Baser på antall klasseinndelinger grupperer metoden verdier som er relativt like, og maksimerer intervallet mellom de ulike klassene [12]. Dette genererer klasser som representerer et ulikt antall verdier. Klassifiseringsmetoden for naturlige inndelinger som benyttes i ArcGIS Pro er basert på "Jenks Natural Breaks algorithm". Metoden for algoritmen er beskrevet under *Univariate classification schemes* i *Geospatial Analysis—A Comprehensive Guide, 6th edition* [13] og består av følgende fire steg:

[1] Angår bare én variabel.

1. Velger hvilken variabel, x, som skal klassifiseres samt antallet klasser k.

2. Et utvalg på $k-1$ tilfeldige eller like verdier blir valgt blant alle verdier av x. Disse verdiene representerer midlertidige klassegrenser.

3. Gjennomsnittsverdien og standardavviket innad blant verdiene i hver klasse beregnes. Også summen av standardavvikene fra samtlige klasser (TSSD) beregnes.

4. Verdier i nedre og øvre del av klassen blir deretter flyttet til naboklassene gjennom justering av klassegrensene. TSSD blir deretter sjekket for å se om verdien øker eller synker. Prosessen itereres flere ganger helt til TSSD er under en gitt verdi, eller at variansen innad i klassene er så lav som mulig og variansen mellom klassene er så høy som mulig. Dette er ingen optimaliseringsmetode, men steg 1 og 2 kan gjentas for deretter å sammenligne TSSD-verdier.

Ved bruk av naturlig inndeling er det viktig å kjenne til den innbyrdes spredningen i verdiene i datasettet, og studere hvordan disse grupperer seg ved valg av antallet klasser. Koroplekart basert på klasser som har en naturlig inndeling egner seg ikke til sammenligning med andre tematiske kart hvor det er brukt et annet datagrunnlag [12].

2.1.2 Statistikk

2.1.2.1 Median og gjennomsnitt

Medianen er den midterste verdien i en tallrekke som er sortert i stigende rekkefølge. Ved odde antall tallverdier er den midterste verdien medianverdien. I tilfeller der antallet observasjoner i tallrekken er partall, blir medianen lik gjennomsnittsverdien av de to midterste tallene [14].

Gjennomsnitt angir den verdien i en tallrekke som er "mest typisk". Det finnes ulike måter å estimere gjennomsnittet, men den vanligste metoden er *artimetisk gjennomsnitt* [15]. Denne formen for gjennomsnitt estimeres ved å ta summen av alle verdier i tallrekken og dele på antallet verdier i samme rekke. Formelen for å estimere gjennomsnittet ($\bar{x}$) for en tallrekke med n verdier er:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} \tag{2.1}$$

Gjennomsnitt omtales også som middeltall eller middel. I forbindelse med undersøkelse av en tallrekke benyttes både medianen og gjennomsnittet. Dette er fordi verdiene hver for seg, og ved sammenligning, forteller mye om balansen i tallrekken. Fordelen med median kontra gjennomsnitt er at medianverdien ikke lar seg påvirke av ekstremverdier i like stor grad [14].

2.1.2.2 Varians og standardavvik

Varians og standardavvik er mål på spredningen for verdiene i et datasett. Mellom gjennomsnittsverdien for datasettet og hver enkelt observasjon er det et avvik. Ved å danne et kvadrat (avvikskvadratet) med sidelengder på størrelse med dette avviket for hver enkelt observasjon, kan variansen og standardavviket beregnes. Variansen er arealet på gjennomsnittskvadratet som beregnes ut ifra alle avvikskvadratene fra samtlige observasjoner i datasettet. Standardavviket er kvadratroten av variansen og representerer dermed kantlengden på det gjennomsnittlige avvikskvadratet. Ved stor spredning blant verdiene i datasettet, vil observasjonene ha en større differanse fra gjennomsnittsverdien og dette vil føre til kvadrater med lengre sidelengder. Dette fører igjen til at verdien på variansen og standardavviket blir høyere. I tilfeller der observasjonene i datasettet er tettere på gjennomsnittsverdien vil de målene ha en mindre verdi [14].

2.1.2.3 Pearsons korrelasjonskoeffisient

Pearsons korrelasjonskoeffisient, kjent som Pearson produkt-moment korrelasjonskoeffisient, er et kvantitativt mål på det lineære forholdet mellom to variabler, x og y. Som et mål på korrelasjon og samvariasjon blir ofte Pearsons korrelasjonskoeffisient benyttet. Verdiene for korrelasjonskoeffisienten spenner fra -1 som tilsvarer perfekt negativ korrelasjon, via 0 som tilsier ingen lineær relasjon mellom de to variablene, og til 1 som er perfekt positiv korrelasjon. Ved perfekt korrelasjon ligger alle observasjonene på en rett linje. Er koeffisienten positiv, øker y når x øker. Er den derimot negativ, synker y når x øker [16]. Benevningen for Pearsons korrelasjonskoeffisient er r.

$$r = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sqrt{SS_{xx}}} \frac{(y_i - \bar{y})}{\sqrt{SS_{yy}}} = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}} \tag{2.2}$$

$$der \quad SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2,\ SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2,\ SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

SS_{xx} og SS_{yy} er mål på de totale variasjonene i datasettet. Dette gjøres ved å summere arealet av avvikskvadratene mellom x- og y-verdier for alle observasjonen, og den tilhørende gjennomsnittsverdien, henholdsvis $\bar{x}$ og $\bar{y}$.

2.1.2.4 Regresjonsanalyse

Regresjonsanalyse er samlebetegnelsen på metoder anvendt i statistikken for å beskrive en variabel basert på én eller flere andre variabler [17]. Verktøyet som benyttes er funksjoner (regresjonsmodeller) som forsøker å fremstille en mest mulig virkelighetskorrekt beskrivelse av responsvariabel "y". Denne variabelen er i alle regresjonsmodeller avhengig av to parametere: den forventede verdien av y (*E(y)*) og summen av de tilfeldige, uforklarlige feilene (ε) [16].

$$y = E(y) + \varepsilon \tag{2.3}$$

Parameteren *E(y)* består av en eller flere uavhengige variabler som utgjør modellens grunnlag for å predikere y. Disse variablene har notasjonen x_1, x_2, x_3, osv. I tilfeller der y predikeres basert på kun én predikasjonsvariabel x er det benyttet en **enkel** lineær regresjonsmodell. Disse modellene er alltid lineære ettersom *E(y)* kun består av konstantleddet β_0 som er punktet på modellen skjærer y-aksen, og β_1 som definerer verdien y øker eller synker med for hver økning i x. Alle modeller bestående av mer enn én predikerende variabel omtales som **multiple** regresjonsmodeller. Multiple regresjonsmodeller kan også bestå av interaksjonsledd (produktet av to uavhengige variabler x), samt opptre som et polynom med høyere ordens ledd. Den generelle formen for multiple regresjonsmodeller hvor y er forklart basert på k predikerende variabler er vist i likning 2.4.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \tag{2.4}$$

I et datasett hvor det ikke forekommer perfekt korrelasjon, vil selv ikke den beste regresjonsmodellen utledet fra dataene representere disse perfekt. Variabelen ε inngår som et ledd for å representerer avviket mellom modellen og datasettet. Målefeil og manglende variabler kan være noen av de uforklarlige faktorene som inngår i dette avviket. For at en regresjonsmodell skal være gyldig må derfor følgende antagelser fra *A second course in statistics: regression analysis* [16] side 110 for ε være gjeldende:

1. Gjennomsnittsverdien av sannsynlighetsfordelingen til ε er 0.
2. Variansen for sannsynlighetsfordelingen til ε er konstant for alle verdier av den uavhengige variabel x. Dette kan skrives som $Var(\varepsilon) = \sigma^2$, hvor σ^2 er konstanten.
3. Sannsynligheten for verdien av ε er normalfordelt for hver x-verdi.

4. De ulike feilene er uavhengige av hverandre. ε for en predikert verdi av y er uavhengig fra andre predikerte verdier av y.

Gyldigeheten av antagelsene skal alltid vurderes i forbindelse med utledning av en regresjonsmodell. Dette gjøres ved å analysere residualene og vurdere disse opp imot antagelsene. Residualer ($\hat{\varepsilon}$ eller e) er regresjonsmodellens predikasjon på ε og er avstanden, gitt i y-verdi, mellom modellen og observasjonen. En residualanalyse kan avdekke mangel på gyldighet av antagelsene gjennom feil knyttet til modellvalg og ekstremverdier i datasettet som kan skyldes feil i selve målingen, eller under loggføring av data.

2.1.2.5 Middelkvadrat feil og rot-middelkvadrat feil

Middelkvadrat feilen (engelsk: mean square error (MSE)) er variansen til regresjonsmodellen og dermed den estimerte variansen for feilleddet ε. Som et estimat på den sanne variansen (σ^2) mellom modellen og populasjonen, benyttes summen av avvikskvadratene for observasjonene i datasettet modellen er estimert ut ifra. Summen av avvikskvadratene (engelsk: sum of squares error (SSE)) beregnes ved å summere de kvadrerte verdiene av residualene:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.5}$$

Når SSE er bestemt, kan middelkvadrat feilen ($\hat{\sigma}^2$ eller s^2) beregnes ved å benytte likning (2.6) hvor n er antallet observasjoner i datasettet.

$$MSE = s^2 = \frac{SSE}{n - Antallet\ estimerte\ \beta\ parametere} \tag{2.6}$$

Fra middelkvadrat feil kan rot-middelkvadrat feilen (engelsk: root mean square error (RMSE)) utledes ved å ta kvadratroten av MSE som vist i likning 2.7. RMSE er det estimerte standardavviket (s) til regresjonsmodellen og gir ofte en mer meningsfull beskrivelsene av usikkerheten. Intervallet $\pm 2s$ fungerer som et nøyaktighetsmål for modellen. Predikerte verdier av y er forventet å falle innenfor dette intervallet [16].

$$RMSE = \sqrt{s^2} = \sqrt{MSE} \tag{2.7}$$

2.1.2.6 Minste kvadraters metode

Minste kvadraters metode (engelsk: Ordinary Least Squares (OLS)) er en mye brukt metode for å tilpasse enkle og multiple lineære regresjonsmodeller til et datasett. Modellen estimeres basert på prinsippet om å minimere summen av alle kvadratene (SSE) mellom observasjonene og den tilpassede modellen. OLS vil på bakgrunn av dette prinsippet tilpasse den lineære modellen best mulig til observasjonene i datasettet [16].

2.1.2.7 Forklaringskraften R^2

Forklaringskraften R^2 (engelsk: Coefficients of Determination) er et mål på hvor godt regresjonsmodellen er tilpasset observasjonene i datasettet. R^2 er en verdi mellom 0 og 1 som forklarer hvor stor andel av den totale variasjonen blant observasjonene i datasettet som forklares av modellen. Er $R^2 = 0.8$ forklarer modellen 80 % av den totale variasjonen mellom observasjonene. Ved bruk av R^2 som mål for nytten[1] av modellen er det viktig at det er betydelig flere observasjoner enn β-parametere i modellen.

$$R^2 = 1 - \frac{SSE}{SS_{yy}} \tag{2.8}$$

Et alternativ til R^2 er å benytte *justert* R^2. Justert R^2 tar høyde for størrelsen på datasettet og antallet β-parametere i modellen. Justert R^2 vil alltid ha en lavere verdi enn R^2. Fordelen med justert R^2 er at den aldri kan "tvinges" til være lik 1 ved å legge til ytterligere uavhengige variabler til modellen. Justert R^2 benyttes av den grunn ofterer enn R^2 i forbindelse med modellvurdering.

$$R_a^2 = 1 - \left[\frac{(n-1)}{n - Antallet\ estimerte\ \beta\ parametere}\right] \times (1 - R^2) \tag{2.9}$$

Nytten av modellen burde ikke kun vurderes basert på R^2. En global F-test av alle β-parametere i modellen bør gjennomføres på forhånd for å bekrefte modellens statistiske nytte. Dette gjelder først og fremst multiple regresjonsmodeller med mange β-parametere. Deretter kan R_a^2 brukes til å tolke hvor mye av variansen som fanges opp av samme modell [16].

2.1.2.8 Kvalitative variable

I multiple regresjonsmodeller kan kvalitative (kategoriske) uavhengige variabler inngå for at modellen skal kunne ta hensyn til forhold som vanskelig lar seg inndele ut ifra en

[1] Hvor godt responsvariabelen y forklares av de uavhengige variablene.

numerisk skala. Ved å tilføre to nye binære variabler (0, 1) som kode for om en egenskap er ikke-gjeldende eller gjeldende for en observasjon, kan modellen ta høyde for kjente forskjeller i datasettet som er grunnlaget for modellen. Hvis hver observasjon i datasettet er delt inn i en av tre kategorier, eksempelvis A, B eller C, er det nødvendig med to kvalitative variable, x_1 og x_2. De tre klassene kan representeres i modellen på følgende måte:

$$x_1 = \begin{cases} 1 & \text{if A} \\ 0 & \text{if } not \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if B} \\ 0 & \text{if } not \end{cases}$$

Utgangspunkt = C

De kvalitative variablene x_1 og x_2 går også under betegnelsen *dummy variable* ettersom verdiene variablene har kun representerer en kode og av den grunn er vilkårlig valgt.

2.1.2.9 Kvadratrot transformasjon

I tilfeller hvor antagelsen om konstant varians for ε ikke er innfridd kan dette skyldes at observasjonene i datasettet har en Poisson-fordeling. Data som er Poisson-fordelt fører til lav varians for lave verdier av $\hat{y}$ og økende varians etterhvert som $\hat{y}$ øker. Etterhvert vil økningen i variansen for høyere verdier av $\hat{y}$ avta noe. I slike tilfeller er y en funksjon av størrelsen på $E(y)$ og variansen er proporsjonal med verdien av $E(y)$. For å oppnå konstant varians for et slikt datasett kan en **varians-stabiliserende transformasjon** benyttes. For data som er Poisson-fordelt er det vanlig å sette responsvariabelen y i kvadratrot:

$$y^* = \sqrt{y} \tag{2.10}$$

og deretter tilpasse modellen [16]:

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \tag{2.11}$$

2.1.2.10 F-test

Modellens nytte kan testes ved bruk av F-statistikk. Det er ulike måter å teste modellen på ved bruk av F-statistikk. I denne studien er det benyttet **global** og **delvis** F-test. Ved global F-test testes nytten av alle β-parametere sett under ett. Følgende hypoteser er gjeldende for en slik test av en multippel regresjonsmodell:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_k = 0$$

$$H_a : Minst\ en\ \beta_i \neq 0$$

Verdien for F beregnes for å avgjøre utfallet av hypotesetesten:

$$F = \frac{(SS_{yy} - SSE)/k}{SSE/[n-(k+1)} = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \frac{MSR}{MSE} \tag{2.12}$$

hvor n er antall observasjoner i datasettet og k er antallet uavhengige variabler i modellen. $k+1$ er dermed antallet β-parametere i modellen.

H_0 forkastes dersom $\alpha >$ p-verdien. p-verdien representerer sannsynligheten for at tabellverdien for F på gitt signifikantnivå er større enn den beregnede verdien av F fra likning 2.12. Dersom H_0 forkastes er det 95 % sikkerhet (signifikansnivå $\alpha = 0{,}05$) for at minst én av β-parameterne i modellen er nyttig for å predikere verdien av y. Ettersom F beregnes basert på forholdet mellom variasjonen blant verdier av y som forklares av modellen, MSR[1], og de uforklarlige variasjonene blant verdier av y som ikke forklares av modellen, MSE, så omtales F-tester som "analyse-av-variansen", forkortet ANOVA [16].

I en delvis F-test testes utvalgte β-parameteres bidrag til modellen. En *redusert* modell (2.13) bestående av et utvalg av de uavhengige variablene sammenlignes med den *komplette* modellen (2.14. Det er nytten av variablene som ikke er med i den reduserte modellen som testes i delvis F-test.

$$y = \beta_0 = \beta_1 x_1 + ... + \beta_g x_g + \varepsilon \tag{2.13}$$

$$y = \beta_0 = \beta_1 x_1 + ... + \beta_g x_g + \beta_{g+1} x_{g+1} + ... + \beta_k x_k + \varepsilon \tag{2.14}$$

[1] "Mean square regression" er gjennomsnittet av summen av kvadratene fra differansen mellom $\hat{y}$ og $\bar{y}$.

Følgende hypoteser gjelder for en delvis F-test:

$$H_0 : \beta_{g+1} = \beta_{g+2} = ... = \beta_k = 0$$

$$H_a : Minst\ en\ \beta_j \neq 0\ for\ j > g$$

F-verdien beregnes på følgende måte:

$$F = \frac{(SSE_R - SSE_C/(k-g)}{SSE_C/[n-(k+1)]} \tag{2.15}$$

$(k-g)$ er antallet β-parametere som blir testet. Senket c representerer verdier fra den komplette modellen, og senket R representerer den reduserte modellen. H_0 blir forkastet dersom beregnet verdi av F er større, eller lik, tabellverdien av F. Tabellverdien er $F_{\alpha,k-g,n-(k+1)}$. H_0 forkastes også dersom α er høyere enn p-verdien. Forkastes H_0 er det 95 % sannsynlighet ($\alpha = 0.05$) at minst en av de testede β-parameterne har utgjør en signifikant nyttighet i å predikere y [16].

2.2 Datakilder

2.2.1 Strava

Strava er en treningsapplikasjon som lar brukere registrere og dele fysisk aktivitet ved å logge GNSS-spor. Sporet består av X- og Y-verdier som er lagret med et tidsstempel [18].

2.2.1.1 Strava Metro

Strava Metro er en nettløsning for nedlasting av data fra Strava. Alle GNSS-spor korrigeres og aggregeres til linjesegmenter som representerer statistikk for tilknyttede spor. Tilgang på aktivitetsdata fra Strava Metro leveres kun ut til brukere som blir godkjent gjennom en enkel søknadsprosess. Bakgrunnen for dette er at Strava Metro kun ønsker å levere ut data til organisasjoner som vil jobbe for en positiv utvikling for Strava-brukere og samfunnet forøvrig. Dette er organisasjoner som er aktive i planlegging og vedlikehold av infrastruktur tilknyttet samferdsel, eller som jobber for å forbedre planleggingsprosesser [19].

En søknad ble sendt inn til Strava Metro den 14.01.2022 med begrunnelse i denne masteroppgaven, og dens tilknytning til NMBU. I søknaden må ønsket studieområde (AOI[1]) spesifiseres i form av land og fylke. Det er også mulig å spesifisere kommune, men dette er ikke et krav [19]. Til denne oppgaven ble fylket Oslo i Norge valgt som AOI, definert av fylkesgrensen i OpenStreetMap[2]. I senere versjoner av Strava Metro skal det bli mulig å definere egne grenser for studieområdet, men for gjeldende versjon er dette verktøyet ikke tilgjengelig [22].

Tilgang til nettjenesten Metroview fra Strava Metro ble tildelt 19. januar 2022 på bakgrunn av innsendt søknad. Metroview består av tre verktøy for henholdsvis analyse, visualisering og nedlasting av posisjonsdata aggregert fra Strava-brukere:

Dashboard Viser data tilgjengelig fra valgt studieområde presentert gjennom ulike statistiske diagramtyper. Data om antall turer og brukere samt aldersfordeling kan analyseres innad og i mellom ulike år.

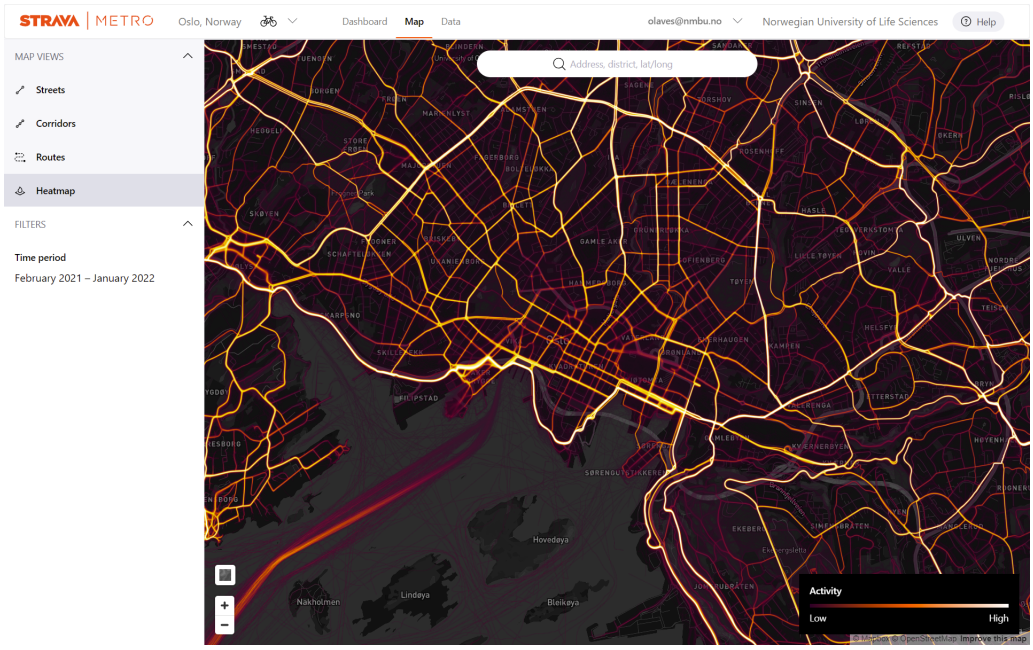
Map Karttjeneste med mulighet for å visualisere og laste ned posisjonsdata. Tilgjengelige kartlag er *Streets*, *Cooridors*, *Routes* og *Heatmap* som viser henholdsvis data på gatenivå, populære traseer, egendefinerte ruter og koropletkart (varmekart). Andre verktøy er endring av bakgrunnskart og filtrering av data basert på tidsperiode og/eller aktivitetsformål[3]. Figur 2.2 viser et skjermbilde av den aktuelle karttjenesten.

Data Side for nedlasting av data bestilt fra *Map*. Informasjon som filnavn, beskrivelse av innholdet i datasettet og filstørrelse er tilgjengelig her.

[1] Area-Of-Interest [20]

[2] OpenStreetMap (OSM) er en åpen internasjonal kartløsning som baserer store deler av sin datainnsamlingen på frivillige bidrag i form av brukerdelte posisjonsdata [21].

[3] Pendling eller fritid.



Figur 2.2: Kartutsnitt av koropletkart (type varmekart) med data fra februar 2021 til januar 2022, skjermbilde av Oslo Sentrum, i Metroview [23].

Fra kartet i Metroview er det mulig å laste ned data innenfor hele studieområdet, eller fra et selvdefinert område. Når et område er valgt kan tidsintervallet avgrenses. Dette intervallet strekker seg fra et år, som det mest generelle, til en avgrensning på år, måned, dato og time. Det er også mulig å velge kun hele dager eller måneder. Etter at et område og tidsrom er spesifisert blir en forespørsel om dataeksportering opprettet. Dataene blir deretter levert som en komprimert mappe (ZIP-fil) bestående av en Shapefil og en CSV-fil. Shapefilen inneholder linjer som er basert på samtlige veier og stier som finnes i databasen til OpenStreetMap innenfor det valgte område. En linje er definert mellom to punkt som enten deler endepunkt med andre linjer (vei-/stikryss), eller er frittstående (blindvei). Hver linje har en ID som kobler den til tilhørende aktivitetsdata i CSV-filen [24]. Et utvalg av data som er lagret i CSV-filen er beskrevet i tabell 2.1.

Sett bort ifra ID-kode, aktivitetstype og datatidspunkt så består CSV-filen av totalt 26 kolonner med data. I tillegg til totalverdiene i tabell 2.1 finnes data om alder- og kjønnsfordelinger, samt turens formål, i begge retninger for samtlige linjesegmenter[1]

[1]Linjesegmenter, også kalt linjestykker, er en del av en rett linje, men som er begrenset mellom to endepunkter [25]. I vårt tilfelle er ikke nødvendigvis linjen rett. Linjen kan for eksempel være en vei som er delt opp i flere linjesegmenter med endepunkter i veikryss, veidelinger eller veiens ende.

Tabell 2.1: Beskrivelse av utvalgte kolonnekoder i CSV-fil fra Metroview [26].

Kolonne	Beskrivelse
edge_uid	Kode for kobling til linjesegment i shapefil.
activity_type	Beskriver aktivitetstypen[1].
hour/date/month/year	Definerer tidsperioden for datafangst i aktuelt linjesegment. Hver tidsenhet representert i egen kolonne.
forward_trip_count	Antall passeringer som er registrert i retning **fra start- til endepunkt** i linjesegmentet.
reverse_trip_count	Antall passeringer som er registrert i retning **fra ende- til startpunkt** i linjesegmentet.
forward_people_count	Antall unike brukere som er registrert i retning **fra start- til endepunkt** i linjesegmentet.
reverse_people_count	Antall unike brukere som er registrert i retning **fra ende- til startpunkt** i linjesegmentet.
	...
forward_average_speed	Gjennomsnittlig fart oppgitt i meter per sekund beregnet fra passeringer som er registrert i retning fra start- til endepunkt i linjesegmentet.
reverse_average_speed	Gjennomsnittlig fart oppgitt i meter per sekund beregnet fra passeringer som er registrert i retning fra ende- til startpunkt i linjesegmentet.

2.2.2 Sykkeltellere

Sykkeltellere består av ett eller flere sykkelregistreringspunkt hvor retning og fart på syklister som passerer blir detektert. Sykkelregistreringspunkt inngår som en del av en trafikkregistreringsstasjon som også kan være tilknyttet ett eller flere registreringspunkt for motorkjøretøy. En slik stasjon kan typisk bestå av en eller flere fysiske sensorer som registrerer passeringer, og et skap som inneholder en datalogger og en ruter. Tilknyttet stasjonen er det ofte oppført en skjerm som normalt viser antall passeringer den aktuelle dagen, totalt antall passeringer inneværende år i tillegg til gjeldende dato og tidspunkt. I dataloggeren lagres registrerte passeringer som deretter deles via ruteren til et trafikkdatasystem i sanntid. På den måten kan systemet raskt kontrollere innkommende data i henhold til fastsatte kvalitetskrav, samt aggregere dataene før publisering [27].

Et eksempel på dataflyten i et trafikkregistreringssytem, her representert av Statens vegvesen sin løsning, er vist i figur 2.3. Figuren viser et system satt opp for registrering av motorkjøretøy, men prinsippet for dataoverføringen er det samme for sykkelregistreringspunkt [27].

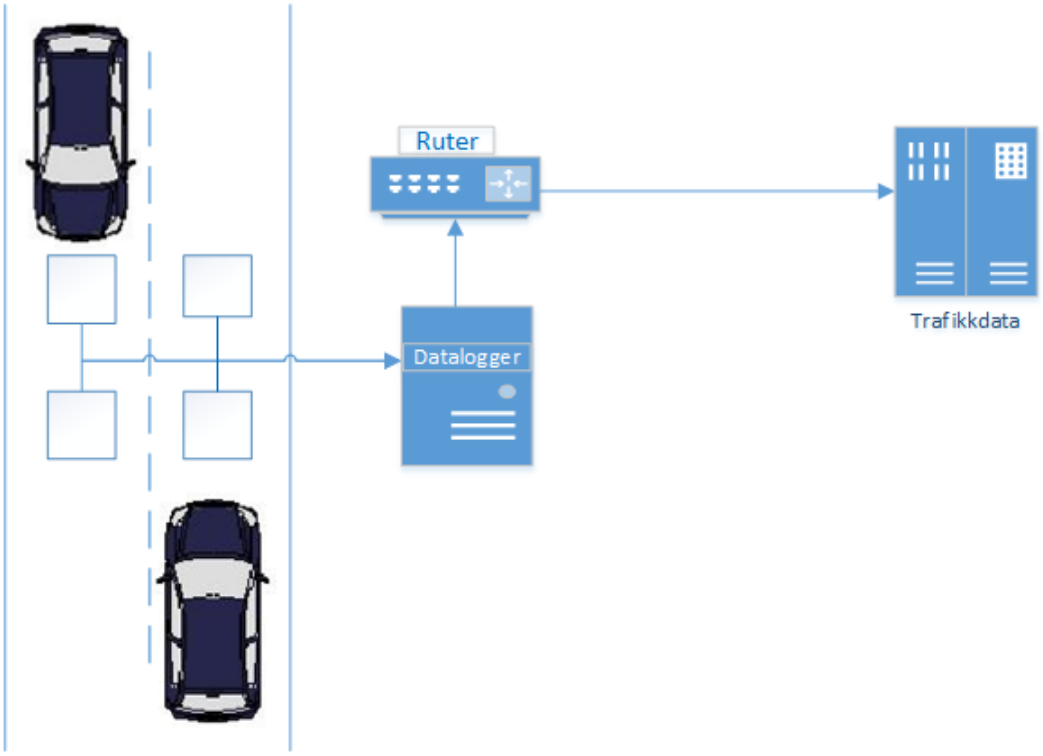
[1]Sykling eller løping/gå/tur

I Norge har det vært praktisert bruk av flere ulike typer sensorer for maskinell sykkelregistrering. Typiske sensorer som er, eller har vært i bruk er:

- Induktive sløyfer
- Piezoelektriske kabler
- Luftfylte gummislanger
- Radarsensor
- Videoanalysering av bilder

Av disse er det i dag induktive sløyfer og piezoelektriske kabler som benyttes ved nye installasjoner. I en SINTEF rapport fra 2009, «*Utprøving av utstyr for å registrere sykkeltrafikk*», ble det konkludert med at induktive sløyfer og en gummislangeløsning ga best resultat for registrering av sykkeltrafikk uten motoriserte kjøretøy i samme kjørebane. I studiet ble det testet åtte ulike oppsett ved bruk av fire forskjellige sensorer. I tillegg til de to nevnte sensortypene, ble også infrarød sensor og radarsensor testet. Blant sykkelregistreringspunkt som er etablert og operative i Oslo kommune per nå, så er den domminerende sensortypen nesten utelukkende induktive sløyfer.

Statens vegvesen har fastslått at alle sykkelregistreringspunkt de etablerer skal registrere trafikken kontinuerlig. Dette nivået, såkalt registreringsnivå 1, representerer den hyppigste registreringen et registreringspunkt kan ha. I andre enden av skalaen, på registreringsnivå 4, gjøres det kun trafikkregistrering i forbindelse med spesialundersøkelser som forekommer ved ujevne mellomrom. Bakgrunnen for valg av registreringsnivå kom som et resultat av et ønske fra 2003 om at det skulle etableres minst 25 registreringspunkt for sykkeltrafikk totalt i en samling på 13 fylker med kontinuerlige målinger. På sikt var målet å utarbeide en landsomfattende sykkelindeks basert på innsamlet data fra de ulike sykkeltellerne. Sykkelindeksen skulle vise utviklingen i sykkeltrafikken og på den måten fungere som et hjelpemiddel i beslutningsprosesser knyttet til utbygging av nye sykkeltraseer [29].



Figur 2.3: Illustrasjon hentet fra *Statens vegvesen - Om trafikkdata* [27]. Fra et kjøretøy er registrert av sensoren, til aggregert data er tilgjengelig i trafikkdataportalen til Statens vegvesen tar det omlag to til tre timer.

2.2.2.1 Induktive sløyfer

I *Veileder i trafikkdata* [29] er induktive sløyfer definert slik:

"Induktive sløyfer er elektriske ledninger som legges i nedfreste spor i vegbanen på en slik måte at de danner en spole. Ledningene påføres en vekselspenning, og når et kjøretøy passerer sløyfene vil metallet i kjøretøyet bryte det magnetfeltet som er dannet over sløyfene. På denne måten blir det enkelte kjøretøy registrert."

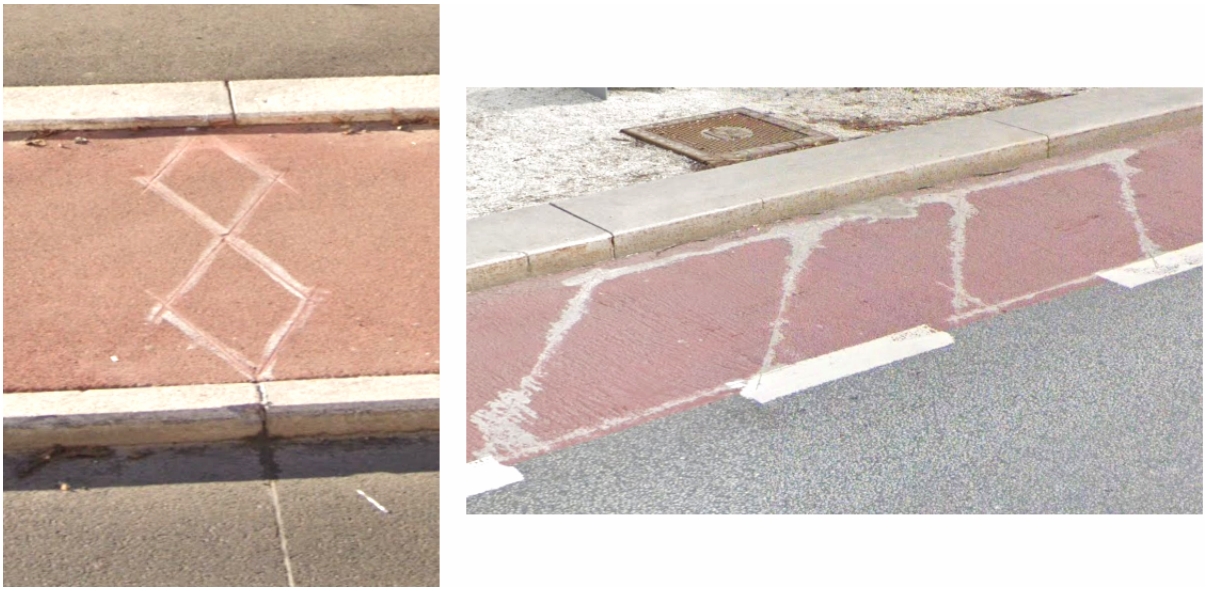
Figur 2.3 viser bruk av fire induktive sløyfer som sensor for registrering av passerende kjøretøy. Med et oppsett der fire sløyfer er benyttet er muligheten for å logge lengden og farten på kjøretøy i begge kjøreretninger til stede. For sykkelregistrering er det flere tilsvarende løsninger der sløyfene, i likhet med registrering av motorkjøretøy, også legges i asfalten. Det som derimot er ulikt er mønsteret som tegnesnår sløyfene skal legges. Blant utstyret som anvendes i sykkelregistreringspunkt i Oslo legges sløyfene etter et mønster som tar utgangspunkt i et parallellogram. Valg av sidelengder på parallellogrammene varierer mellom de ulike løsningene leverandørene av sykkeltellere leverer. De to vanligste mønsterløsningene i Oslo er vist i figur 2.4.

I retningsbestemte sykkelfelt, der trafikken er lagt opp til å gå i en kjøreretning, er det vanlig å installere kun én sløyfe for å registrere passeringen, og eventuelt farten på passerende syklister. Er det i tillegg ønskelig å fastslå retningen på syklisten over registreringspunktet, kreves det installasjon av minimum to sløyfer. Dette kan være aktuelt på gang- og sykkelveier der syklister i begge kjøreretninger deler samme trase [28] [29].

Eksemplene i figur 2.4 viser sykkelfelt med påbudt kjøreretning. I Dronning Eufemias gate er det til tross for retningsbestemmelsen, installert to sløyfer. Dette gjør det mulig å registrere syklister som sykler mot påbudt kjøreretning. Måles det høye verdier her, kan det være nødvendig å vurdere tiltak for å begrense mengden syklister mot påbudt kjøreretning. Typiske tiltak kan være forbedret skilting eller fysiske hindringer. En reduksjon av syklister mot kjøreretningen vil bidra til å minske sjansen for ulykker på strekningen.

I Oslo er det to ulike leverandører av registreringsapparat for syklister. På oppdrag fra Bymiljøetaten i Oslo kommune utplasserte det kanadiske selskapet ECO-Counter 43 tellere per september 2017 [30]. For registreringspunktene Statens vegvesen har ansvaret for er ulike versjoner av Datarec fra Aanderaa Data Instruments benyttet. Valg av versjonstype avhenger i mange tilfeller av om registreringspunktet skal installeres på en adskilt gang- og sykkelvei eller i blandet trafikk med motorkjøretøy. De fleste sykkelregistreringspunkt befinner seg på gang- og sykkelveier. Punkt som er plassert i blandet trafikk skal like fult være et selstendig registreringspunkt som ikke inngår i registreringspunktet for motorkjøretøy. Bakgrunnen for dette er blant annet at sykkelveier, uavhengig av tilknytning til andre veityper, inngår i NVDB[1] som selvstendig vei [29].

[1]Nasjonal vegdatabank



Figur 2.4: Eksempler på induktive sløyfer. Fra venstre: ECO-Counter installasjon på østgående sykkelfelt i Åkebergveien, og Datarec installasjon i vestgående sykkelfelt i Dronning Eufemias gate. Begge gatene ligger i Oslo kommune. Begge skjermbilder er hentet fra *Google Maps* [31].

ECO-Counter er leverandør av sykkeltellere til Oslo kommune. Sløyfene legges ofte slik som i figur 2.4 med utforming som to tilstøtende parallellogram dersom sykkeltraseen er bred nok. Er den smalere er det ofte brukt kun én sløyfe med lengre sidelengder. I SINTEF's studie [28] ble ECO-Counters løsning for induktive sløyfer testet i sykkelfelt ment kun for syklister og i veibanen der både gående, syklister og motorkjøretøy ferdes. Nøyaktigheten[1] ble målt til å være 97,5 % i sykkelfeltet og 83,5 % i gaten med ulike trafikktyper. Virkningsgraden[2] ble målt til henholdsvis 97,8 og 96,3 prosent. På bakgrunn av rapportens målinger for nøyaktighetsmål og virkningsgrad, er sykkeltellerne fra ECO-Counter å regne som troverdie datakilder.

Datarec 7 og Datarec 410 er begge løsninger som er tatt i bruk av Statens vegvesen i deres sykkeltellere. Fellesnevneren mellom disse to systemene er at de begge benytter induktive slynger med samme type mønster. Figur 2.5 illustrer hvordan slyngene legges i forhold til hverandre og veidekke. For å sikre at slyngen skal registrere alle forbipasserende syklister er det viktig at kortsidene (a-d og b-c) strekkes helt ut mot kanten av banen og at disse har en lengde på 1,2 meter som tilsvarer den normale akselavstanden på en sykkel. Bredden ("Width" i figuren) på slyngene bør ligge på mellom 1,5 og 3,3 meter. →

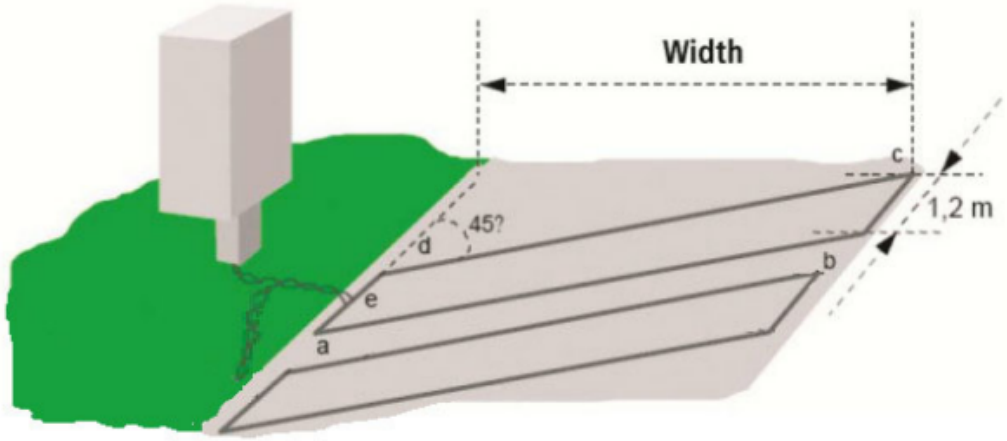
[1]Forholdet mellom *Riktig registrering av utstyr* og *Antall passeringer* i prosent. Nøyaktigheten er et mål på hvor godt utstyret unngår å registrere objekter som ikke er sykler.

[2]Forholdet mellom *Totalt registrert* og *Antall passeringer* i prosent. Virkningsgraden er et mål på hvor godt utstyret er til å fange opp objekter som passerer. Objektene som blir talt blir registrert som sykler.

Er sykkelfeltet mye bredere enn denne maksverdien bør det vurderes å heller legge to smalere slynger parvis i bredden. De to spisse vinklene i punkt a og c skal legges på 45 grader og avstanden mellom to slyngepar skal være 1,0 meter. Slyngene legges maksimalt 2 cm under overdekket [32]. I Dronning Eufemias gate (figur 2.4) har Statens vegvesen installert slynger etter nevnte kriterier.

I studien til SINTEF[28] ble både Datarev 7 og 410 testet på flere ulike gang- og sykkelveier der både gående og syklister ferdes. Resultatet fra manuelle kontroller viser at gjennomsnittet av nøyaktigheten fra tre ulike studieområder var på 94,1 % for Datarec 7 og 97,4 % for Datarec 410. Tilhørende gjennomsnitt av virkningsgraden fra samme tilfeller var henholdsvis 95,9 % og 99,2 %. Testene viser at begge versjonene av Datarec gir tilfredstillende nøyaktighet, men at Datarec 410 er 3,3 prosent mer nøyaktig. I studien konkluderes det også med at de løsningene som gir høyest nøyaktighet i trafikk uten motoriserte kjøretøy er ECO-Counter og Datarec 410. Virkningsgraden til Datarec 410 er 3,3 prosent bedre enn med Datarec 7 og førstnevnte kan derfor sies å være mer følsom for passeringer.

Fordelen med Datarec 7 er at den kan brukes i kombinasjon med sensorer for registrering av motorkjøretøy. Figur 2.6 viser fylkesvei 160 i Bærum kommune hvor dette er gjort. Ved å kombinere to trafikksensorer så kan kostnader reduseres ved at utstyret for lagring og overføring av data halveres sammenlignet med å installere to stasjoner separat.



Figur 2.5: Krav til dimensjoner på oppsett av to induktive slynger for Datarec 7/410. Med to slynger er det mulig å fastslå retningen på passerende syklister. Illustrasjon hentet fra *Inductive Loops - Technical note*[32].

Figur 2.6: Trafikkregistreringspunkt for både motorkjøretøy og sykkeltrafikk. Fire induktive sløyfer er installert i veibanen (rød sirkel) på samme måte som i figur 2.3. På den separate gang- og sykkelstien er sløyfene lagt som i figur 2.5 for å registrere passeringer og fastslå retningen (grønn sirkel). Skjermbilde fra *Google Maps* [31].

2.2.2.2 Innhenting av data

Statens vegvesen har en webløsning kalt *Trafikkdata*[33] som inneholder en landsdekkende karttjeneste med alle deres trafikkregistreringspunkt. Både registreringspunkt for motorkjøretøy og sykler er samlet i denne karttjenesten. Her er det mulig å filtrere punkter basert på *Vegkategori*, *Fylker*, *Registreringshyppighet* og *Trafikanttype*. Tilknyttet siden finnes informasjon om datainnsamlingsprosessen, API[1] og side for eksportering av data. Ved eksportering av data markeres registreringspunkt i kartet, tidsintervallet defineres og ønsket aggregeringsnivå velges. På bakgrunn av valgenen som er gjort generere tjenesten en CSV-fil med data klar for nedlasting. Kolonner og en kort beskrivelse av disse er vist i tabell 2.2.

Oslo kommune sin webløsning for deres trafikkregistreringspunkt er levert av ECO-Counter [35]. I likhet med trafikkdata fra Statens vegvesen består siden av en karttjeneste som viser tellepunkt for gående og syklende. Her er det mulig å definere tidsperiode og vise statistikk over antall syklister i ulike retninger i et tellepunkt. Til mange av tellepunktene er det også lagt ved ett eller flere bilder med beskrivelse av hvor sensorene, som i de fleste tilfeller er induktive sløyfer, er frest ned i veien. Disse bildene er nyttige i prosessen med å matche tellepunktene med korrekt linjesegment fra Strava.

ECO-Counter har ingen tilsvarende åpen funksjon for direkte nedlasting av data fra sin nettjeneste slik som Statens vegvesen. Bymiljøetatens avdeling for *Trafikkstyring og analyse* har derimot tilgang på en egen nedlastningsportal for data. Dette gjør det mulig å få tilsendt data fra avdelingen på forespørsel. Struktur på mottatt data er vist i tabell 2.3. Data blir levert på XLSX-format.

[1]Forkortelse for "Application Programming Interface". API er et hjelpeverktøy for programmerere utenfor utviklergruppen av koden til en tjeneste. API beskriver hvordan førstnevnte gruppe kan anvende funksjonene i tjenesten til bruk i andre applikasjoner uten å måtte sette seg inn i tjenestens kildekode [34].

Tabell 2.2: Relevante kolonner i CSV-filer fra dataeksporteringssiden til Statens vegvesen. Beskrivelsene er sitert fra *Statens vegvesen - Om trafikkdata* [27].

Kolonne	Beskrivelse
Trafikkregistreringspunkt	Punktets unike ID.
Navn	Punktets navn.
Vegreferanse	Punktets plassering på vegnettet.
Fra	Starttidspunktet for periodens aggregater.
Til	Sluttidspunktet for periodens aggregater.
Dato (kun for time- og døgntrafikk)	Dato for periodens aggregater.
Felt	Definerer en gitt retning i registreringspunktet eller markert **Totalt** der alle retninger er inkludert. *(Omformulert)*
Volum	Trafikkmengden for den aktuelle perioden og retningen (eller totalt).
Dekningsgrad (%)	Andel tid det er gode data for i perioden.
Antall timer eller døgn total	Antall timer eller døgn i perioden.
Antall timer inkludert	Antall timer som har data i perioden.
Antall timer ugyldig	Antall timer i perioden med ukjent kvalitet.

Tabell 2.3: Kolonner i XLSX-fil fra dataeksportportalen til Bymiljøetaten (Oslo kommune) av sykkeltellerdata fra ECO-Counter.

Kolonne	Beskrivelse
Periode	Tidsintervall for eksportert data.
Time	Data og tidspunkt for telledata.
Første retning	Beskrivelse av en av retningene.
Andre retning	Beskrivelse av den motsatte retningen til første retning.

2.2.3 Begrensninger

Det er enkelte begrensninger med dataene fra Strava som er verdt å nevne:

- Det kan forekomme demografiske skjevheter i dataene fra Strava sammenlignet med resterende populasjon. Mange studier har påpekt at det er en overrepresentasjon av menn i datasettene fra Strava [36].

- Det er flere aktive syklister som benytter Strava enn det er "vanlige" syklister. Aktive syklister prioriterer ikke nødvendigvis å sykle de samme stedene som "vanlige syklister". Syklister med Strava har vist seg å oppsøke gater som legger til rette for høy fart, i motsetning til "vanligesyklister som prioriterer de korteste og tryggeste veiene.

- Ettersom posisjonsdataene blir registrert med en treningsklokke eller mobil så er ikke nøyaktigheten alltid like god. Når Strava Metro benytter sine algoritmer for å matche aktiviteter i Strava med linjesegmenter kan det forekomme koblinger til feil linjesegment. Dersom linjesegmentene ligger tett, eksempelvis for en vei med fortau, kan unøyaktige data føre at data blir tildelt linjesegmentet som representerer fortauet til tross for at syklisten syklet i veien [18].

KAPITTEL

3

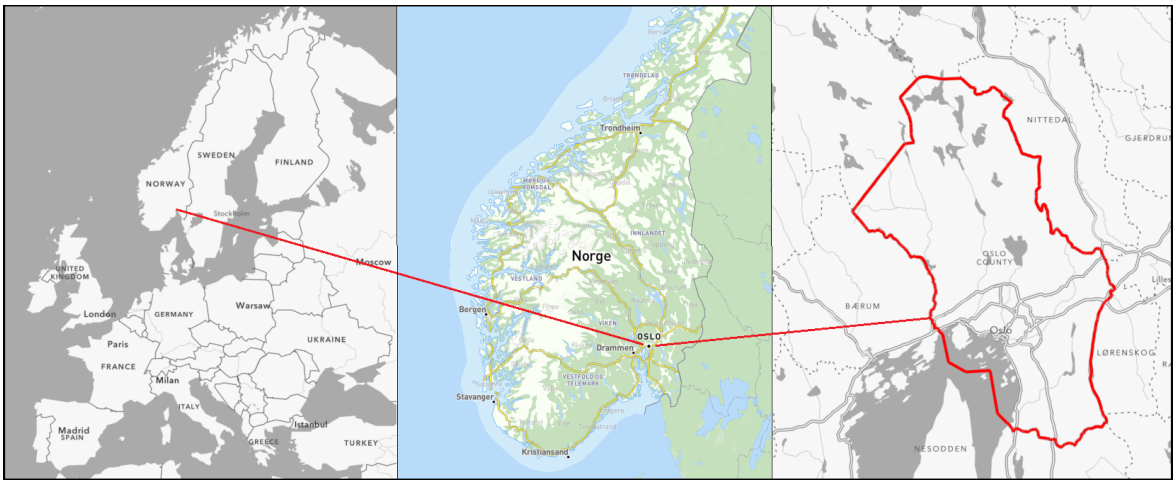
METODE

I kapittelet *Metode* presenteres studiens fremgangsmåte kronologisk. Første del av kapittelet inneholder beskrivelse av datasett og studieområder. Så følger prosessen som er benyttet i forbindelse med utvalg, behandling og undersøkelse av data. Videre foreligger metoden i regresjonsanalysen for estimering og validering av modeller. Avslutningsvis benyttes modellene til å beregne antallet syklister i en gate med nye sykkelfelt.

3.1 Studieområder

Det overordnede studieområde i denne oppgaven var gitt gjennom kommunegrensen til Oslo (figur 3.1). Oslo er Norges hovedstad og består av et areal på 454 kvadratkilometer, hvorav bydelene tilsammen utgjør 137 kvadratkilometer [37]. Tall fra SSB[1] [38] viser at befolkningen i Oslo i 2021 var på totalt 697 010, hvorav den demografiske kjønnsfordelingen var tilnærmet 50/50. Samme år målte Oslos værstasjon, Meteorologisk institutt på Blindern, temperaturer fra -14,5 °C til 30,2 °C, med en gjennomsnittstemperatur på 7,2 °C. Samme stasjon målte 94 nedbørsdøgn og 93 døgn med snø på bakken totalt samme år [39].

I en holdningsundersøkelse fra 2020 om sykling i Oslo utført av Opinion AS for Bymiljøetaten i Oslo kommune, oppga 400 av de totalt 800 spurte Oslo-beboerne at de syklet ofte eller alltid/nesten alltid i Oslo gjennom sommerhalvåret. Kun 13 % av den samme gruppen ga tilsvarende svar på spørsmål om sykling i vinterhalvåret [40]. Statistikk fra *Map* i Metroview viser at det i 2021 var registrert 774 110 unike sykkelaktiviteter i Oslo (innenfor AOI) dette året. Bak dette tallet var det 38 270 unike brukere av Strava som bidro til å registrere en eller flere av disse sykkelaktivitetene. Dette tilsvarer rett i overkant av 20 aktiviteter per bruker.



Figur 3.1: Oslo kommune. Kartutsnitt hentet fra Esri mfl. [41] og Norkart [42].

I alle bydeler i Oslo, utenom Bydel Stovner, er det utplassert en eller flere sykkeltellere som enten er driftet av Statens vegvesen eller Oslo kommune. Dette danner et godt grunnlag for god geografisk spredning i datasettet. Som tidligere nevnt er fylkesgrensen til Oslo valgt som AOI i Strava Metro. Fra Metroview kan derfor alle tilgjengelige data tilknyttet linjesegmentene definert i OSM tilbake til 2017 lastes ned på ulike aggregeringsnivåer for hele Oslo. →

[1]Statistisk sentralbyrå

Alle sykkelveier i Oslo som er av interesse i forbindelse med denne studien er kartlagt i OSM.

En av de viktigste grunnene til at Oslo er valgt som studieområde, er tilgangen på relevante data. Omfanget og størrelsen på dataene er et resultat av følgende:

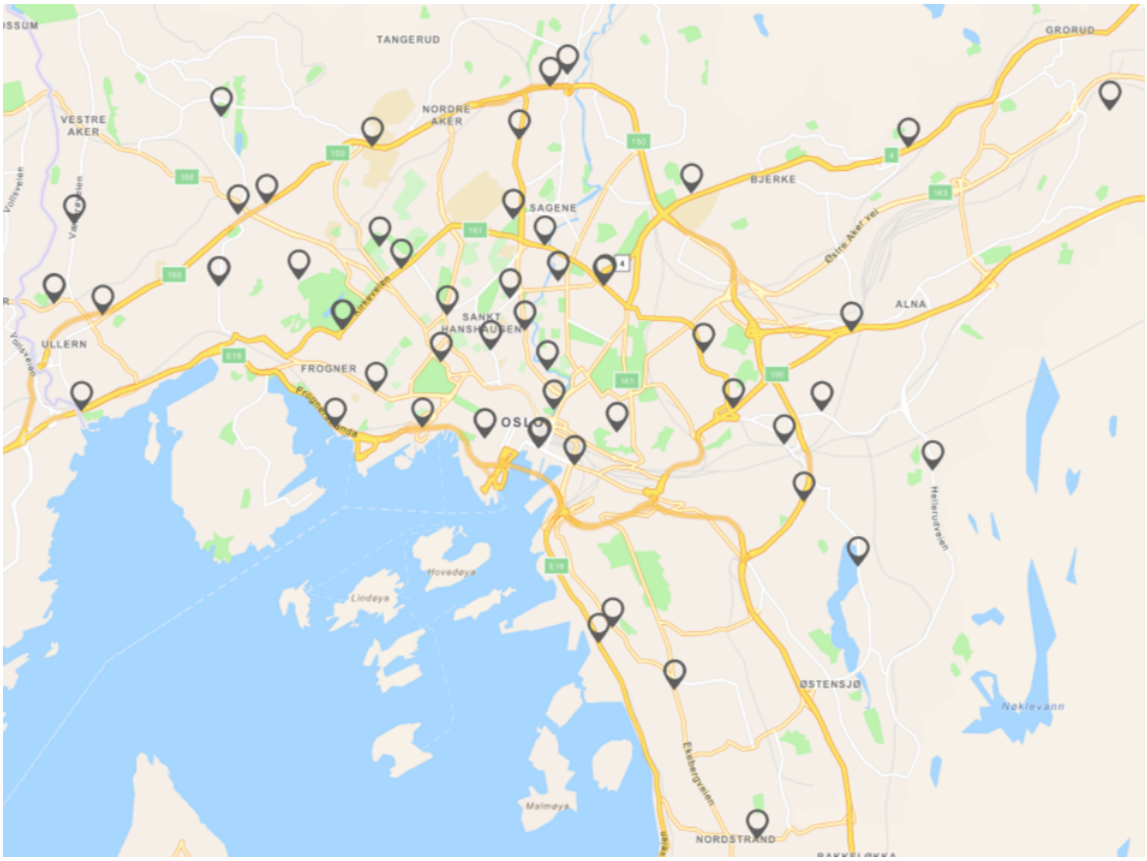
- Flertallet av Oslos innbyggere oppgir at de benytter sykkel som fremkomstmiddel.
- Brukere av Strava tilsvarende 18.2 % av alle innbyggerne i Oslo har logget en eller flere sykkelaktiviteter i løpet av et år (2021).
- Sykkeltellere installert på ulike steder av byen har registrert syklister daglig over flere år.

På bakgrunn av overnevnte forhold, samt andre faktorer nevnt i dette delkapittelet, vurderes Oslo til å være et godt egnet studieområde sett i lys av studiens formål.

Innenfor studieområdet er det valgt ut to fokusområder som hver for seg har hvert sitt formål i studien. Fokusområde 1 (3.1.1) er valgt ut med den hensikt å gi best mulig grunnlag for å kunne besvare forskningsspørmål 1. Dette området er også benyttet i forbindelse med modellbygging. Fokusområde 2 (3.1.2) er valgt ut for testing og kontroll av modellen. Resultater fra dette området skal bidra til å gi svar på forskningsspørmål 2.

3.1.1 Fokusområde 1

Fokusområde 1 ble valgt med tanke på å generere størst mulig datasett for å kunne studere sammenhengen mellom data fra sykkeltellere og Strava. Med bakgrunn i dette, og med mål om god geografisk spredning, ble alle sykkeltellere i hele Oslo kommune vurdert. Figur 3.2 viser den geografiske avgrensningen til fokusområde 1, samt plasseringen til alle sykkeltellerne som er oppført i webtjenestene til Oslo kommune (*Sykkeltellere i Oslo kommune* [35]) og Staten vegvesen (*Trafikkdata* [33]).

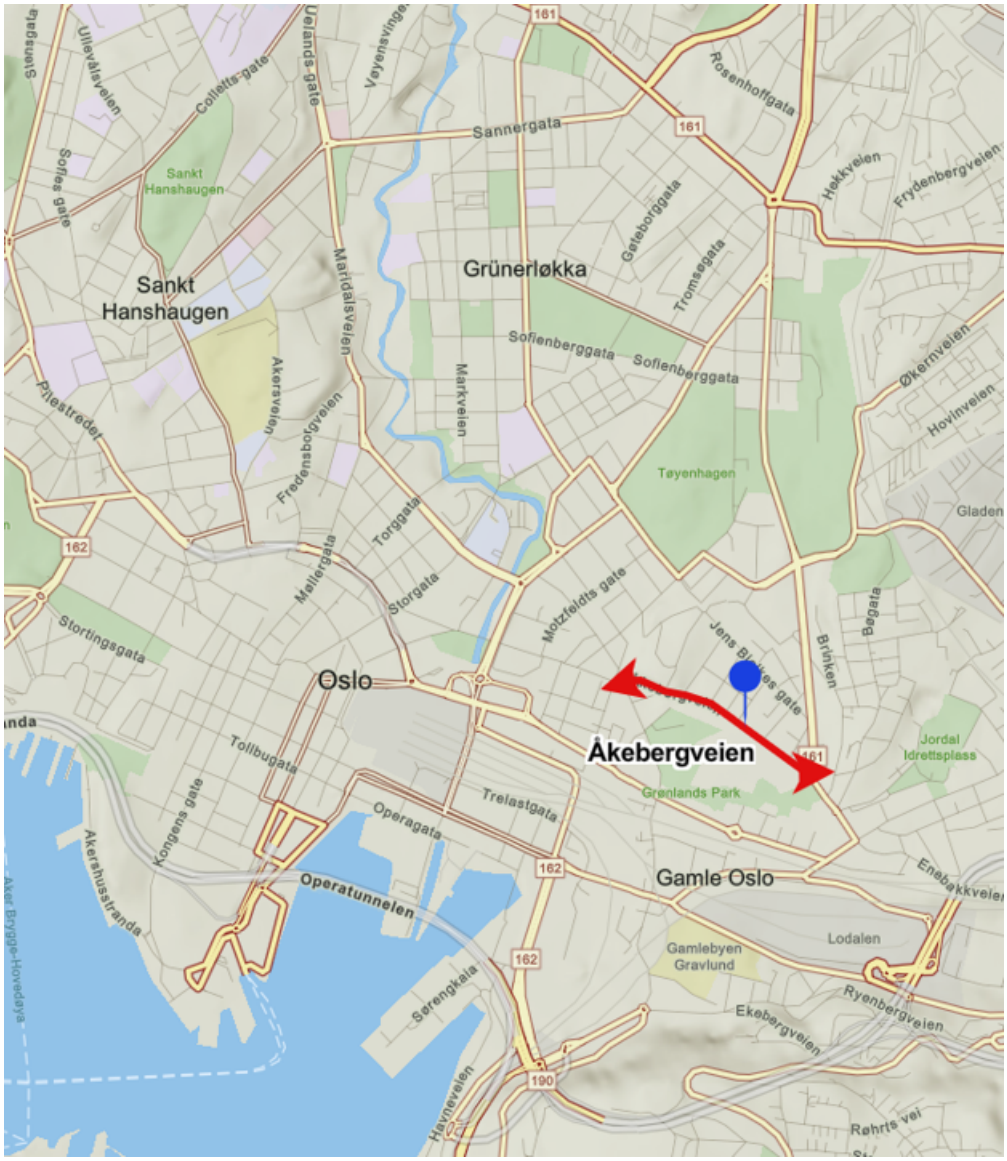


Figur 3.2: Kart over aktuelle sykkeltellere i Oslo. Konstruert i ArcGIS Mapviewer.

3.1.2 Fokusområde 2

I fokusområde 2 inngår deler av Åkebergveien i Oslo. Åkebergveien, som er markert i figur 3.3, tilhører bydelen Gamle Oslo og ligger nord for Grønlandsparken. I Åkebergveien ble det mellom august 2018 og august 2020 anlagt envegsregulert sykkelvei, videre omtalt som opphøyde sykkelfelt, mellom Grønlandsleiret og Kjølberggata. Strekningen med sykkelfelt er på omlag 700 meter og inngikk i Statens vegvesens prosjekt "Sykkelpilot" [43] som en av de første sykkelveiene med denne type utforming i Oslo. Før prosjektet hadde gaten ingen sykkelfelt, men bilparkering i vestgående veibane. Mellom sommeren 2017 og fram til august 2018, ble en midlertidig løsning innført der parkeringsplassene ble fjernet og et ikke-opphøyd sykkelfelt ble merket opp i østgående veibane. I dag har sykkelfeltene i Åkebergveien en bredde på 2,2 meter og ligger på et eget nivå i forhold til både veibanen og fortauet. De to trafikale løsningene i Åkebergveien, før og etter prosjektet, kan studeres i figur 3.4. I følge "Oslostandarden for sykkeltilrettelegging" [44] er opphøyde sykkelfelt egnet for sykling på flere ulike hastigheter og har stor kapasitet i forhold til antall brukere. Denne løsningen er derfor ment å dekke behovet hos ulike typer syklister. Sykkelfeltene i Åkebergveien representerer den østlige enden av Byrute 2 mellom Galgeberg og Smestad, og er derfra tilknyttet sykkelfeltene vestover i Grønlandsleiret [45].

Rett vest for Åkebergveien 28 (59°54'38.3"N 10°46'24.3"E) er det oppført en sykkelteller som registrerer syklister i begge retninger. Sykkelteller, tilhørende Oslo kommune, har data fra før og etter utbyggingen av de opphøyde sykkelfeltene. I utbyggingsfasen var sykkelteller midlertidig demontert. Data fra perioden før og etter demonteringen er fortsatt tilgjengelig på *Sykkeltellere i Oslo* [35]. Strava Metroview har tilgjengelig data for aktiviteter tilhørende linjesegmentene i Åkebergveien tilbake til 2018. Sykkelfeltløsningen som er anlagt i Åkebergveien gjør gaten interessant for å teste ut modellen. Sykkelteller i gaten gjør det også mulig å sammenligne det reelle antallet syklister med estimerte verdier fra modellen. I fokusområde 2 inngår også enkelte sidegater som er tilknyttet Åkebergveien.



Figur 3.3: Plassering og utstrekning for sykkelfeltene i Åkebergveien. Sykkeltellerens plassering er markert med blå markør.

Figur 3.4: Over: Åkebergveien før oppgradering, mai 2017. Under: Åkebergveien med opphøyde sykkelfelt, november 2020. Skjermbilde fra *Google Maps* [31].

3.2 Datasett

3.2.1 Aggregering i tidligere studier

Valg av tidsperiode for datasettet var noe av det første som ble vurdert. I tidligere studier som har brukt data fra Strava for sammenligning med tellerdata har det vært store variasjoner knyttet til gruppering med hensyn på tid. I Livingston mfl. [36] ble data aggregert fra timer og opp til to-dagers perioder. Aggregering på timer ga lavest korrelasjon, mens to-dagers aggregering resulterte i høyest verdi. Korrelasjonskoeffisienten var på henholdsvis 0,781 og 0,887. Roy mfl. [8] sammenlignet data fra Strava med telledata gjennom aggregering på daglig, månedlig og årlig nivå. Høyeste R^2-verdi ble oppnådd ved en modell som benyttet årlig aggregert data.

Hong, McArthur og Livingston [9] studerte den lineære relasjonen mellom data fra sykkeltelling og Strava. Det ble aggregert på flere nivåer, fra time-for-time opp til to og to dager. Også her resulterte høyere grad av aggregering til høyere r- og R^2-verdier. For grupperinger på timer ble korrelasjonskoeffisienten beregnet til 0,816, og for dag-for-dag nivå var verdien 0,908. For å bekrefte den sterke relasjonen ble det satt opp lineære regresjonsmodeller. Modellene fra data gruppert på timer og dager hadde R^2-verdi på henholdsvis 0,67 og 0,82. Studien vurderte aggregering på månedlig nivå til å gi en ytterligere økning i korrelasjonen og at studiens modell dermed ville kunne gi resultater som var representative for hele populasjonen i studieområdet. I studien ble det brukt en Poisson modell med månedlige aggregerte data som input og underbygget dette med det faktum at modellen kun gjør sammenligninger ved bruk av én enhet, som er antallet syklister som benytter Strava.

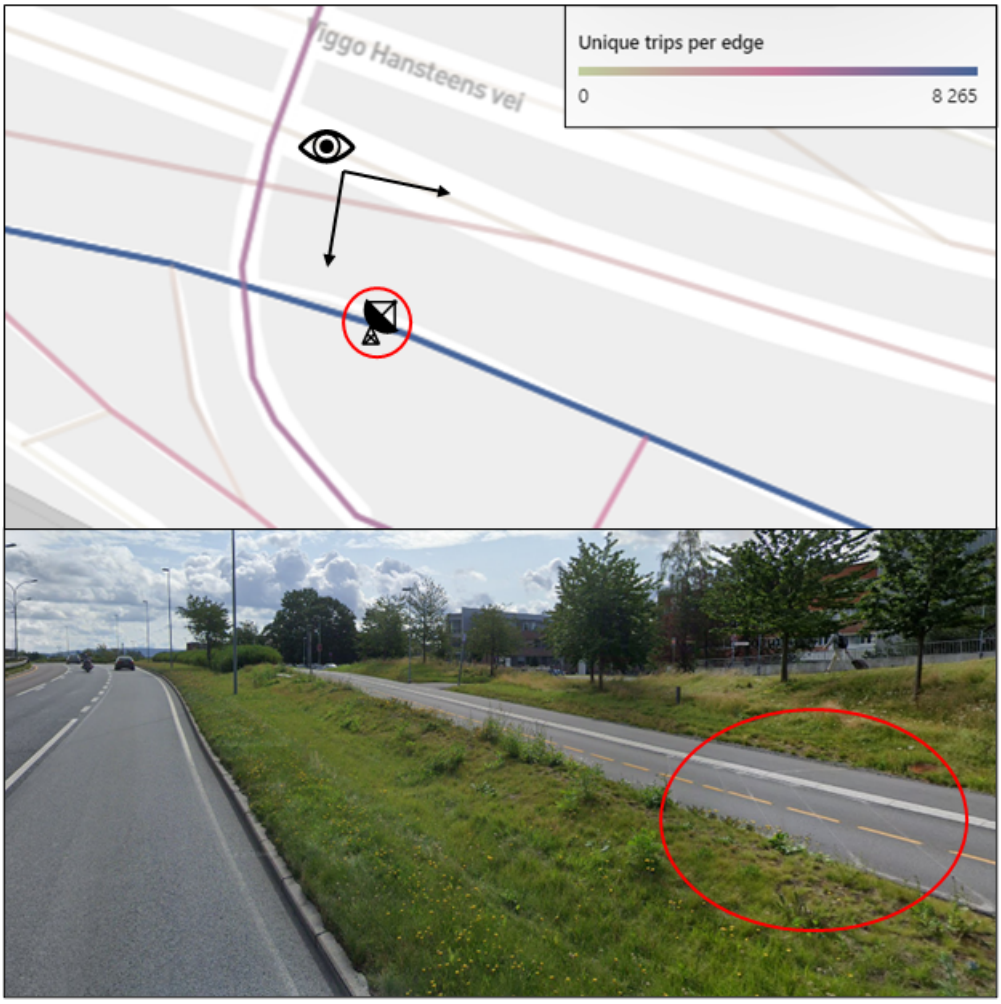
3.2.2 Utvalg

Valg av aggregeringsnivå for datasettet ble tatt på bakgrunn av flere vurderinger. Tidligere studier har vist at datasett bestående av målinger gjort over en lengre tidsperiode har høyere korrelasjon enn kortere perioder. "Trafikkdata" fra Statens vegvesen leverer data fra sykkeltellere helt ned på timesnivå. Tellerdata fra Oslo kommune er derimot kun tilgjengelig på et dag-for-dag nivå. Et datasett med lavere aggregering enn på dette nivået ble derfor utelukket ettersom Oslo kommunes tellere representerer nærmere 80 % av alle sykkeltellerne i Oslo. Statistikk fra tellere på Trafikkdata viste at månedene i sommerhalvåret, fra og med april, til og med september skilte seg ut fra resterende måneder med et markant høyere antall registrerte syklister. Dette bekreftes også av Opinions spørreundersøkelse [40]. I Livingston mfl. [36] kom det frem at data fra områder med få syklister (under 50 registrert), som ble aggregert på et lavt nivå, resulterte i svak korrelasjon mellom data fra Strava og sykkeltellerne. Utvalg til datasett i denne studien falt derfor på månedlig aggregering av data, fra mai måned.

Bakgrunnen for dette valget er at mai er den første måneden i året hvor sykkeltrafikken tar seg kraftig opp, og historisk sett gir et stabilt høyt antall snittpasseringer ved sykkeltellere i Oslo. Måneden inneholder også flere offentlige fridager som Arbeidernes internasjonale dag (1. mai), Norges grunnlovsdag (17. mai) og Kristi himmelfartsdag (dato varierer). Nevnte dager har vist seg å bidra til en variasjon i datasettene i form av en markant økning, eller reduksjon i antallet syklister. Dette er variasjoner det er interessant å studere nærmere med hensyn på om de gjenspeiles i datasettene fra Strava.

Fra Strava Metro var det for valgt AOI mulig å laste ned data som ble registrert tilbake i 2017. Alle tellestasjonene på Trafikkdata [33] og ECO-Counter [35] ble derfor undersøkt for data fra mai 2017 og fram til mai 2021. Stolpediagram over dataene ble visuelt kontrollert etter prinsipper fra Turner og Lasley [46] side 64-65 for detektering av mulige feil. Åpenbare feil i data fra sykkeltellere skyldes ofte defekter på registreringsutstyret. I tilfeller med hull i dataserien skyldes dette som regel vedlikehold eller utbedring av veibanen der sykkeltelleren er plassert.

En utfordring knyttet til utvalget for datasettene fra Strava var å velge det linjesegmentet som representerer syklistene som hadde passert registreringssensoren i Metroview. Noen tellestasjoner består utelukkende av én registreringssensor (sløyfer) installert på en separat gang- og sykkelvei, som tilfellet i figur 3.5. I kartutsnittet øverst i figuren er et tilfelle med lav usikkerhet knyttet til valget om hvilken linje som representerer syklistene som passerer tellepunktet. I dette eksempelet fra Gaustad ligger gang- og sykkelveien med god avstand til nærliggende veier, og dermed andre linjesegmenter. Dette gjør at syklister som logger en aktivitet gjennom Strava, og som passerer tellepunktet (rød sirkel) på Gaustad vil med høy sannsynlighet få tilknyttet denne aktiviteten til dette linjesegmentet (blå linje). Mange aktiviteter på linjen, sett i sammenheng med tilsvarende lave verdier på "konkurrerende" linjer tilfører høy kvalitet til disse dataene fra Strava.



Figur 3.5: Gaustad i Oslo. Øverst: utsnitt fra Metroview av linjesegmenter i området rundt Statens vegvesens sykkelteller "Gaustad sykkel". Fargeverdiene på linjene representerer tettheten av data tilknyttet hver linje fra mai 2021. Nederst: bilde fra Google Maps [31] der rød sirkel viser plasseringen til tellersensoren.

I andre tilfeller er det mer usikkert hvorvidt alle syklister som registrerer turen med Strava, og sykler over et tellepunkt, blir tilknyttet det samme linjesegmentet. Sør-vest for Ankerbrua i Oslo er det installert to tellesensorer i hvert sitt respektive sykkelfelt i veibanen. Utenfor sykkelfeltene ligger det fortau som er definert som egne linjer i OSM og Metroview. I nederste bilde i figur 3.6 illustreres nevnte tilfelle. Alle tre linjene, i utsnittet øverst i samme figur, som krysser broen (fortau, veibane, fortau) er tilknyttet et betydelig antall turer. Ut ifra fargegraderingen som representerer antall aktiviteter, har likevel linjen for veibanen markant flest turer sammenlignet med de to andre. Enkelte syklister som har benyttet seg av Strava kan ha valgt å sykle på fortauene i dette området, men det kan også være en ukorrekt tilknytning til linjesegmentet for syklister som har syklet i sykkelfeltene, til tross for at disse tilhører veibanen. Dette er en usikkerhet i dataene fra Strava som er beskrevet nærmere i 2.2.3.

Figur 3.6: Ankerbrua i Oslo. Øverst: utsnitt fra Metroview av linjesegmenter som krysser Ankerbrua og Oslo kommunes sykkelteller "Eventyrbrua". Fargeverdiene på linjene representerer tettheten av data tilknyttet hver linje fra mai 2021. Nederst: bilde hentet fra ECO-Counter som viser registreringssensorenes plassering i sykkelfeltene [35].

For å styrke kredibiliteten i datasettet ble følgende kvalitetsvurderinger gjort trinnvis for hver av tellestasjonene:

1. Studere om det er flere nærliggende linjesegmenter til det linjesegmentet som representerer veien registreringssensoren er montert, og undersøke om disse har verdier som kan tyde på feil tildeling av data.

2. Studere gatebilder i Google Maps [31] og i ECO-Counter for å sammenligne registreringssensorenes fysiske plassering i forhold til de ulike veiene linjesegmentene representerer.

3. Studere omgivelsene rundt tellesensoren for å se om objekter, som høye bygninger og vegetasjon, kan ha negativ innvirkning på nøyaktigheten til posisjonsbestemmelsen på stedet.

Ut ifra nevnte vurderingspunkter ble hver stasjon gradert **A** eller **B**, der A-stasjoner ble ansett som *sikkert* og B-stasjoner som *noe mindre sikkert*. Ettersom graderingen er gjort basert på subjektive vurderinger og uten en absolutt terskel, er bruken av denne inndelingen begrenset til å kunne benyttes i forbindelse med sammenligning av korrelasjonsverdier.

For utvalget av sykkeltellere var også kompletthet i form av datasett fra et flertall av årene en faktor som ble vektet. Enkelte sykkeltellere var i løpet av perioden 2017-2021 midlertidig eller permanent demontert som følge av asfalteringsarbeid, utbedringer eller omlegging av veibanen. Det har også vært tilfeller hvor sykkeltellere har blitt demontert og flyttet av andre årsaker. Sykkeltelleren i Hoffsveien[1] ble demontert i oktober 2017 og flyttet 250 meter lengre syd i gaten. Den nåværende sykkeltelleren leverte sin første dagsmåling av syklister 20. juni 2018. Den tidligere sykkeltelleren hadde registreringssensorer installert på det vestvendte fortauet, og i begge kjøreretninger i veibanen. Den nåværende telleren har en registreringssensor montert i sørgående veibane, og en på adskilt gang- og sykkelvei øst for veibanen [35].

Data fra den nåværende sykkeltelleren er inkludert i datasettet, mens data fra den tidligere telleren er ekskludert. Dette skyldes at den tidligere telleren har målinger fra kun ett av de fem valgte årene, mens den nåværende telleren har data fra tre av årene. Registreringssensorenes plassering er også mer gunstig ved den nåværende sykkeltelleren med tanke på tilknytning til korrekt linjesegment. Fravær av muligheten for sammenligning, og dermed kvalitetskontroll av datasett mellom år, samt høy usikkerhet knyttet til valg av linjesegment gjør at data fra den tidligere telleren i Hoffsveien og sykkeltellere i tilsvarende tilfeller, er utelatt fra datasettet.

Etter en gjennomgang av alle sykkeltellere tilhørende Oslo kommune og Statens vegvesen ble data fra de aktuelle tellerne lastet ned, tilsendt eller notert manuelt på filformatet CSV. For samtlige sykkeltellere ble tilhørende linjesegmenter valgt og data for hvert av de fem årene lastet ned. Dette ble gjort gjennom *Data*-fanen i Metroview. En gjennomgang av datasettene fra Strava viste at det er betydelige mangler i dataene fra 2017. Linjesegmenter tilknyttet flere tellerstasjoner har ingen eller få dager med data fra dette året. På bakgrunn av dette er alle data fra mai 2017 utelatt fra utvalget i denne studien.

[1] Gate mellom Skøyen og Smestad.

Følgende antall sykkeltellere med tilgjengelig data fra mai måned for årene 2018, 2019, 2020 og 2021, og som dermed danner datagrunnlaget i denne studien er:

2018 28 sykkeltellere totalt (23 fra Oslo kommune, 5 fra Statens vegvesen)

2019 36 sykkeltellere totalt (25 fra Oslo kommune, 11 fra Statens vegvesen)

2020 34 sykkeltellere totalt (26 fra Oslo kommune, 8 fra Statens vegvesen)

2021 36 sykkeltellere totalt (26 fra Oslo kommune, 10 fra Statens vegvesen)

3.2.3 Databehandling

Datasettene ble kvalitetskontrollert, tilpasset og slått sammen gjennom koding i Spyder og Jupyter Notebook. Koden ble skrevet i programmeringsspråket Python. I forbindelse med behandling og analyse av dataene i de to programmene ble Python-pakken *pandas* importert [47]. CSV-filene med data fra Strava og sykkeltellerne, ble lest inn på et todimensjonale tabellformat kalt *DataFrame*. DataFrame er den foretrukne måten å oppbevare data på ved bruk av pandas ettersom pakken inneholder flere integrerte funksjoner med høy ytelse for manipulering av hele, eller deler av et datasett. Pakken *NumPy* [48] ble importert og benyttet ved konvertering av datatyper, datalagring i matriser (*ndarray*) og i forbindelse med matematiske operasjoner. For visualisering av grafer og diagrammer ble funksjoner fra *matplotlib.pyplot* [49] anvendt.

Flere funksjoner ble kodet for å kontrollere, trekke ut og omstrukturere dataene. For datasettene fra Trafikkdata ble kun dager med dekningsgrad på 95 % eller mer inkludert. Krav til verdien under "Antall timer ugyldig" ble satt til å være absolutt null. For datasettene fra Strava ble antallet dager med data per linjesegment kontrollert. Ved et høyere antall rader med data per unike linje-ID ("edge_uid") enn 31[1] så kunne dette tyde på at data hadde blitt duplisert. I slike tilfeller ble datasettet det gjaldt undersøkt og korrigert om nødvendig.

Datasett fra Strava inneholder to verdier for aktivitetsregistrering; "forward_trip_count" og "reverse_trip_count". De aller fleste sykkeltellerstasjoner registrerer passeringer i begge retninger, både på gang- og sykkelveier, separate sykkelanlegg og i veier med blandet trafikk. I denne studien er det valgt å sammenligne totalverdier, summen av begge retninger fra de to datasettene. Dette er gjort for å utelukke feil knyttet til retningsbestemmelse av passerende syklister over registreringssensorene.

[1] Mai måned består av 31 dager.

Enkelte unntak er gjort, som i Hoffsveien der en registreringssensor er installert i den ene kjøreretningen i veibanen og den andre på en separat gang- og sykkelvei. I dette tilfellet ble det trukket ut to linjesegmenter som hver for seg representerte stedet en av sensorene er plassert. På gang- og sykkelveien blir det registrert trafikk i begge retninger, så her ble totalverdier beregnet. I veibanen der sensoren kun er ment å registrere sørgående syklister måtte plasseringen av linjesegmentets start- og endepunkt undersøkes. Dette ble gjort ved å importere datasettets tilhørende Shapefil i ArcGIS Pro, for deretter å studere retningen på det aktuelle linjesegmentet. For veibanen i Hoffsveien er linjesegmentet definert med startpunkt sør for, og endepunkt nord for, registreringssensoren. Som følge av dette ble kun verdiene fra "reverse_trip_count" benyttet, ettersom sensoren er plassert i veibanen med sørgående trafikk. Tilsvarende situasjon er også gjeldende for sykkeltelleren i Bærumsveien og på Ring 2 nord-øst for Majorstuen.

Både datasettene fra Strava og de to aktørene av sykkeltellere ble kontrollert for manglende verdier (NaN[1]-verdier). Ved manglende verdi(er) fra enten Strava og/eller sykkeltellere for en gitt dag, ble raden fjernet fra begge datasettene. Det samme ble gjort på dager der antallet registrerte passeringer for en sykkelteller var på 50 eller lavere (ref side 1237 i Livingston mfl. [36]). Tilfeller hvor det var registrert flere aktiviteter i Strava enn det var registrert syklister ved tilhørende tellestasjon på samme dag ble også utelatt. Typiske årsaker til dette kan være at passerende syklister ikke registreres av sykkeltelleren, eller at aktiviteter blir registrert ved et uhell i Strava. Det kan også være at store grupper med syklister passerer tellesensoren og den dermed ikke klarer å skille syklistene fra hverandre.

For sammenslåing av datasettene ble linjesegmentets ID (edge_uid) og navnet på sykkeltelleren benyttet som koblingsnøkkel. Dette ble gjort ved å importere en egen CSV-fil hvor alle sykkeltellerne var ført opp med tilhørende ID-verdi. Datasettene for sykkeltellerne ble iterert og hver tellerstasjon ble tildelt korrekt ID ut ifra koblingen i den eksterne CSV-filen. De to datasettene ble så slått sammen basert på ID-verdiene. Etter sammenslåing besto datasettet av kolonnene "Navn", "Dato", "Volum teller", "Volum strava" og "edge_uid". For en ryddigere struktur ble datasettet omgjort til en pivottabell (Tabell 3.1) med de to førstnevnte kolonnene som indeksering. Etter kvalitetskontroll og øvrig databehandling så besto tabellen av totalt 4045 rader (uten NaN-verdier) med data.

[1]Not-a-number

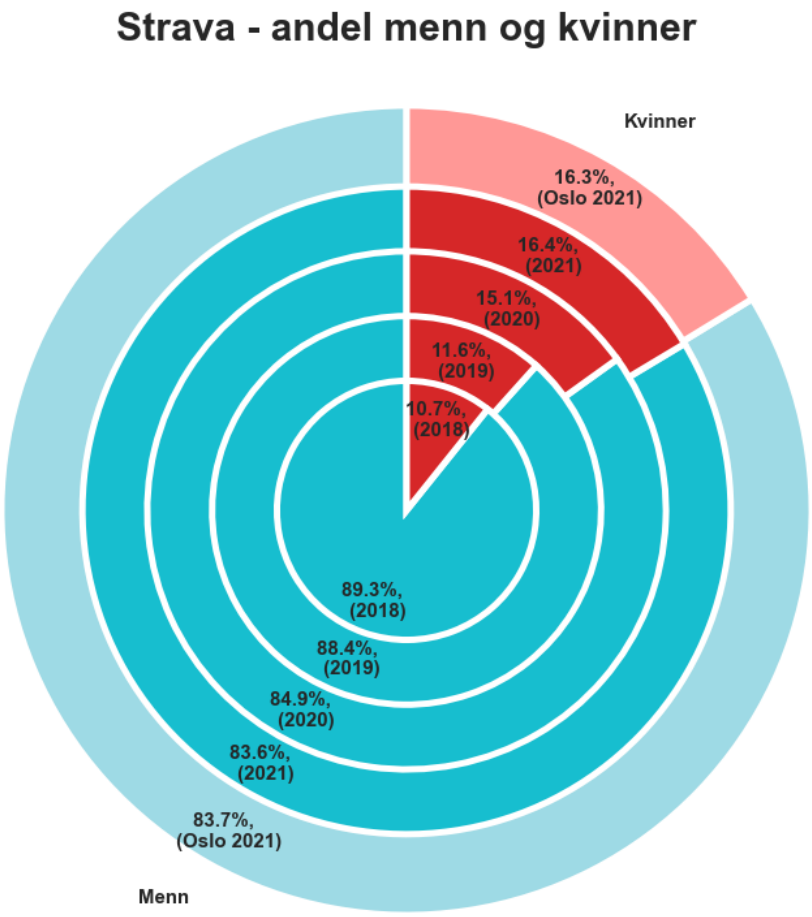
Tabell 3.1: Fem første radene med data fra pivottabell som resultatet av gjennomført databehandling. Pivottabellen er lagd ved bruk av *pandas.pivot_table* i Python.

Navn	Dato	Volum strava	Volum teller	edge_uid
AKERSYKE. SYKKEL 03	01.05.2018	15	151	182239380
	02.05.2018	130	1333	182239380
	03.05.2018	85	1052	182239380
	04.05.2018	130	1295	182239380
	05.05.2018	75	678	182239380

3.2.4 Utforskning av data

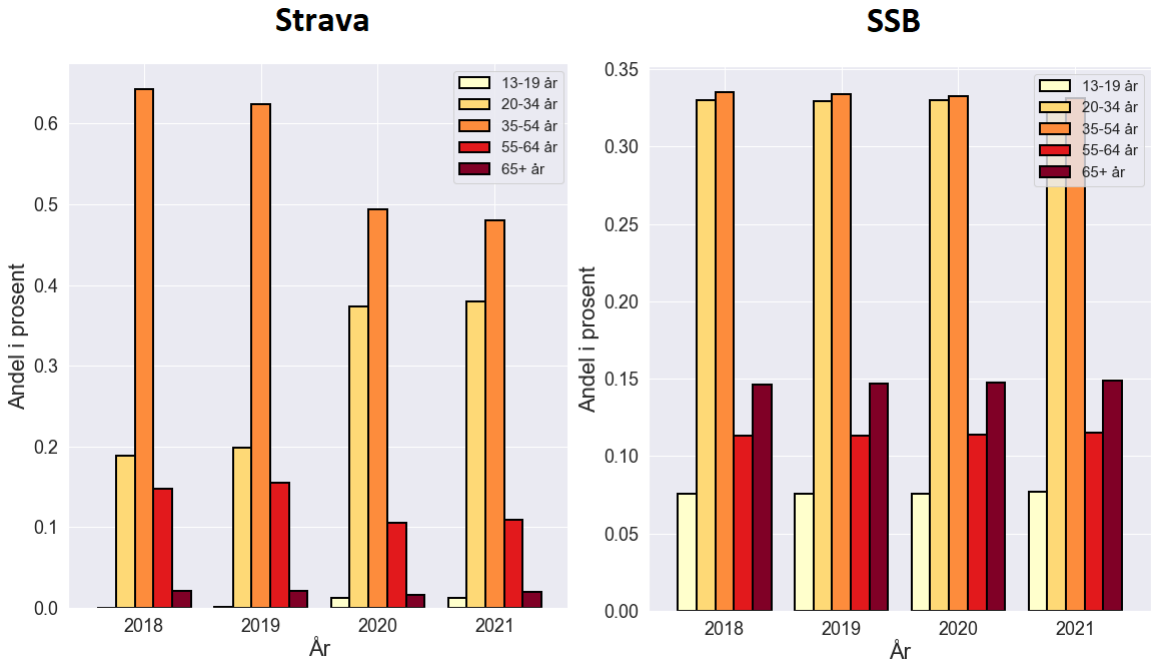
3.2.4.1 Strava

For å undersøke den demografiske fordelingen i dataene fra Strava ble kjønns- og aldersfordeling mellom de årlige datasettene studert. Det er gjennom tidligere studier slått fast at det er en demografisk skjevhet blant brukerne av Strava sammenlignet med populasjonen forøvrig. Dette er en begrensning som er ytterligere beskrevet i 2.2.3. Andelen mannlige og kvinnelige brukere i studiens data fra Strava kan er vist i figur 3.7. Representasjonen av kvinner i datasettene øker hvert år, fra 10,7 % i 2018 til 16,6 % i 2021, med en markant økning fra mai 2019 til mai 2020. Sett opp imot data fra alle linjesegmentene i mai 2021 så er fordelingen blant de utvalgte linjene samme år tilnærmet identisk. Dette indikerer at de utvalgte linjesegmentene er representative for all sykkelaktivitet på Strava i Oslo fra samme periode med tanke på kjønnsfordeling. Fra holdningsundersøkelsen til Opinion [40] kommer det fram at 35 % av alle spurte menn sier de sykler ofte, mens blant de kvinnelige deltagerne svarer 27 % det samme. For de som aldri sykler, eller nesten aldri sykler er andelene tilnærmet motsatt ved at 38 % av kvinnene avga dette svaret mot 26 % av mennene. Dataene fra Strava viser samme trend, bare i forsterket skala.



Figur 3.7: Andelen menn og kvinner i de ulike datasettene fra Strava presentert som et flerlags sektordiagram. Mørke sirkler (fra innerst: 1-4) representerer fordelingen basert på linjesegmentene i datasettet fra sitt respektive år. Den lyse sirkelen (5) representerer fordelingen basert på data tilknyttet sykkelaktiviteter fra mai 2021 for alle linjesegmenter i Oslo (AOI).

Figur 3.8 viser aldersfordelingen i datasettene fra Strava, og den reelle fordelingen basert på tall hentet fra SSB. Det er en parvis likhet mellom årene 2018-2019 og 2020-2021 blant alderfordelingen i datasettene fra Strava. Det er signifikant mindre avvik i dataene fra Strava sett opp imot tallene fra SSB for aldersgruppene 20-34 og 35-54 år i 2020 og 2021 sammenlignet med de to årene før. Brukere av Strava i alderen 35-54 utgjør hele 60 % av den totale brukermassen i 2018 og 2019, mens de i alderen 20-34 år utgjør rett i underkant av 20 % samme år. I 2020 og 2021 er derimot forskjellen i andelene mellom de to nevnte gruppene betraktelig redusert ved at andelen 35-54 åringer har minket, og andelen 20-34 åringer har økt. Brukerne mellom 55-64 år utgjorde i overkant av 10 % i 2020 og 2021, bare et par prosentpoeng fra "fasiten" til SSB. Samme aldersgruppen representerte en høyere andel de to første årene, omkring 15 %. Aldersgruppene 65+ og 13-19 år er sterkt underrepresentert i datasettene fra alle de fire årene. Sett under ett representerer datasettene fra 2020 og 2021 likevel Oslos befolkning på en bedre måte enn datasettene fra 2018 og 2019.



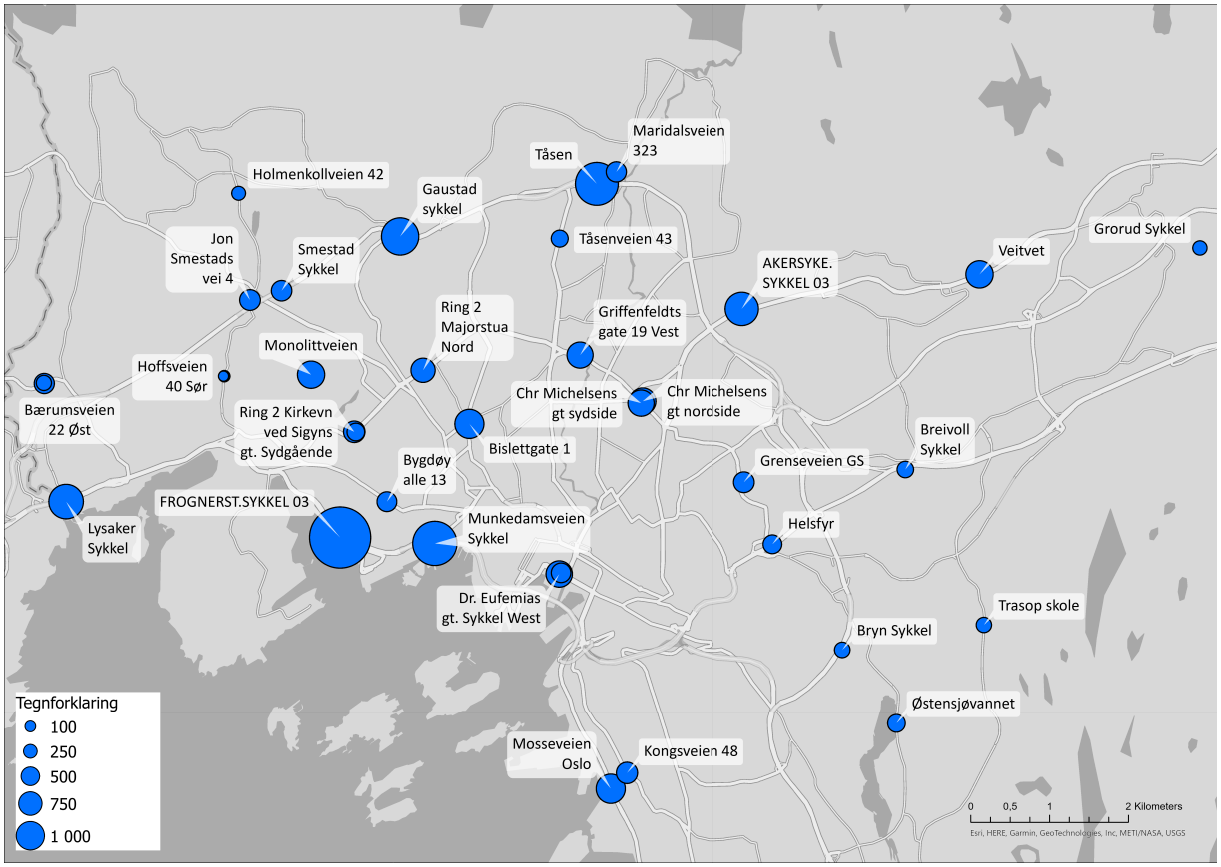
Figur 3.8: Søylediagram over aldersfordeling. Fra venstre: aldersrepresentasjonen i data fra linjesegmentene tilknyttet sykkeltellerne, alderfordelingen for befolkingen i Oslo (Data: *Statistisk sentralbyrå* [38]).

I holdningsundersøkelsen [40] fra 2020 var det deltagere i alderen 18-25 og 65-79 år som hadde lavest svarprosent tilknyttet svarene "alltid" og "nesten alltid" på spørsmålet om vedkommende sykler i sommerhalvåret, henholdsvis 6 og 8 prosent. Dette forholdet mellom de to aldersgruppene er også å finne i studiens data fra Strava, til tross for at aldersintervallene ikke er direkte sammenlignbare. Sett svarene fra holdningsundersøkelsen og statistikken fra dataene fra Strava under ett så er det flest syklister, og dermed også brukere av Strava, i øvre to tredjedel av aldersspennet 20-34 år. Fra tilnærmet alle studier Lee og Sener [18] har gjennomgått kommer det frem at kjernealderen i datasett fra Strava er mellom 25 og 44 år, og at disse utgjør omtrent halvparten av brukerne. I datasettene benyttet i denne studien utgjør brukere av Strava i alderen 20-54 år omtrent 76-84 prosent av alle brukerne.

Til tross for en overvekt i brukere av Strava for alle år i alderen 20-54 år er det likevel en trend at differansen i andelene minker for alle aldersgrupper sett opp imot SSBs tall. Dette skyldes trolig at det var en markant økning i antallet unike brukere av Strava i Oslo de to seneste årene sammenlignet med de to foregående. Henholdsvis 18 317 og 17 922 unike brukere registrerte en eller flere aktiviteter i mai 2020 og 2021 i Oslo, mot 13 631 og 12 683 i 2018 og 2019 [23]. Et bredere utvalg i form av flere unike brukere gir derfor en mindre skjevhet i datasettet knyttet til demografiske faktorer som alder- og kjønnsfordeling. Et slikt datasett vil kunne gjenspeile det totale antallet syklister i Oslo bedre og bidra til å bygge en mer korrekt modell.

3.2.4.2 Sykkeltellere

Tabellen etter databehandlingen ble importert i ArcGIS Pro som CSV-fil. Her ble verdier for gjennomsnitt, varians og standardavvik beregnet ved bruk av vektøyet *Summary Statistics* [50], for deretter å bli undersøkt. Nevnte verdier ble studert internt blant hver enkelte sykkelteller og linjesegment, samt i grupperinger basert på sykkeltellernes romlige tilhørighet til bydelene. Figur 3.9 viser gjennomsnittlig daglige passeringer for de ulike sykkeltellerne. Høyest antall passeringer per dag, 2691 syklister, har telleren "FROGNERST.SYKKEL 03" som registrerer syklister på gang- og sykkelveien ved Frognerstranda. Veien ligger parallelt med E18 og er en populær pendlerrute inn til Oslo sentrum fra vest. Færrest antall passeringer daglig har sykkeltelleren i sørgående veibane i Hoffsveien med et gjennomsnitt på 82 syklister. Snittet blant alle målinger fra samtlige sykkeltellere som er med i datasettet er på 761 syklister daglig. Ut ifra figur 3.9 er trenden at sykkeltellere med høyest antall daglige passeringer er plassert på gang- og sykkelveier tilknyttet ring- og europaveier, samt enkelte større gater i Oslo sentrum.



Figur 3.9: Antall passeringer i gjennomsnitt per sykkelteller basert på data fra alle fire år, visualisert som skalerte sirkler. Det er ikke benyttet psykologisk skalering (Flannery). Kartutsnittet er beregnet og fremstilt i ArcGIS Pro.

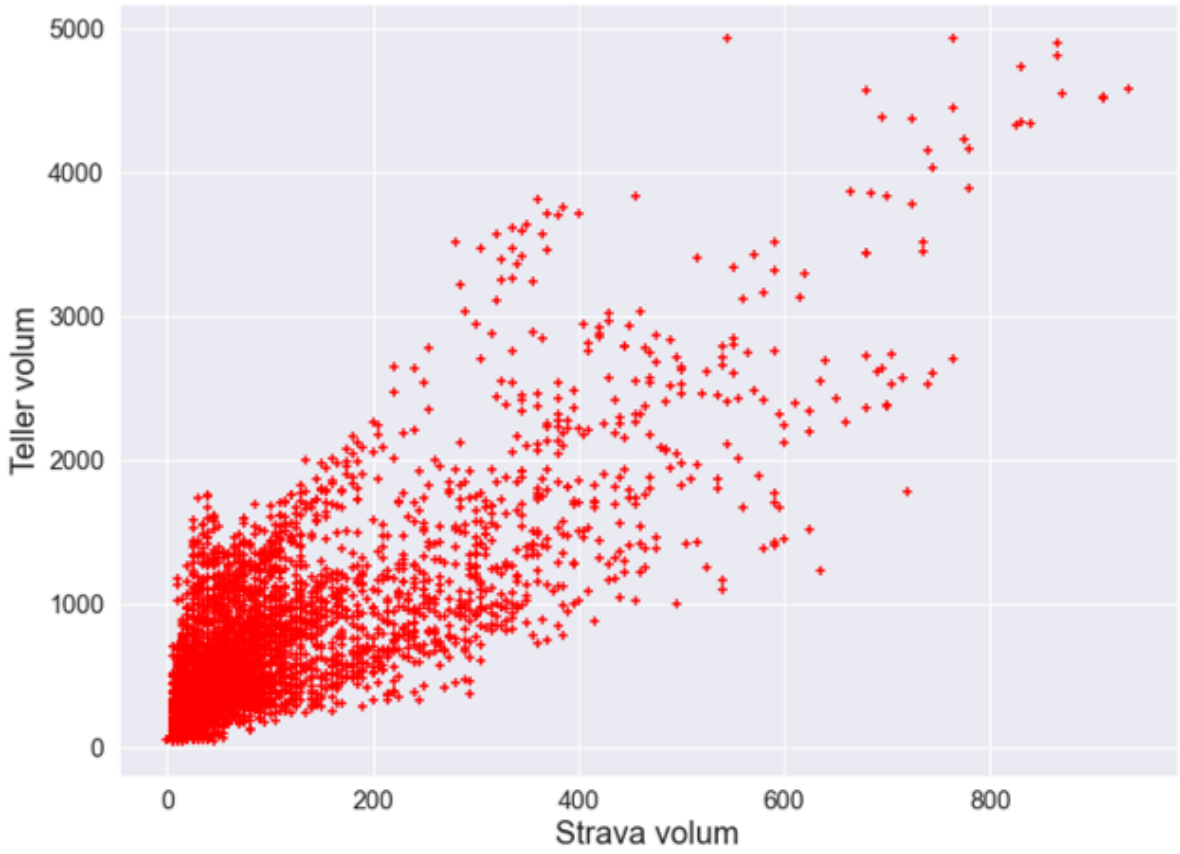
Data fra Strava for de fire årene viser at også linjesegmentet tilknyttet sykkeltelleren ved Frognerstranda hadde flest registrerte aktiviteter med et snitt på 485 per dag. Sykkeltelleren "Bygdøy allé 13" hadde færrest registrerte aktiviteter med et snitt på kun 14 syklister per dag. Dette gjenspeiles i figur 3.10 hvor forholdet mellom gjennomsnittsverdiene fra de to datakildene er sammenlignet. Bislettgata, hvor sykkeltelleren "Bislettgate 1" er plassert, har den laveste tettheten i antall brukere av Strava blant alle stedene i fokusområde 1. Sammen med sykkeltelleren i Bygdøy allé så skiller disse to seg markant ut fra resten i figuren. I Bislettgate 1 og Bygdøy allé er det henholdsvis 38 og 35 syklister per syklist som bruker Strava, men for alle øvrige stasjoner er samme verdi på under 20. Gjennomsnittet er på 9 syklister per syklist som bruker Strava, med et standardavvik på 7. Det store intervallet fra gjennomsnittsverdien, sett i sammenheng med standardavviket, kan indikere en systematisk underregistrering av syklister som benytter Strava i linjesegmentet for Bislettgate 1 og Bygdøy allé.

Figuren viser også høyere tetthet av registrerte aktiviteter i Strava for områdene utenfor sentrumsbydelene i Oslo. Ved å sammeligne tellepunktene "Smestad sykkel", "Gaustad sykkel" og "Tåsen", alle tilknyttet Ring 3, mot "Ring 2 Majorstua Nord", "Griffenfeldts gate 19 Vest" og "Chr Michelsens gt nordside" som er plassert langs Ring 2 (nærmere sentrum), så er intervallet kortere mellom hver syklist som bruker Strava i førstnevnte gruppe. Ses figur 3.10 og 3.9 i sammenheng er det flere syklister som bruker Strava i ytre deler av bykjernen enn nærmere sentrum.

Figur 3.10: Uklassifiserte, fargegraderte sirkler basert på forholdet mellom gjennomsnittet av sykkeltellerverdier og aktiviteter i Strava.

3.3 Korrelasjon

Det lineære forholdet mellom antallet aktiviteter i Strava og antallet syklister telt av sykkeltellerne ble studert på flere nivåer. Korrelasjonskoeffisient (r) ble beregnet for alle år med tilgjengelig data for samtlige tellerstasjoner. Dette ble gjort ved å benyttet funksjonen *pd.DataFrame.corr* [51] fra pandas som beregner korrelasjonen mellom to kolonner i en DataFrame basert på Pearson. Korrelasjonsverdier ble sammenlignet på tabellformat og gjennom visualisering i spredningsdiagram. Alle tellestasjonene hadde positiv korrelasjon mot aktivitetene i Strava. Laveste korrelasjon var det mellom datasettene i 2021 for vestgående veibane i Bærumsveien hvor r-verdien var 0,145. Høyest korrelasjon, og tett opp imot perfekt, var Frognerstranda i dataene fra 2019 hvor r-verdien var på 0,995. Alle observasjonene fra hele datasettet ble visualisert i matplotlib (figur 3.11) og den totale korrelasjonen ble beregnet til 0,785.



Figur 3.11: Spredningsdiagram med data fra Strava mot tilhørende data fra sykkeltellere. Datasett inkluderer kun observasjoner hvor r >0.

3.4 Modell

Utgangspunktet for den statistiske modellen er en multippel regresjonsmodell bygd på en kvantitativ og flere kvalitative variabler. Den kvantitative variabelen er antall sykkelaktiviteter i Strava **per dag**, og de kvalitative variablene definerer geografiske variasjoner. Modellen er bygd basert på data etter 3.2.3.

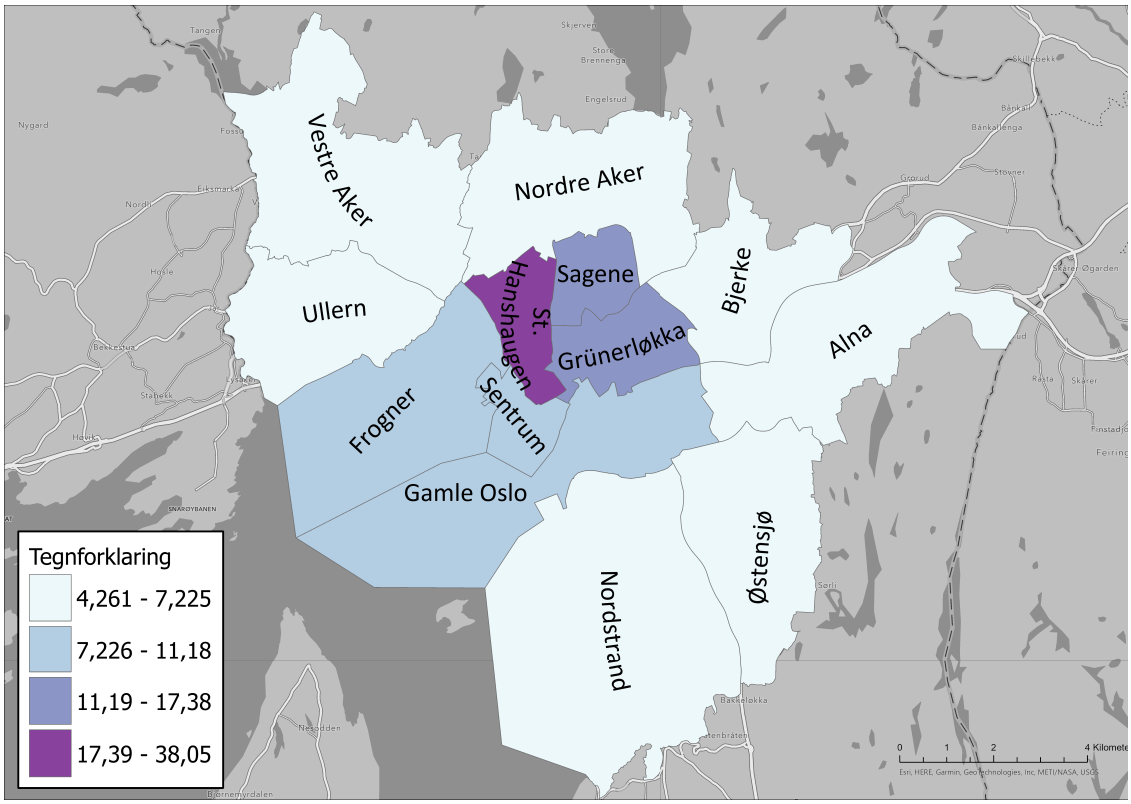
3.4.1 Kvalitative variabler

Kvalitative variabler er tilført modellen for å ta høyde for geografisk plassering ved estimering av antallet syklister. Figur 3.10 viser store variasjoner i forholdet mellom data fra Strava og sykkeltellerne for ulike deler av Oslo. De kvalitative variablene ble derfor definert ut ifra klassifisering av variasjonene i forholdet mellom Strava- og telledata med utgangspunkt i Oslos bydeler som geografisk avgrensing.

Til å finne forholdet mellom Strava- og sykkeltellerdata for hver av bydelene ble ArcGIS Pro benyttet. Fra "ArcGIS Server" ble kartlaget "Bydeler_Oslo" [52], bestående av bydelene i Oslo definert som polygoner, lastet ned og importert. →

En *Spatial join* mellom sykkeltellerne og bydelene ble gjennomført for å tildele alle stasjonene sin respektive bydel basert på den topologiske relasjonen *within*[1]. Gjennomsnittsverdien basert på alle sykkeltellerne innenfor hver bydel ble beregnet, og tilsvarende for linjesegmentene fra Strava. Gjennomsnittsverdiene ble lagt til bydelspolygonene ved bruk av *Join Field* [53]. Forholdet mellom gjennomsnittsverdiene for sykkeltellerne og Strava ble visualisert ved bruk av samme metode som i figur 3.10. Flere ulike klassifiseringsmetoder ble vurdert, samt antall klasser. Til slutt ble fire klasser definert basert på naturlig inndeling (2.1.1.2) i ArcGIS. Intervallene av klassifiseringen per bydel kan studeres i figur 3.12. Det er kun bydeler som inneholder en eller flere sykkeltellere som er inkludert.

Med fire klasser som grunnlag ble tre kvalitative variabler opprettet og definert for hver av bydelene. Variablene ble lagt til datasettet fra tabell 3.1 gjennom bruk av bydelsnavn som nøkkel. Hver tellestasjons verdier for de tre kvalitative variablene er å finne i tabell A.1 under *Tillegg A*. Datasettet ble deretter eksportert som CSV for videre analyse i RStudio.

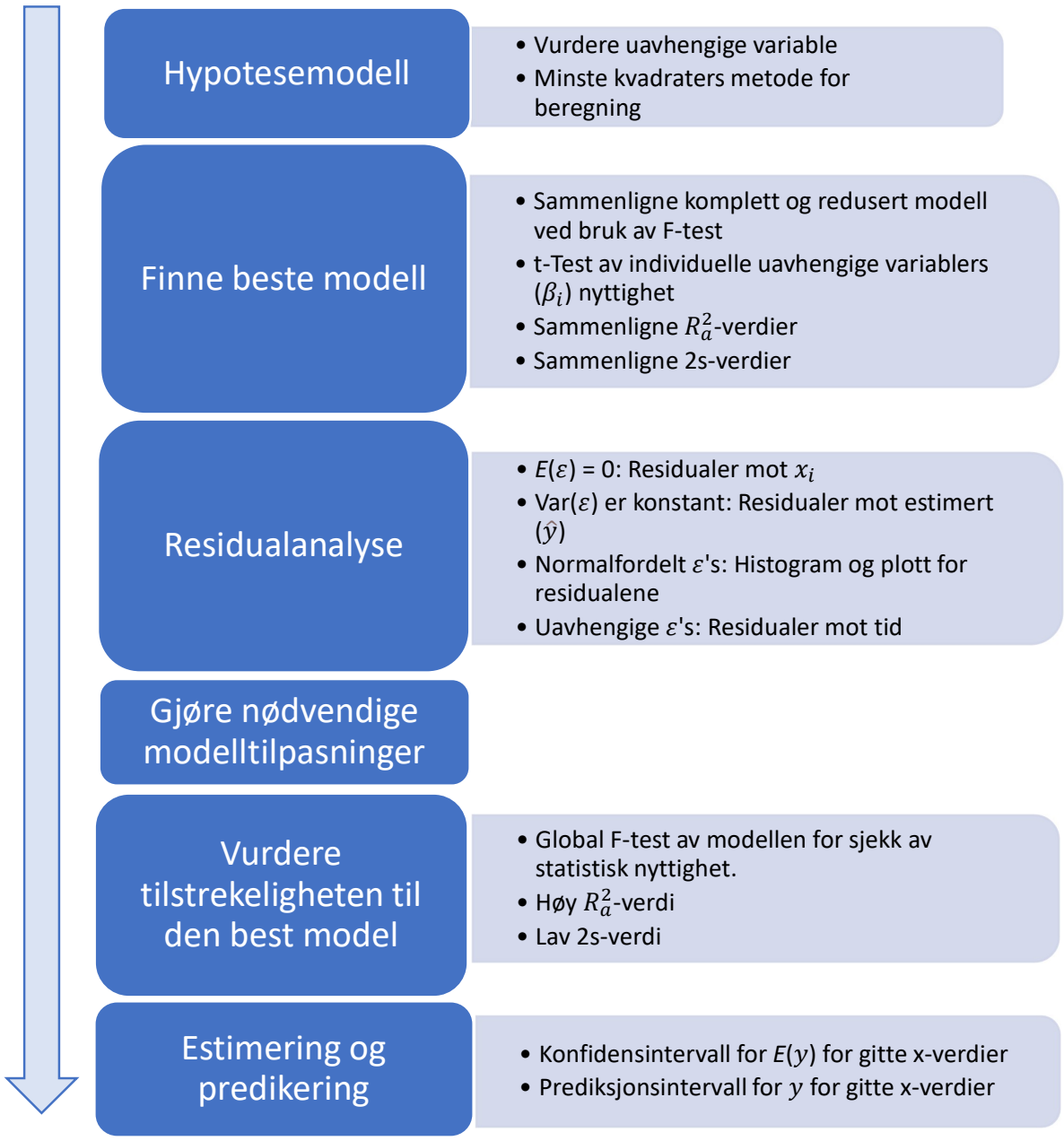


Figur 3.12: Klassifisering etter hvor mange syklister som i gjennomsnitt passerer en sykkelteller mellom hver gang en syklist som benytter Strava passerer. Alle sykkeltellere og linjesegmenter innenfor hver bydel er brukt i beregningen av gjennomsnittsverdier.

[1]Relasjonen er sann hvis objekt A (sykkelteller) er innenfor grensen til objekt B (bydelen).

3.4.2 Valg av modell

Prosessen *Guide to Multiple Regression A second course in statistics: regression analysis* [16] side 242 er benyttet som utgangspunkt for å komme fram til den beste modellen. Data ble lest inn i RStudio og en ordinær minste kvadraters regresjonsmodell ble tilpasset. Deretter ble trinnene i figur 3.13 fulgt kronologisk.



Figur 3.13: *Guide to Multiple Regression*

Etter prosessen med modellberegningen ble følgende to modeller valgt for bruk til estimering og sammenlikning; en ordinær minste kvadraters modell (3.1) og en kvadratrot-transformert modell (3.2).

$$\hat{y} = 132,652 + 4,257x_1 + 328,574x_2 + 540,228x_3 + 795,932x_4 + \varepsilon \tag{3.1}$$

$$\sqrt{\hat{y}} = 15,33 + 0,066x_1 + 5,643x_2 + 10,48x_3 + 14,7x_4 + \varepsilon \tag{3.2}$$

Fra residualanalysen ble det avdekket heteroskedastisitet[1] i residualene til den ordinære minste kvadraters modellen. Antagelsen om konstant varians for ε var dermed antatt å ikke være oppfylt. En **Breusch-Pagan Test**[2] ble gjennomført og nullhypotesen ble forkastet. Som følge av dette ble flere vektede lineære modeller og stabiliserende transformasjoner testet, deriblant kvadratrot-transformasjon. Enkelte av de vektede modellene hadde lavere RMSE sammenlignet med den ordinære lineære modellen. Den ordinære minste kvadraters modellen hadde derimot signifikant høyere R^2-verdi sammenlignet med de øvrige modellene. Tidligere studier, deriblant Hong, McArthur og Livingston [9], har vist at til tross for utfordringer med å oppfylle antagelsene så kan modeller likevel predikere jevne resultater for antallet syklister basert på data fra Strava. Derfor er begge modellene benyttet til videre beregning av syklister i fokusområde 2.

3.5 Beregning av antallet syklister

Etter fullført regresjonsanalyse i 3.4.2 ble antallet syklister beregnet for fokusområde 2. Siden sykkelfeltene ble oppført mellom august 2018 og august 2020 så ble tidsrommet 1. januar til 31. juli i 2018 og 2021 valgt for å representere perioden henholdsvis før og etter oppgraderingen. Data fra linjesegmenter i Åkebergveien, samt alle tilknyttede sidegater ble lastet ned med daglig aggregering fra Metroview for de to periodene. CSV-filene ble lest inn og omstrukturert ved bruk av koding i Python, etter samme prinsipp som i 3.2.3. Linjesegmentene ble undersøkt og slått sammen i de tilfellene hvor det forekom flere inndelinger mellom hvert veikryss. Ved sammenslåing av flere linjesegmenter ble gjennomsnittsverdien beregnet og tilført det nye linjesegmentet. For hver dag i de to periodene ble den totale mengden syklister estimert ved bruk av de to modellene.

[1] Når variansen til ε varierer basert på størrelsen til x-verdien. Motsatte av homoskedastisitet [16].

[2] Breusch-Pagan Test benyttes når det er mistanke om heteroskedastisitet. Følgende hypoteser gjelder for testen:

H_0: Homoskedastisitet er tilstede (variansen til residualene er konstant)

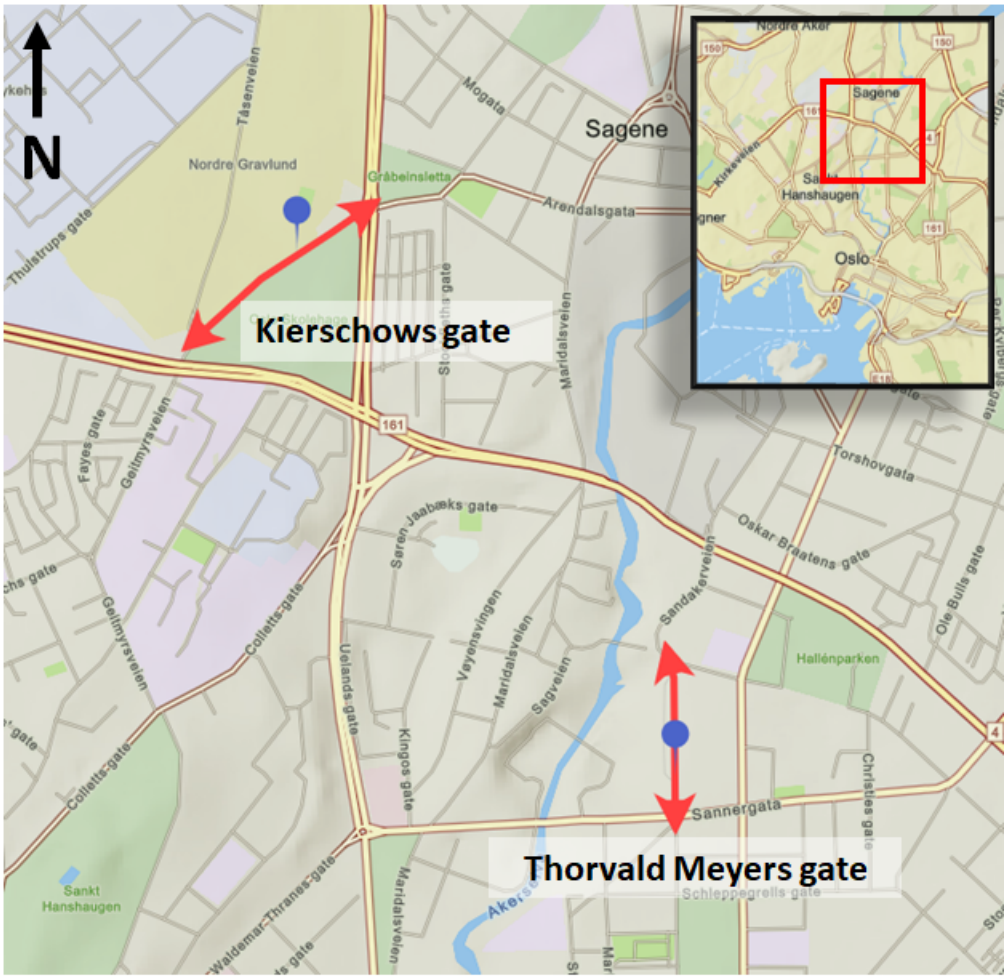
H_a: Heteroskedastisitet er tilstede (variansen til residualene er ikke konstant)

For å korrigere for vekslende faktorer som kan ha innvirkning på antallet syklister ble likning 1 fra Pritchard, Bucher og Frøyen[54] side 118 benyttet. Likningen (3.3) tar høyde for forvirringsfaktorer, deriblant vær- og sesongsvariasjoner mellom de to datasettene. Verdien som beregnes er en positiv eller negativ verdi som representerer differansen mellom antallet syklister fra perioden før til perioden etter sykkelfeltutbyggingen for et linjesegment x. Variabelene num_{x1} og num_{x2} representerer antallet syklister for linjesegmentet som studeres i henholdsvis perioden før (1) og etter (2) utbyggingen. Σnum_{N1} og Σnum_{N2} er summen av alle linjesegmentene i fokusområde for periode 1 og 2. Etter beregning av differansene for de ulike linjesegmentene ble verdiene visualisert i ArcGIS Pro. Data fra sykkeltelleren i Åkebergveien ble hentet fra ECO-Counter og benyttet som kontroll av resultatene.

$$\Delta Adjusted\ bicycle\ volume_{linjesegment\ x} = \left(\frac{num_{x2}}{\Sigma num_{N2}} - \frac{num_{x1}}{\Sigma num_{N1}}\right) \times \Sigma num_{N1} \tag{3.3}$$

3.6 Ekstra studieområder

På bakgrunn av resultatene fra beregningene gjort for fokusområde 2 var det behov for ytterligere testing av modellen. To nye områder ble valgt; en strekning på 350 meter i Kierschows gate mellom Ring 2 og Uelands gate, og en 250 meter lang del Thorvald Meyers gate mellom Sannergata og Biermanns gate. Områdenes plassering kan studeres i figur 3.14. Aktuell del av Kierschows gate tilhører St. Hanshaugen bydel. Thorvald Meyers gate er en del av bydelen Grünerløkka. For gjeldende område i Kirshows gate er det oppmerkede sykkelfelt i begge kjøreretninger for hele sin utstrekning. Studieområde i Thorvald Meyers gate er enveiskjørt for motorkjøretøy og har derfor kun nordgående trafikk. Det er likevel motstrøms sykkelfelt anlagt i hele strekingen for sørgående sykkeltrafikk . For syklister som sykler nordover er det sykkelfelt store deler av strekningen, bortsett fra de 40 første meterne fra Sannergata hvor det er parkeringsplasser. På denne delen må syklister sykle i veibanen. Begge gatene har installert sykkeltellere som markert i figur 3.14 med plassering 59°56'12.4"N 10°44'51.9"E og 59°55'43.1"N 10°45'32.7"E for henholdsvis Kirshows gate og Thorvald Meyers gate. Sykkeltellerne registrerer syklister i begge retninger og driftes av Oslo kommune. Resultater fra modellen ble kontrollert opp med data fra sykkeltellerne i begge områdene. Aktuell del av Kierschows gate tilhører St. Hanshaugen bydel. Thorvald Meyers gate er en del av bydelen Grünerløkka.



Figur 3.14: Plassering og utstrekning til studieområdene i Kierschows gate og Thorvald Meyers gate.

KAPITTEL

4

RESULTATER

I kapittelet *Resultater* presenteres svar fra beregninger og analyser som ble gjennomført i *Metode*. Resultatene er presentert i figurer og tabeller, og deretter beskrevet.

4.1 Korrelasjon

I 3.3 ble korrelasjonen mellom antallet syklister som brukte Strava og antall syklister totalt som passerte de 37 utvalgte sykkeltellerne beregnet for hver mai måned. Dette ble gjort for å kunne bidra til svar knyttet til forskningsspørsmål 1, og samtidig danne et grunnlag for forståelse av dataene før regresjonsanalysen. Alle korrelasjonsverdiene er ført inn i tabell 4.1. Verdiene er beregnet ut ifra Pearsons korrelasjonskoeffisient. Hver stasjon er oppnevnt med tildelt gradering fra 3.2.2 (**A** eller **B**) og stasjonene tilhørende Statens vegvesen er markert med "(SV)". Verdiene er gradert på farge hvor $r > 0,75$ er grønt, $0,75 > r > 0.5$ er gult og $r < 0,5$ er rødt.

Høyest gjennomsnittsverdi for alle tellestasjonene var i mai 2019. Samme verdi for 2018 er tilnærmet lik, bare 0,002 lavere. Medianverdien for 2018 er derimot høyere enn i 2019. For 2020 og 2021 var det mindre differanse mellom de respektive gjennomsnitts- og medianverdiene. Trenden for hele perioden er at gjennomsnittsverdien for korrelasjonsverdiene synker. Enkelte sykkeltellere, som "Bærumsveien 22 Øst", "Helsfyr" og "Grorud Sykkel", har stabilt høye verdier for alle fire år. For sykkeltelleren "Jon Smestads vei 4" har det derimot vært en nedgang i korrelasjon for alle årene etter 2018. Dette er gjeldende hos flere tellestasjoner og bidrar til trenden i nedgangen for de årlige gjennomsnittsverdiene. Det er likevel flere unntak, som for "Maridalsveien 323" og "Griffenfeldts gate 19 Vest" hvor høyeste korrelasjonsverdi er fra 2021 datasettet. Dette bidrar til å avkrefte mistanke om beregningsfeil eller feil ved datasettene.

Sykkeltellerne med svakest relasjon mellom Strava og telledataene var "Bærumsveien 22 Vest", "Hoffsveien 40 Nord" og "Hoffsveien 40 Sør". Felles for alle tre er at de har få registrerte aktiviteter i Strava. Maks antall aktiviteter i Strava for en dag er henholdsvis 55, 40 og 55 syklister. Dette er de laveste maksimalverdien for Strava, i tillegg til "Bygdøy alle 13" med 35 og "Bislettgate 1" med 50. De to sistnevnte tellestasjonene hadde, som tidligere nevnt i 3.2.4, et betydelig høyere forholdstall mellom data fra Strava og sykkelteller enn de øvrige tellestasjonene. Dette forklares av generelt få passeringer blant brukere av Strava og relativt høye passeringer totalt med 496 og 1045 i gjennomsnitt for henholdsvis "Bygdøy alle 13" og "Bislettgate 1". De to sistnevnte tellestasjonene hadde også lave korrelasjonsverdier, med enkelte unntak for "Bislettgate 1", målt mot verdiene til de øvrige tellerne.

Tabell 4.1: Korrelasjonsverdier beregnet mellom sykkeltellere og tilknyttede linjesegmenter.

Tellestasjon	2018	2019	2020	2021	Gj. snitt teller
AKERSYKE. SYKKEL 03 (**B**)(SV)	0.945	0.841	0.802	0.734	0.83
Smestad Sykkel (**A**)(SV)	0.974	0.934	0.822	0.635	0.841
Gaustad sykkel (**A**)(SV)	0.947	0.882	0.867	0.722	0.854
Bryn Sykkel (**A**)(SV)	0.938	0.912	0.858	NaN	0.903
Dr. Eufemias gt. Sykkel West (**B**)(SV)	0.946	0.916	NaN	0.732	0.865
Holmenkollveien 42 (**B**)	0.753	0.706	0.802	0.724	0.741
Bærumsveien 22 Øst (**A**)	0.985	0.967	0.928	0.919	0.95
Bærumsveien 22 Vest (**B**)	0.303	0.572	0.407	0.145	0.357
Jon Smestads vei 4 (**A**)	0.96	0.87	0.831	0.695	0.839
Bislettgate 1 (**B**)	0.750	0.807	0.729	0.587	0.718
R2 Kirkevn ved Sigyns gt. Syd (**A**)	0.876	0.889	0.862	0.811	0.86
R2 Kirkevn ved Sigyns gt. Nord (**A**)	0.941	NaN	0.800	0.785	0.842
Monolittveien (**A**)	0.963	0.955	0.94	0.885	0.936
Chr Michelsens gt nordside (**A**)	0.736	0.79	0.744	0.756	0.756
Tåsen (**B**)	0.953	0.900	0.926	0.719	0.875
Griffenfeldts gate 19 Vest (**A**)	0.670	0.761	0.805	0.84	0.769
Chr Michelsens gt sydside (**A**)	0.853	0.763	0.82	0.695	0.783
Maridalsveien 323 (**B**)	0.878	0.830	0.918	0.947	0.893
Veitvet (**A**)	0.666	0.893	0.852	0.841	0.813
Tåsenveien 43 (**A**)	0.741	0.545	0.745	0.680	0.678
Mosseveien Oslo (**B**)	0.932	0.941	0.781	0.895	0.887
Bygdøy alle 13 (**B**)	0.635	0.603	0.466	0.353	0.514
Kongsveien 48 (**B**)	0.816	0.785	0.642	0.651	0.724
Nordstrandveien 59 (**B**)	0.604	0.803	0.647	0.873	0.732
Helsfyr (**A**)	0.975	0.972	0.876	0.905	0.932
Grorud Sykkel (**A**)(SV)	NaN	0.907	0.947	0.899	0.918
Dr. Eufemias Gt. Sykkel East (**B**)(SV)	NaN	0.882	NaN	0.617	0.75
Breivoll Sykkel (**B**)(SV)	NaN	0.976	0.898	0.942	0.939
Lysaker Sykkel (**A**)(SV)	NaN	0.985	0.954	0.930	0.956
Munkedamsveien Sykkel (**A**)(SV)	NaN	0.981	NaN	0.950	0.966
FROGNERST.SYKKEL 03 (**A**)(SV)	NaN	0.995	0.933	0.964	0.964
Hoffsveien 40 Nord (**A**)	NaN	0.152	0.77	0.492	0.471
Hoffsveien 40 Sør (**A**)	NaN	0.494	0.399	0.51	0.468
Grenseveien GS (**A**)	NaN	0.947	0.85	0.805	0.867
Østensjøvannet (**B**)	0.948	0.879	0.759	0.704	0.822
Trosop skole (**B**)	0.270	0.670	0.685	0.682	0.577
Ring 2 Majorstua Nord (**B**)	0.953	0.845	0.815	0.76	0.843
Gjennomsnitt per år og totalt	0.818	0.82	0.791	0.743	0.795
Median per år og av gj. snitt	0.904	0.881	0.817	0.745	0.84

4.2 Regresjonsanalysen

4.2.1 Ordinær minste kvadraters modell

Hypotesemodellen, utgangspunktet for regresjonsanalysen, var en ordinær minste kvadraters modell, videre omtalt som OLS-modellen. Grunnlaget for den tilpassede modellen var to kvantitative variabelene "Volum_teller" og "Volum_strava", samt tre kvalitative variabler "x1", "x2" og "x3". Figur 4.1 viser statistikk for OLS-modellen.

En delvis F-test ble gjennomført for å sjekke nytten av de kvalitative variablene. Figur 4.2 viser henholdvis F- og p-verdien (grønn rubrikk) etter ANOVA. Basert på p-verdien kan null hypotesen forkastes, og med 99,9 % sikkerhet si at minst en av de kvalitative variablene utgjør et nytte for modellen med tanke på å redusere residualene. Basert på t-verdiene, og tilhørende p-verdier i figur 4.1 (oransje rubrikk) er det høy sannsynlighet ($>$ 99,9 %) for at samtlige av de uavhengige variabelene bidrar positivt til modellen med tanke på nytten. Justert R^2 (rød rubrikk) viser at modellen forklarer 72 % av variansen for datapunktene. Til sammenligning var justert R^2 for den reduserte modellen på 0,62. Det estimerte standardavviket til modellen, RMSE, er på 351 (blå rubrikk). Den gjennomsnittlige usikkerheten i forbindelse med estimering av antallet syklister totalt basert på antallet syklister med Strava, ved gitt geografisk plassering, er dermed 351 syklister. For modellen uten de kvalitative variablene var RMSE = 413. Basert på modellens standardavvik vil en prediksjon av antallet syklister totalt for en ny gate, ut ifra et gitt antall syklister med Strava trolig kunne avvike med maksimalt 702 syklister fra modellen.

```
Call:
lm(formula = Volum_teller ~ Volum_strava + x1 + x2 + x3, data =

Residuals:
     Min       1Q   Median       3Q      Max
-1607.93  -183.51   -46.94   146.63  2146.63

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  132.65194    8.79644   15.08   <2e-16 ***
Volum_strava   4.25714    0.04255  100.05   <2e-16 ***
x1           328.57376   13.25713   24.79   <2e-16 ***
x2           540.22781   19.78106   27.31   <2e-16 ***
x3           795.93241   32.74036   24.31   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 351.7 on 4040 degrees of freedom
Multiple R-squared:  0.7214,    Adjusted R-squared:  0.7211
F-statistic:  2615 on 4 and 4040 DF,  p-value: < 2.2e-16
```

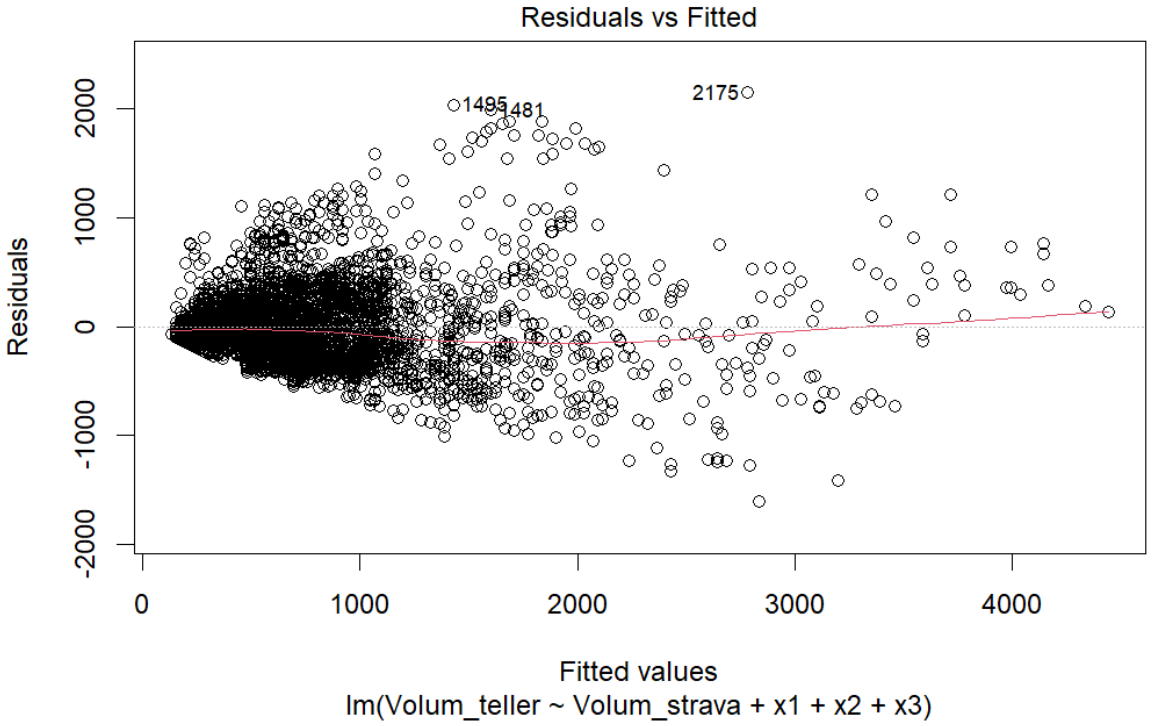
Figur 4.1: OLS-modell med tilhørende statistikk. Utskrift fra RStudio.

```
Analysis of Variance Table

Model 1: Volum_teller ~ Volum_strava
Model 2: Volum_teller ~ Volum_strava + x1 + x2 + x3
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1   4043 689514214
2   4040 499826027  3 189688187 511.07 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figur 4.2: Resultat fra delvis F-test av OLS-modell.

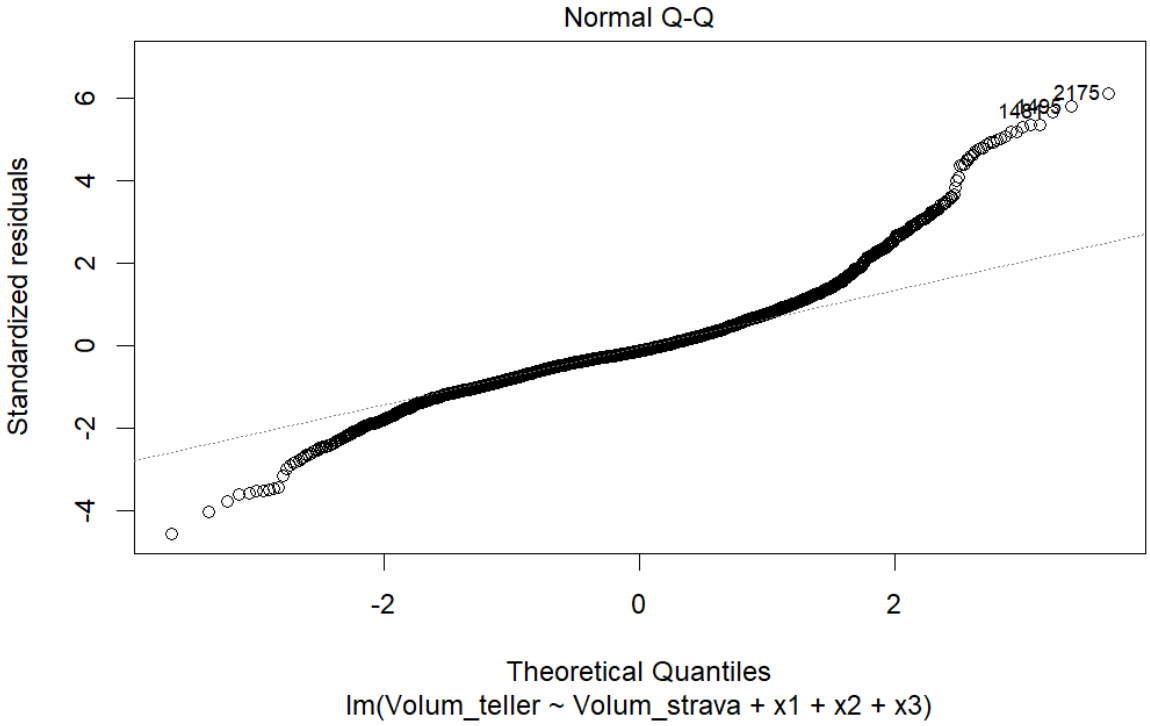
4.2.1.1 Residualanalyse

I residualanalysen av OLS-modellen ble gyldigheten av antagelsene om feilleddet ε undersøkt. I første omgang ble antagelse **1** og **2** fra 2.1.2.4 sjekket ved å studere plottet "Residuals vs Fitted" i figur 4.3. Residualene viser en høy grad av symmetri mellom området over og under linjen som representerer $\hat{\varepsilon} = 0$. Det er samtidig flere ekstremverdier blant de positive residualverdiene enn for de negative. Likevel er tyngdepunktet av residualene omkring $\pm$ 500 og disse bidrar til at ekstremverdiene har liten påvirkning på gjennomsnittsverdien av $\hat{\varepsilon}$. Plottet viser også at variansen øker med økende verdier av $\hat{t}$. Dette bryter med antagelse 2 om konstant varians for ε. Den positive stigning på linjen i "Scale-Location"-plottet (Figur A.1) underbygger dette.



Figur 4.3: Residualer mot estimerte y-verdier ($\hat{y}$) for OLS-modellen.

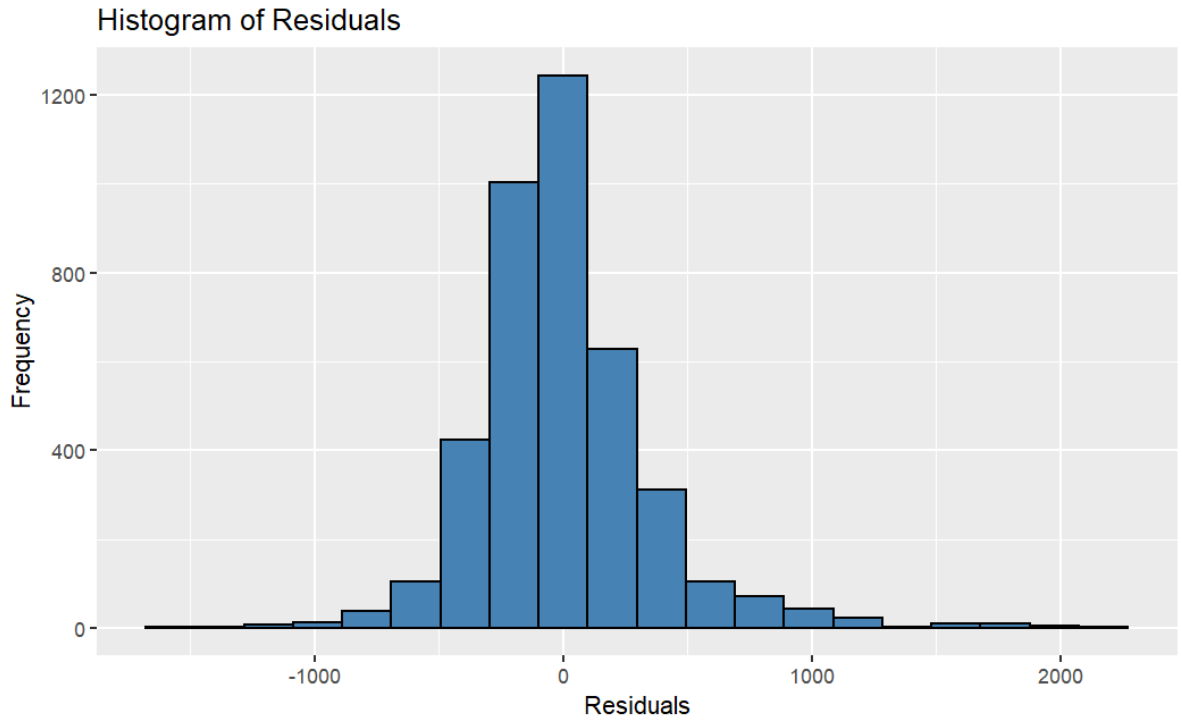
Antagelse **3** om normalfordeling av ε for OLS-modellen ble sjekket ved å studere "Normal Q-Q"-plottet i figur 4.4 og histogram over fordelingen av residualene i figur 4.5. Ved normalfordelte residualer ligger punktene i Normal Q-Q-plottet langs den stiplede linjen, med "haler" på hver ende av punktrekken som kan avvike *noe* fra linjen på grunn av naturlige variasjoner. For OLS-modellen har vi haler som avviker relativt mye fra linjen. Plottet viser også at det er flere *uteliggere* (engelsk: outliers) blant observasjonene ettersom det er flere punkter over tre i absoluttverdi for **standardiserte residualer**[1]. Statistikk i RStudio viste at 69 observasjoner i datasettet var å regne som uteliggere. For å undersøke innflytelsen disse hadde på estimeringen av β-parameterne i modellen ble **Cook's distanse**[2] beregnet og plottet i figur A.2. Siden uteliggerne tilsammen representerte under 2 % av alle observasjonene, og den høyeste verdien av Cook's distanse for en observasjon var 0,03, ble ingen av disse fjernet fra datasettet. Histogrammet over residualene skal være tilnærmet normalfordelt. For OLS-modellen er det flest residualer omkring null. Det er flere observasjoner med residualer i intervallet [-500, 0] enn for [0, 500], henholdsvis omtrent 1420 og 920 observasjoner (sett bort ifra observasjonene tett på null). Histogrammet viser, i likhet med Normal Q-Q-plottet, at det er flere og mer ekstreme uteliggere med positive residualverdier enn negative.



Figur 4.4: Normal Q-Q-plott med fordelingen av ε for OLS-modellen.

[1]Observasjon i har standardisert residualverdi $z_i = \hat{\varepsilon}_i/s = (y_i - \hat{y}_i/s)$. Dersom $|z_i| > 3$ for en observasjon i er den å regne som en uteliggere [16].

[2]Observasjon i har Cook's distanse (engelsk: Cook's distance) $D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1)MSE}\left[\frac{h_i}{(1-h_i)^2}\right]$. Observasjon i har en stor innflytelse på β-parameterne dersom $D_i > 1$.



Figur 4.5: Histogram med frekvensfordeling av residualene i OLS-modellen.

Antagelse **4** om ε er uavhengig med hensyn på tid ble testet med **Durbin-Watson Test**[1]. d- og p-verdien fra testen var på henholdsvis 0,72 og 0. Dermed ble nullhypotesen forkastet og autokorrelasjon mellom residualene i modellen var påvist. Autokorrelasjon mellom observasjoner er å forvente ettersom datasettet består av tidsseriedata[2] med registrerte verdier fra samme måned over flere år.

Som en følge av ikke-konstant varians for ε ble en Breusch-Pagan Test gjennomført for OLS-modellen. p-verdien fra testen var tett på 0 og dermed ble null-hypotesen om konstant varians blant residualene forkastet. Spredningsmønsteret til residualene i figur 4.3 indikerte av observasjonene hadde en *Poisson* fordeling. Derfor ble en ny modell med kvadratrot-transformasjon (beskrevet i 2.1.2.9) estimert i RStudio.

[1]Durbin-Watson test sjekker graden av autokorrelasjon (korrelasjon på tvers av ulike observasjoner i datasettet) mellom residualene i datasettet.

H_0: Det er ingen korrelasjon mellom residualene.

H_a: Residualene er autokorrelerte.

Durbin-Watson beregnes ved $d = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2}$ hvor $0 \leq d \leq 4$, der $d \approx 2$ betyr ingen korrelasjon, $d < 2$ betyr positiv korrelasjon og $d > 2$ betyr negativ korrelasjon. H_0 forkastes dersom $\alpha > p$-verdi [16].

[2]Tidsseriedata er data som er samlet i serie over et gitt tidsintervall.

4.2.2 Kvadratrot-transformert modell

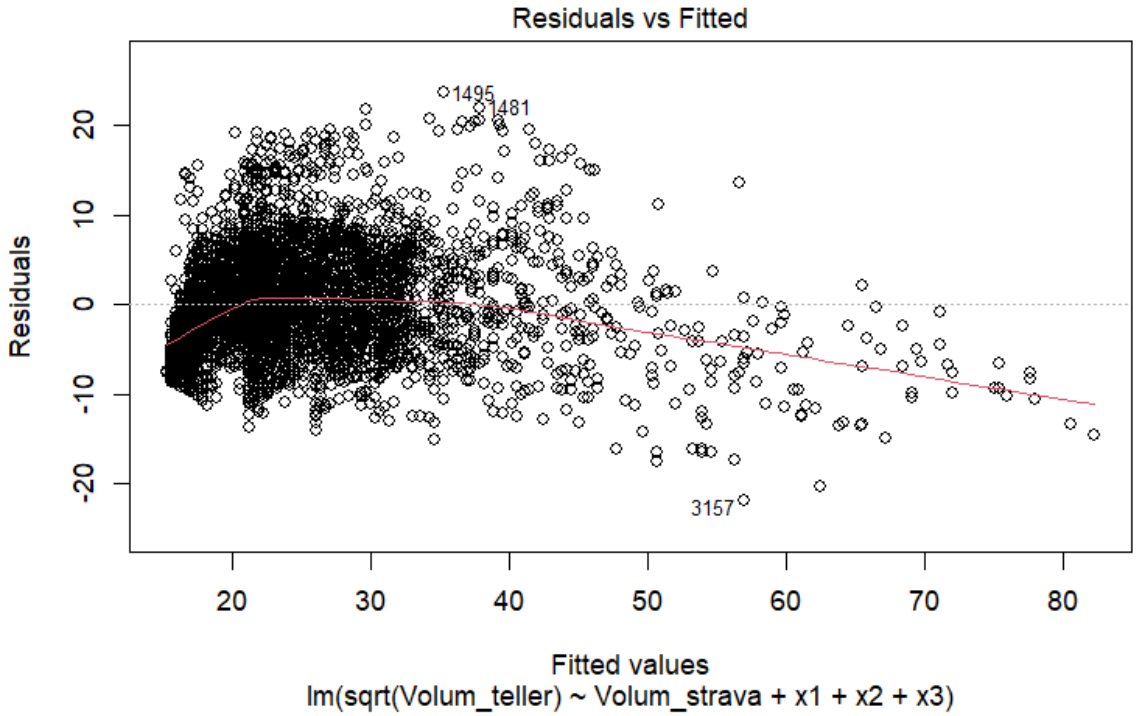
Den transformerte modellen, videre omtalt som SQRT-modellen, ble estimert i RStudio. Modell og statistikk er vist i figur 4.6. Siden responsvariabel $\hat{y}$ opptrer som kvadratroten av den estimerte y-verdien er β-parameterne ikke sammenlignbare med OLS-modellens β-parametere. Resultatet fra den globale F-testen (nederste linje i figur 4.6) viser at minst én av modellens estimerte parametere er nyttige i å predikere antallet syklister. De individuelle t-testene, med tilhørende p-verdier, bekrefter at ingen av de uavhengige parameterne er null. Siden det kun er fire parametere som testes er sannsynligheten lav for Type 1 feil[1]. SQRT-modellen har en justert R^2-verdi på 0,68. Modellen fanger dermed opp omtrent 4 % mindre av variasjonen blant observasjonene enn det OLS-modellen gjør.

```
Call:
lm(formula = sqrt(Volum_teller) ~ Volum_strava + x1 + x2 + x3,
    data = teller_strava_data)

Residuals:
    Min      1Q  Median      3Q     Max
-21.915  -4.324  -0.603   3.877  23.575

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.533e+01  1.512e-01  101.37   <2e-16 ***
Volum_strava 6.556e-02  7.314e-04   89.64   <2e-16 ***
x1           5.643e+00  2.279e-01   24.77   <2e-16 ***
x2           1.048e+01  3.400e-01   30.81   <2e-16 ***
x3           1.470e+01  5.628e-01   26.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.046 on 4040 degrees of freedom
Multiple R-squared:  0.6826,    Adjusted R-squared:  0.6822
F-statistic:  2172 on 4 and 4040 DF,  p-value: < 2.2e-16
```

Figur 4.6: Estimert SQRT-modell med tilhørende statistikk. Utskrift fra RStudio.

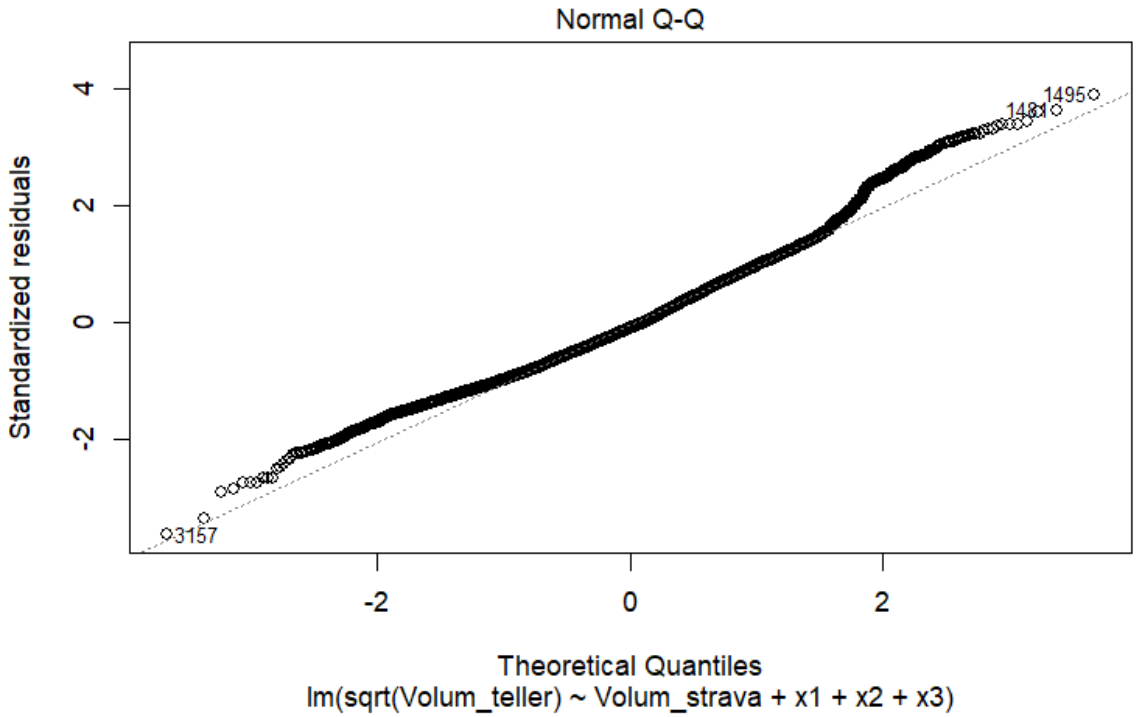
[1]Type 1 feil er at H_0 forkastes til tross for at den er korrekt. Et tilfelle av falsk-positiv.

"Residuals VS Fitted"-plott for SQRT-modellen er presentert i figur 4.7. Her vises tydelig effekten av transformasjonen på variansen for $\hat{\varepsilon}$. Residualene er mer spredt ut på lave verdier av $\sqrt{\hat{y}}$ og variansen er mer konstant sammenlignet med OLS-modellen. En bredere spredning av residualene mot OLS-modellen vises også i histogrammet over residualfordelingen i figur 4.9. Antagelsen om konstant varians for ε er i større grad innfridd med SQRT-modellen. Plottet viser også at vekten av residualene mot modellen er mer varierende sammenlignet med OLS-modellen. Den rød linjen i "Residuals vs Fitted"-plottene viser trenden i residualene blant observasjonene. Det er en overvekt av negative residualer i intervallet [0,22] for $\sqrt{\hat{y}}$. Modellen vil predikere et høyere antall syklister enn det observasjonene skulle tilsi i dette intervallet. Det samme gjelder i tilfeller hvor verdien av $\sqrt{\hat{y}}$ er 40 eller høyere. Dette er et forventet resultat ettersom kvadratrot-transformasjonen komprimerer høye verdier av $\hat{y}$ og strekker ut lave verdier. Siden det er betydelig flere observasjoner med relativt lave verdier sammenlignet med høye, og SQRT-modellen er lineær så legger observasjonene i førstnevnte gruppe mye av føringen for modellens parametere fordi responsvariabelen er kvadratroten av $\hat{y}$ vil residualer med samme verdi utgjøre et betydelig større avvik i prediksjonen av antallet syklister på høyere verdier av $\hat{y}$ enn på lave. Til sammenligning er trenden i residualene for observasjonene mer jevn med OLS-modellen.



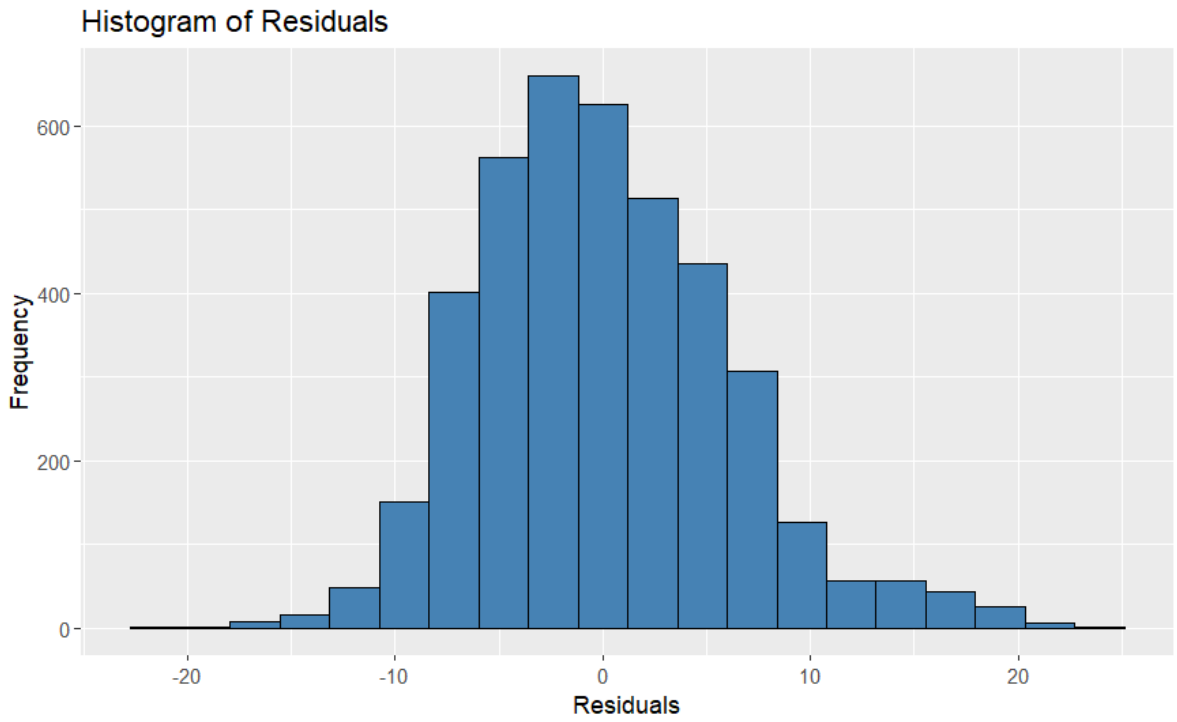
Figur 4.7: Residualer mot estimerte y-verdier ($\hat{y}$) for SQRT-modellen.

"Normal Q-Q"-plottet i figur 4.8 viser at residualene er tilnærmet normalfordelt for SQRT-modellen. Det er en større andel av residualene som ligger i ett med linjen sammenlignet med OLS-modellen. Halene i begge ender avviker mindre og består av færre ekstremverdier med lavere verdi. Enkelte av de mest negative verdiene for residualene er mindre avvikende enn forventet for en normalfordeling. I den positive enden er de mest positive verdiene noe mer forsterket enn forventet. Halene er likevel mye mindre markante enn ved samme plott for OLS-modellen (4.4). Blant observasjonene er det 32 uteliggere i SQRT-modellen. Dette er langt færre enn for OSL-modellen der 69 av observasjonene var uteliggere. Figur A.4 viser at Cook's distanse for observasjonene i SQRT-modellen er lavere enn OSL-modellen og uteliggerne har derfor en lav påvirkning på estimeringen av β-parameterne.



Figur 4.8: Normal Q-Q-plott med fordelingen av ε for SQRT-modellen.

Histogrammet over residualene i figur 4.9 bekrefter antagelsen om en mer korrekt normalfordeling av residualene sammenlignet med OLS-modellen. Det er en liten overvekt av lave negative residualer og høye positive residualer. Andelen lave negative residualverdier i histogrammet gjenspeiler konsentrasjonen på lave $\sqrt{\hat{y}}$ (< 20) i figur 4.7.



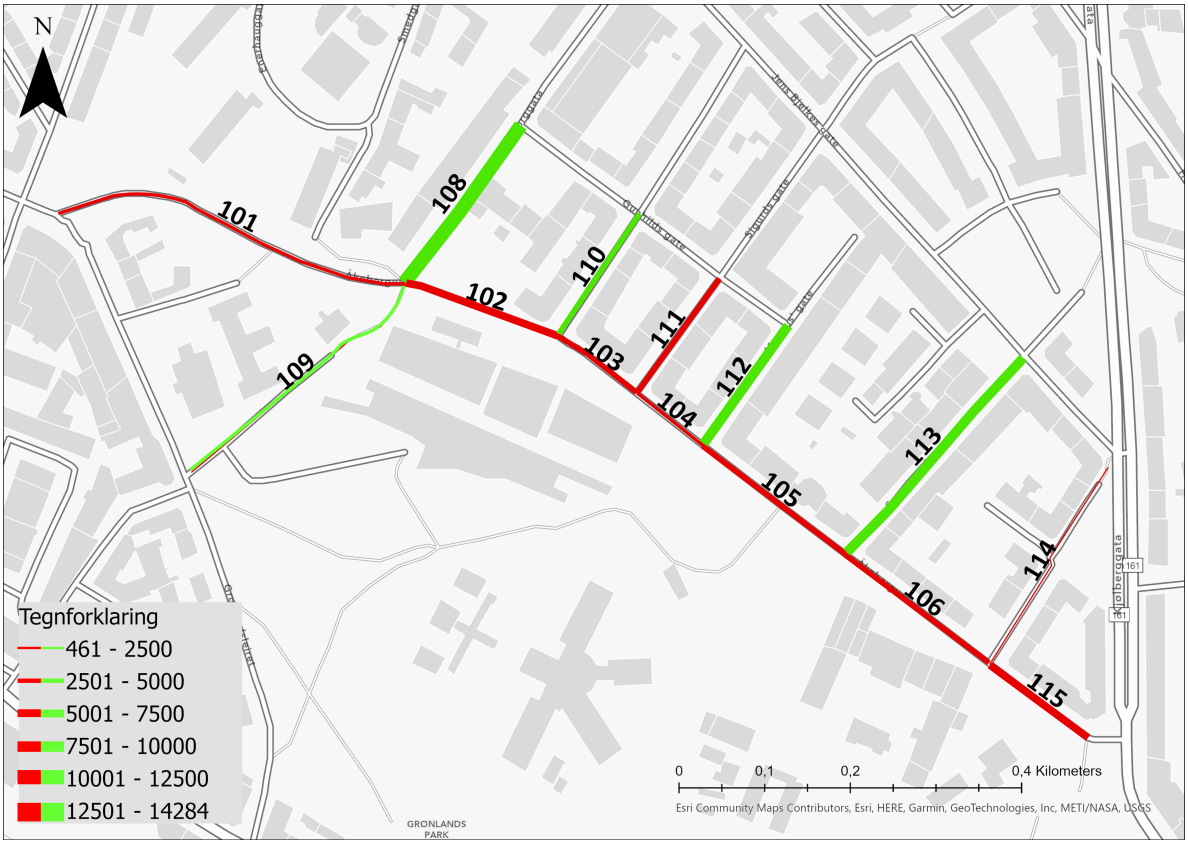
Figur 4.9: Histogram med frekvensfordeling av residualene i SQRT-modellen.

4.3 Resultater fra prediksjon av antallet syklister

4.3.1 Åkebergveien

Figur 4.10 viser differansene i antallet syklister mellom årene 2018 og 2021 for perioden 1. januar til 31. juli for de ulike linjesegmentene i Åkebergveien og tilknyttede gater. Differansene som er beregnet basert på de to modellenes predikerte verdier og formel 3.3 er vist i tabell 4.10. Verdiene som presenteres i figur 4.10 er fra OLS-modellens predikasjon. Tilsvarende kartutsnitt for SQRT-modellen er utelatt ettersom mange av verdiene faller inn under samme klasse som OSL-modellen. De verdiene som ikke er i samme klasse er merket med "*" i tabell 4.10.

Fra perioden før de opphøyde sykkelfeltene var bygget (periode 1), til perioden etter de var oppført (periode 2) predikerer begge modellene en nedgang i antall syklister for alle deler av Åkebergveien. For fem av syv sidegater er det derimot en økning i syklister i følge begge modellene. Borggata (linje 108) har den største økningen fra periode 1 til periode 2. En større nedgang for linje 102 kontra 101 kan tyde på at flere av syklistene som kommer syklende østover fra Grønlandsleiret på linje 101 oftere velger å sykle opp Borggata, og i enkelte tilfeller ned linje 109 enn videre på Åkebergveien i periode 2. Samtidig er økningen for linje 108 og 109 tilsammen betydelig høyere enn reduksjonen for 101 og 102. →



Figur 4.10: Beregnede differanser fra OLS-modellen i antallet syklister for Åkebergveien og nærliggende gater. Positive og negative verdier er markert henholdsvis grønt og rødt.

Tabell 4.2: Differansen mellom de to periodene i 2018 og 2021 for linjesegmentene i figur 4.10. De uthevede verdiene er linjesegmentene som en en del av Åkebergveien.

Linje	OLS-modell	SQRT-modell
101	**-3243**	**-3344**
102	**-7527**	***-6967**
103	**-5096**	***-4596**
104	**-4992**	**-4509**
105	**-5314**	***-4838**
106	**-5647**	***-4978**
108	14284	13064
109	2306	1931
110	6745	6271
111	-6310	-5990
112	11209	10425
113	11753	10871
114	-461	-445
115	**-7705**	***-6892**

Dette kan ha flere årsaker, som at syklister kan ha valgt andre ruter enn de som er studert i figur 4.10 fra det aktuelle krysset, eller har avsluttet aktiviteten innenfor et av linjesegmentene 108 og 109. Sett i lys av størrelsen på verdiene i tabell 4.2 er det lite trolig at de to nevnte årsakene forklarer den totale mengden "manglende" syklister mellom periode 1 og 2.

For alle linjesegmenter utenom 101 er absoluttverdien av differansene som baserer seg på prediksjoner fra SQRT-modellen lavere enn OLS-modellens i tabell 4.2. Høyeste og laveste differanse mellom differansene var for henholdsvis linje 108 og 114. Differanseverdien mellom differansene fra de to modellene økte som forventet mer enn størrelsen på differansene.

4.3.1.1 Kontroll mot sykkelteller

Data fra sykkelteller i Åkebergveien (beskrevet i 3.1.2), samt OLS- og SQRT-modellens prediksjon av det totale antallet syklister basert på data fra sykkeltellerens tilknyttede linjesegment er presentert i tabell 4.3. Der tabell 4.10 viser en beregnet nedgang i antallet syklister for linjesegmentet (105) fra 1. til 2. periode, så viser data fra sykkelteller derimot en økning på 55,6 %. Bak økningen på 56 266 syklister totalt mellom perioden 1 og 2 ligger en økning for alle månedene fra januar til juli. Både OLS-modellen og SQRT-modellen predikerer et høyere antall i de tre første månedene av periode 1 enn det tall fra sykkelteller tilsier. For resterende måneder i periode 1 (bortsett fra juli måned for OLS-modellen) predikerer begge modellene færre syklister enn det som er telt. Totalt for periode 1 predikerer OLS- og SQRT-modellen henholdsvis 4 514 og 13 034 færre syklister enn det som er telt for perioden. Disse tallene tilsvarer et periodisk prosentavvik på henholdsvis -4,5 og -12,9 prosent.

For periode 2 underpredikerer begge modellene antallet syklister sammenlignet med tall fra sykkelteller. Av de to modellene har SQRT-modellen størst avvik for samtlige måneder, med et periodeavvik på -33,5%. OSL-modellen har er mindre periodeavvik med -27,3 %, som tilsvarer 70 590 syklister. Der modellene i periode 1 blir "reddet" ved den periodiske sammenligningen gjennom blant annet overpredikering i enkelte måneder er ikke dette tilfelle i periode 2 hvor avvikene er betydelig høyere. Avvikene i periode 2 skyldes en kombinasjon av at et høyere antall syklister er registrert av sykkelteller, samt en nedgang i antallet syklister som benytter Strava i perioden sett under ett. For å avkrefte mistanke om feil i datasettet eller beregningsfeil i modellene ble data fra de to periodene undersøkt i Metroview. Metroview viste en nedgang i registrerte aktiviteter i Strava, fra 3510 i periode 1 til 2720 i periode 2. Trenden fra periode 1 til periode 2 er ulik med hensyn til antallet syklister som hadde passert tellepunktet sammenlignet med

antallet syklister som hadde registrert turen på Strava. Dette fører til at et i utgangspunktet negativt avvik i periode 1 utvikler seg til et forsterket negativt avvik i periode 2 som følge av færre syklister på Strava, men flere syklister totalt.

OLS-modellen predikerer høyere verdier enn SQRT-modellen for alle månedene i tabell 4.3. Siden begge modellene underpredikerer for en overvekt av månedene kommer derfor OLS-modellen nærmere sykkeltellerens verdi i begge periodene.

Tabell 4.3: Statistikk fra sykkelteller i Åkebergveien fra periode 1 og 2. Grønn og rød markering representerer henholdsvis positivt og negativt avvik fra data fra sykkelteller.

Måned	Sykkelteller	OLS-modell	Avvik OLS	%-avvik OLS	SQRT-modell	Avvik SQRT	%-avvik SQRT
jan.18	5253	11658	6405	121,9	10802	5549	105,6
feb.18	4295	10451	6156	143,3	9733	5438	126,6
mar.18	4991	10525	5534	110,9	9783	4792	96,0
apr.18	16201	14333	-1868	-11,5	13017	-3184	-19,7
mai.18	26810	16780	-10030	-37,4	14996	-11814	-44,1
jun.18	29043	17629	-11414	-39,3	15801	-13242	-45,6
jul.18	14630	15333	703	4,8	14057	-573	-3,9
Totalt 2018	101223	96709	-4514	-4,5	88189	-13034	-12,9
jan.21	11020	6393	-4627	-42,0	5974	-5046	-45,8
feb.21	10662	8355	-2307	-21,6	7809	-2853	-26,8
mar.21	18579	13719	-4860	-26,2	12705	-5874	-31,6
apr.21	25485	15542	-9943	-39,0	14195	-11290	-44,3
mai.21	27601	15215	-12386	-44,9	13851	-13750	-49,8
jun.21	39275	16651	-22624	-57,6	15094	-24181	-61,6
jul.21	24867	15538	-9329	-37,5	14316	-10551	-42,4
Totalt 2021	157489	91413	-66076	-42,0	83944	-73545	-46,7
Totalt begge år	258712	188122	-70590	-27,3	172133	-86579	-33,5

4.3.2 Thorvald Meyers gate

Data i perioden 1. mai 2018 til 30. april 2022 for linjesegmentet tilknyttet sykkeltelleren i Thorvald Meyers gate dannet grunnlaget for beregningene i tabell 4.4. For måneder hvor data manglet eller var av dårlig kvalitet ble disse utelatt fra estimering i modellene. Siden Thorvald Meyers gate ligger i bydel Grünerløkka har begge modellene dummy variabel $x_3 = 1$ og får dermed 540,228 og 10,48 som konstantledd for henholdsvis OLS- og SQRT-modellen.

Sett hele perioden under ett er avviket fra sykkeltellerens totalverdi på -16,1 og -18,8 prosent for henholdsvis OLS- og SQRT-modellen. Begge modellene underpredikerer også på årsnivå for samtlige år, med -13 % som minste avvik i 2021 (OLS-modellen). I vintermånedene (november-mars), med lav sykkeltrafikk, predikerer modellene forhøyede verdier sammenlignet med sykkelteller. I sommermånedene (april-oktober) underpredikerer modellene mer enn de overpredikerer i vintermånedene. →

Tabell 4.4: Statistikk for sykkeltelleren i Thorvald Meyers gate fra måneder med tilgjengelig data i perioden 2018-2022. Grønn og rød markering representerer henholdsvis positivt og negativt avvik fra data fra sykkeltelleren.

Måned	Sykkelteller	OLS-modell	Avvik OLS	%-avvik OLS	SQRT-modell	Avvik SQRT	%-avvik SQRT
mai.18	35573	24512	-11061	-31,1	23615	-11958	-33,6
jun.18	40349	25218	-15131	-37,5	24223	-16126	-40,0
jul.18	28861	25058	-3803	-13,2	24163	-4698	-16,3
aug.18	29918	25569	-4349	-14,5	24604	-5314	-17,8
sep.18	34468	24321	-10147	-29,4	23442	-11026	-32,0
okt.18	31452	24462	-6990	-22,2	23652	-7800	-24,8
nov.18	20164	23045	2881	14,3	22351	2187	10,8
des.18	9731	20199	10468	107,6	19658	9927	102,0
Totalt 2018	230516	192384	-38132	-16,5	185708	-44808	-19,4
aug.19	38738	24238	-14500	-37,4	23381	-15357	-39,6
sep.19	39587	23177	-16410	-41,5	22460	-17127	-43,3
okt.19	30916	23249	-7667	-24,8	22618	-8298	-26,8
nov.19	16766	21290	4524	27,0	20771	4005	23,9
des.19	9560	18696	9136	95,6	18295	8735	91,4
Totalt 2019	135567	110650	-24917	-18,4	107525	-28042	-20,7
jan.20	15072	22761	7689	51,0	22205	7133	47,3
feb.20	14935	21397	6462	43,3	20860	5925	39,7
mar.20	15041	23635	8594	57,1	22945	7904	52,5
apr.20	26117	25092	-1025	-3,9	24123	-1994	-7,6
mai.20	40013	26719	-13294	-33,2	25626	-14387	-36,0
jun.20	49522	26045	-23477	-47,4	24975	-24547	-49,6
jul.20	30474	24486	-5988	-19,6	23672	-6802	-22,3
aug.20	56167	26655	-29512	-52,5	25578	-30589	-54,5
sep.20	56160	25898	-30262	-53,9	24843	-31317	-55,8
okt.20	40358	24294	-16064	-39,8	23510	-16848	-41,7
nov.20	25100	22880	-2220	-8,8	22209	-2891	-11,5
des.20	14216	19603	5387	37,9	19154	4938	34,7
Totalt 2020	383175	289465	-93710	-24,5	279700	-103475	-27,0
jan.21	9931	22698	12767	128,6	22152	12221	123,1
feb.21	9843	19795	9952	101,1	19315	9472	96,2
mar.21	16289	23697	7408	45,5	22998	6709	41,2
apr.21	27608	24729	-2879	-10,4	23805	-3803	-13,8
mai.21	26325	25595	-730	-2,8	24649	-1676	-6,4
jun.21	43727	25752	-17975	-41,1	24700	-19027	-43,5
jul.21	28341	24273	-4068	-14,4	23486	-4855	-17,1
aug.21	47811	26254	-21557	-45,1	25206	-22605	-47,3
sep.21	43782	24470	-19312	-44,1	23579	-20203	-46,1
okt.21	32186	23550	-8636	-26,8	22871	-9315	-28,9
nov.21	26383	22302	-4081	-15,5	21722	-4661	-17,7
des.21	8902	16320	7418	83,3	16002	7100	79,8
Totalt 2021	321128	279435	-41693	-13,0	270485	-50643	-15,8
jan.22	9885	20673	10788	109,1	20256	10371	104,9
feb.22	11476	19434	7958	69,3	19012	7536	65,7
mar.22	23635	23466	-169	-0,7	22798	-837	-3,5
apr.22	28519	23751	-4768	-16,7	22969	-5550	-19,5
Totalt 2022	73515	87324	13809	18,8	85035	11520	15,7
Totalt alle år	1143901	959258	-184643	-16,1	928453	-215448	-18,8

Sammenligning av 2021 mot 2020 med -24,5% i årlig avvik, viser følgende:

- I 2021 er det i vintermånedene et **høyere positivt** avvik enn i 2020 grunnet færre syklister totalt i sesongen.
- I 2021 er det i sommermånedene et **lavere negativt** avvik enn i 2020 grunnet færre syklister totalt i sesongen.

Det gjennomsnittlige avviket for de tolv månedene er høyere i 2021 sammenlignet med 2020, men basert på de to nevnte forholdene over så er det totale avviket for året mindre i 2021 enn i 2020. I likhet med Åkebergveien predikerer OLS-modellen nærmere sykkeltellerens verdier i Thorvald Meyers gate enn det SQRT-modellen gjør.

4.3.3 Kierschows gate

Data fra samme periode som for Thorvald Meyers gate ble lastet ned for linjesegmentet ved sykkeltelleren i Kierschows gate. Den delen av Kierschows gate hvor sykkeltelleren er plassert ligger i bydel St. Hanshaugen som gir modellene dummy variabel $x_4 = 1$. Ved beregninger legger modellene til 795,932 og 14,7 som konstantledd for henholdsvis OLS- og SQRT-modellen.

Det er flere likheter og ulikheter mellom beregningene gjort for Thorvald Meyers gate og Kierschows gate i tabell 4.5. Trenden for hvilke måneder modellene avviker positivt og negativt fra sykkeltellernes verdier er, med enkelte unntak, tilnærmet lik som i Thorvald Meyers gate. I vintermånedene predikerer modellene over sykkeltelleren, og motsatt for sommermånedene. Også for Kierschows gate predikerer OLS-modellen mer korrekte verdier enn SQRT-modellen. For Kierschows gate predikerer begge modellene verdier med mindre avvik fra sykkeltelleren på både totalt og årlig nivå enn for Thorvald Meyers gate. For 2020, året hvor begge gatene har data fra samtlige måneder, gir de to modellene en mer nøyaktig prediksjon for året i Kierschows gate enn i Thorvald Meyers gate. Dette til tross for at antallet syklister totalt dette året er betraktelig høyere i Kierschows gate enn i Thorvald Meyers gate. For 2018 og 2019, som også består av data fra samtlige måneder, er avviket på kun -2,9 og -3,5 prosent for OLS-modellen og -6,2 og -5,6 prosent for SQRT-modellen i Kierschows gate. Tabell 4.3 viser at beregninger gjort for 2018 i Åkebergveien også resulterte i et lavt avvik på -4,5 %. For Thorvald Meyers gate samme år var avviket noe større med -16,5 %, men ikke mer enn 0,4 prosentpoeng fra det totale avviket for gaten. Det skal poengteres at de to sistnevnte gatenes totalavvik for 2018 ikke inneholdt data fra alle måneder.

Tabell 4.5: Statistikk for sykkeltelleren i Kierschows gate fra måneder med tilgjengelig data i perioden 2018-2022. Grønn og rød markering representerer henholdsvis positivt og negativt avvik fra data fra sykkeltelleren.

Måned	Sykkelteller	OLS-modell	Avvik OLS	%-avvik OLS	SQRT-modell	Avvik SQRT	%-avvik SQRT
jan.18	11041	28577	17536	158,8	27708	16667	151,0
feb.18	8959	24160	15201	169,7	23430	14471	161,5
mar.18	10297	29268	18971	184,2	28383	18086	175,6
apr.18	31125	32293	1168	3,8	31366	241	0,8
mai.18	55462	37307	-18155	-32,7	36554	-18908	-34,1
jun.18	57251	34174	-23077	-40,3	33375	-23876	-41,7
jul.18	37131	33245	-3886	-10,5	32278	-4853	-13,1
aug.18	53890	35666	-18224	-33,8	34779	-19111	-35,5
sep.18	46967	33039	-13928	-29,7	32124	-14843	-31,6
okt.18	38327	33180	-5147	-13,4	32217	-6110	-15,9
nov.18	24519	29219	4700	19,2	28326	3807	15,5
des.18	10947	24606	13659	124,8	23862	12915	118,0
Totalt 2018	385916	374734	-11182	-2,9	364402	-21514	-5,6
jan.19	13014	28490	15476	118,9	27626	14612	112,3
feb.19	11464	26442	14978	130,7	25641	14177	123,7
mar.19	19075	28350	9275	48,6	27494	8419	44,1
apr.19	36179	32911	-3268	-9,0	31998	-4181	-11,6
mai.19	48781	35496	-13285	-27,2	34626	-14155	-29,0
jun.19	47592	34506	-13086	-27,5	33653	-13939	-29,3
jul.19	35156	33071	-2085	-5,9	32106	-3050	-8,7
aug.19	53126	34413	-18713	-35,2	33482	-19644	-37,0
sep.19	47990	31555	-16435	-34,2	30662	-17328	-36,1
okt.19	35595	31313	-4282	-12,0	30375	-5220	-14,7
nov.19	19268	26030	6762	35,1	25234	5966	31,0
des.19	10900	22515	11615	106,6	21835	10935	100,3
Totalt 2019	378140	365092	-13048	-3,5	354732	-23408	-6,2
jan.20	17413	26826	9413	54,1	26009	8596	49,4
feb.20	16842	28786	11944	70,9	27913	11071	65,7
mar.20	21614	31058	9444	43,7	30135	8521	39,4
apr.20	38790	34697	-4093	-10,6	33852	-4938	-12,7
mai.20	55705	37286	-18419	-33,1	36522	-19183	-34,4
jun.20	68109	37867	-30242	-44,4	37316	-30793	-45,2
jul.20	34157	33860	-297	-0,9	32927	-1230	-3,6
aug.20	72119	38944	-33175	-46,0	38351	-33768	-46,8
sep.20	74954	36294	-38660	-51,6	35559	-39395	-52,6
okt.20	54398	33713	-20685	-38,0	32761	-21637	-39,8
nov.20	34255	30868	-3387	-9,9	29932	-4323	-12,6
des.20	18170	27392	9222	50,8	26560	8390	46,2
Totalt 2020	506526	397591	-108935	-21,5	387837	-118689	-23,4
okt.21	41127	32647	-8480	-20,6	31676	-9451	-23,0
nov.21	34405	30335	-4070	-11,8	29414	-4991	-14,5
des.21	11794	25109	13315	112,9	24350	12556	106,5
Totalt 2021	87326	88091	765	0,9	85440	-1886	-2,2
jan.22	14394	27244	12850	89,3	26418	12024	83,5
feb.22	14970	27542	12572	84,0	26704	11734	78,4
mar.22	31646	32368	722	2,3	31400	-246	-0,8
apr.22	36900	33167	-3733	-10,1	32294	-4606	-12,5
Totalt 2022	97910	120321	22411	22,9	116816	18906	19,3
Totalt alle år	1455818	1345829	-109989	-7,6	1309227	-146591	-10,1

KAPITTEL

5

DISKUSJON

I kapittelet *Diskusjon* drøftes svarene i *Resultater*. Modellene prestasjon evalueres og forslag til forbedringer presenteres. Korrelasjonen i datasettet kommenteres også.

5.1 Modellvurdering

Prediksjoner fra modellene og beregningen av differansene knyttet til antall syklister i fokusområde 2, viste at det lett kan tas feilslutninger på bakgrunn av dataene fra Strava. Området som ble undersøkt hadde generelt lav tilstedeværelse av syklister med Strava. Dette førte til at utvalget som skulle representere populasjonen av syklister ble for lite til å kunne «stå imot» enkeltbrukeres endrede rutevaner. En mulig forklaring kan være at flere syklister kan ha benyttet Åkebergveien som pendlerrute til og fra jobb i perioden før utbyggingen, men ikke vendt tilbake til veien etter at de nye sykkelfeltene var oppført. Dette kan skyldes at gruppen syklister med Strava fant alternative ruter under byggeperioden som de fortsatte å bruke ut i periode 2. Ved å studere gater med en større andel av syklister som benytter seg av Strava, kan nevnte usikkerheter bli betraktelig redusert. Der linjesegment 105 i Åkebergveien hadde 2720 registrerte turer i periode 2, hadde linjesegmentet ved sykkeltelleren på Frognerstranda i samme tidsrom registrert 64 060 passerende syklister som logget turen sin på Strava. I et slikt områder hvor deltagelsen av syklister med og uten Strava er høyt vil et periodisk fravær av individuelle brukere av Strava ha en mindre påvirkning på beregnet avvik.

Begge modellene predikerer i snitt lavere verdier for alle de tre gatene sett i lyst av tall fra sykkeltellerne. I vintermånedene predikerer modellene ofte over sykkeltellernes verdier, og motsatt i sommermånedene. Dette skyldes generelt lite variasjon blant verdiene predikert av modellene. Siden modellene er lineære vil ikke alle variasjonene i forholdet mellom syklister med og uten Strava fanges opp. Avvikene kan derfor bli spesielt store for måneder med et høyt eller lavt volum av totalt antall syklister. Ved sammenligning av år der alle dager i alle måneder er predikert, blir avviket i flere tilfeller ikke like høyt som ved månedssammenstilling. Modellen kan derfor være bedre egnet til å se på trenden for endringer i antall syklister for gater med linjesegmenter som har data fra alle måneder over flere år. Et eksempel på dette vises i tabell 4.5 fra Kierschows gate hvor trenden i års-verdiene fra sykkeltelleren gjenspeiles i modellenes tilsvarende beregninger for årene 2018, 2019 og 2020.

OLS-modellen har lavere snittavvik på års- og totalstatistikk enn det SQRT-modellen har. SQRT-modellen leverer derimot prediksjoner med lavere avvik enn OLS-modellen i enkelte vintermåneder. Det er større variasjon i OLS-modellens prediksjoner, noe som delvis kan forklares med at modellen har en høyere R_2-verdi enn det SQRT-modellen har. Det skal nevnes at ulikhetene mellom de predikerte verdiene fra de to modellene utgjør en liten andel sammenlignet med avviket til sykkeltellernes verdier.

5.2 Forslag til forbedring av modell

For å oppnå mer nøyaktige prediksjoner må modellene som er estimert i denne studien forbedres, og nye modelltyper må testes. Videre må modellenes RMSE reduseres og R_2-verdi økes for for å oppnå bedre prediksjoner. På den måten kan usikkerheten til hver ny prediksjon av $\hat{y}$ reduseres og resultatet i det lange løp bli mer korrekt. Siden alle β-parametere i begge modellene viste høy nytte i forbindelse med estimeringen av y, kan en utvidelse ved å legge til flere uavhengige variabler bidra til å gjøre modellene bedre. Livingston mfl. [36] anvendte modeller med kvalitative variabler for geografisk tilhørighet, år og tidspunkt på dagen (formiddag/ettermiddag/utenom rush). I tillegg ble interaksjonsledd mellom antall aktiviteter i Strava og ulike variabler i de tre nevnte kategoriene testet ut for å redusere RMSE og øke verdien av R_a^2. Lin og Fan [55] benyttet kvalitative variabler for å skille mellom ulike typer sykkelveier som var tatt i bruk under sykkelaktiviteter i Strava. Variabler som tok høyde for ulikheter i demografiske forhold ble også undersøkt i studien. Dette er variabler som også kan vurderes i en modell for prediksjon av sykkeltrafikken i Oslo.

Antall syklister som benytter ulike veier og gater i Oslo varierer. Differansen i volumet av syklister mellom Åkebergveien og Frognerstranda er ett eksempel som er nevnt i studien. Det samme gjelder for antall brukere av Strava for ulike veier. En forbedret modell kunne tatt høyde for denne ulikheten ved å kategorisere veier etter popularitet blant syklister. Et varmekart, tilsvarende *Heatmap* i Metroview, kunne vært klassifisert etter popularitet blant alle syklister og ved estimering av modellen hadde linjesegmentene i datasettet vært knyttet til informasjon om gatens popularitet. Modellen kunne da gjort en mer "skreddersydd" prediksjon for enkeltgater sammenlignet med å benytte en generalisert kvalitativ variabel som representerer en hel bydel.

Alle de tre gatene som er undersøkt i studien ligger i bydeler med lav tetthet av syklister som benytter Strava sett i forhold til antallet syklister totalt. Med utgangspunkt i figur 3.11 hadde det vært interessant å teste modellen i områdene mot ytterkantene av den befolkede delen av Oslo hvor tettheten av Strava-syklister er høyere. Siden påslag fra kvalitative variabler ikke er gjeldende for enkelte av de nevnte områdene er det likevel ikke sikkert modellene hadde predikert med mindre avvik. Det hadde likevel vært interessant å studere bidraget fra de kvalitative variablene opp mot bidraget fra den kvantitative variabelen βx_1.

For videre arbeid hadde det vært interessant å teste andre typer regresjonsmodeller, som for eksempel generaliserte og vektede modeller. Av generaliserte modeller er Poisson regresjon tidligere blitt benyttet i flere studier for prediksjon av syklister basert på data fra Strava. Modeller som korrigerer for romlig autokorrelasjon har også tidligere vært benyttet i denne sammenhengen. Slike modeller kunne vært interessant å teste i et eventuelt videre arbeid med oppgaven.

5.3 Korrelasjonsverdier

Korrelasjonen mellom data fra linjesegmentene i Strava og sykkeltellere går fram i tabell 4.1. Korrelasjonsverdiene varierer mellom de ulike tellestasjonene og på tvers av mai måned hvert år. Felles for mange av sykkeltellerne med høy og stabil korrelasjon over flere av studieårene, er at de har registreringspunkt som er plassert på en adskilt sykkelvei og ikke i blandet trafikk som eksempelvis i veibanen eller på sykkelfelt i veiskulderen. Det finnes imidlertid unntak som for "Maridalsveien 323" og "Monolittveien" hvor registreringssensoren er plassert i sykkelfelt i veibanen og korrelasjonsverdiene relativt sett er stabilt høye over alle fire år. Felles for to av de tre stasjonene ("Bærumsveien 22 Vest" og "Hoffveien 40 Sør") med gjennomsnittlig korrelasjonsverdi på under 0,5, er at begge ligger i veibaner. På spørsmål om sykkeltellere med registreringspunkt i veibanen er mer utsatt for feilregistrering som følge av blandet trafikk og større slitasje kunne overingeniør Per Roy Laudal i avdeling for trafikkstyring og analyse i Bymiljøetaten svare følgende:

"Vi har ikke erfart at tellere med blandet trafikk er mer upresise enn andre tellere. Det kan forekomme at MC blir registrert som sykkel [...] Slitte tellere kan gi mer feil, men siden vi henter ut data månedlig klarer vi å holde oversikt over hvilke sløyfer som må erstattes." (P.R.Laudal, personlig kommunikasjon, 29.04.2022)

Dette tilbakeviser mistanken om at registreringspunktene i blandet trafikk registrerer mindre presist enn adskilte gang- og sykkelveier. Utover dette er det få åpenlyse teorier som kan forklare at korrelasjonsverdiene varierer slik de gjør i denne oppgaven.

Under 3.2.2 ble sykkeltellerne gradert **A** eller **B** basert på en vurdering av hvor sikkert valget av korrekt linjesegment var. I tabell 4.1 er alle sykkeltellerne oppnevnt med gradering. Det finnes sykkeltellere med høy og lav korrelasjon i begge graderingsklassene. Bortsett fra "Breivoll Sykkel" er alle sykkeltellere med en gjennomsnittlig korrelasjon på over 0,9 tilknyttet gruppe A. Sett bort ifra sykkeltellerpunktene i Hoffsveien er sykkeltellere med årlig korrelasjon på under 0,5 kun tilhørende gruppe B. Utover dette er det lite som markant skiller de to gruppene fra hverandre. →

Det er derfor lite som tyder på at tettliggende linjesegmenter og utfordrende omgivelser med tanke på nøyaktighet knyttet til posisjonsbestemmelse, som kan føre til ukorrekt tildeling av aktiviteter, har en vesentlig betydning for korrelasjonen mellom Strava og sykkeltellere på valgt aggregeringsnivå.

For videre arbeid vil aggregering på flere nivåer være viktig for å forstå mer av årsaken bak variasjonene i korrelasjonsverdiene. Sammenligning av beregninger fra flere måneder enn bare mai kan bidra til å forstå hvilke stasjoner og linjesegmenter som jevnt over er stabilt høyt korrelerte og motsatt. Ved å anvende resultater fra korrelasjonsberegninger mer aktivt kan datasett bestå av data som gir bedre forutsetninger for å estimere en forbedret modell.

KAPITTEL

6

KONKLUSJON

I kapittelet *Konklusjon* presenteres slutningene som er gjort for oppgavens innledende forskningspørsmål med bakgrunn i resultatene.

Under introduksjonen i denne masteroppgaven ble følgende to forskningsspørsmål fremlagt:

1. Er det samvariasjon mellom antall registrerte sykkelturer i Strava og antall syklister totalt for et område?
2. Er det mulig å anvende data fra Strava til å predikere det totale antall syklister som tar i bruk nye sykkelveier i Oslo?

Resultater fra korrelasjonsberegningene viser at data fra Strava samvarierer positivt med det totale antall syklister i mange tilfeller. En gjennomsnittlig korrelasjonsverdi på 0,8 beregnet ut ifra data fra mai måned i 4 år, fra 37 ulike sykkeltellere viser at antall sykkelturer som blir registrert i Strava har en signifikant sammenheng med det totale volumet av syklister. Dette er tilfelle over tid og for ulike geografisk områder. Resultatene viser også at korrelasjonen kan forandres over betydelig over tid for enkeltstasjoner og det er derfor ingen garanti for at sykkeltellere som tidligere har hatt sterk positiv korrelasjon med data fra Strava vil ha dette i fremtiden. På bakgrunn av dette må det utvises aktsomhet hvis man skal gjøre antagelser om fremtidige korrelasjonsverdier.

Begge modellene som ble estimert og benyttet til prediksjon av det totale volumet syklister i denne oppgaven viste varierende nøyaktighet. Ved evaluering av utviklingen for sykkeltrafikken i Åkebergveien før og etter innføringen av opphøyde sykkelfelt presenterte begge modellene et lite korrekt bilde av virkeligheten. Dette resultatet påpekte viktigheten av å utforske dataene fra Strava nøye på forhånd og vurdere disse med et kritisk blikk før en prediksjon. For Thorvald Meyers gate og Kierschows gate predikerte modellene i enkelte perioder verdier med mindre avvik, men disse var likevel ikke gode nok til å studere detaljene i den totale sykkeltrafikken for ulike perioder.

Modellene som er brukt i denne oppgaven kan benyttes til å studere variasjoner og trender tilknyttet rutevalg i områder der volumet av syklister er relativt høyt. Dette forutsetter imidlertid en kontroll mot komplette år (alle måneder) med sykkeltellerdata for å stadfeste modellens gjennomsnittlige avvik.

Skal modellene benyttes til sitt tiltenkte formål, som er å predikere det totale antall syklister for flere perioder, for så å sammenligne periodene og studere utviklingen i sykkeltrafikken må modellene videreutvikles og testes ytterligere. Selv ved en optimalisering med utgangspunkt i oppgavens modeller, eller estimering av andre typer modeller så er ikke dataene fra Strava per nå en sikker nok representasjon av befolkningen i Oslo til at modellene kan benyttes etter intensjonen i forskningsspørsmål 2. →

I denne intensjonen inngår det at modellen skal kunne benyttes med sikkerhet uten kontroll mot sykkeltellere, kameraregistrering eller manuelle tellinger på daglig basis. Resultatene i oppgaven viser likevel at det ligger mye potensiale i data fra Strava som det er mulig å utnytte på ulike måter. På sikt, med en økning i antall brukere og en bredere demografisk fordeling som gir en bedre representasjon av befolkningen vil forutsetningene for å utvikle en modell som kan predikere volumet av sykkeltrafikken nøyaktig nok til å erstatte faste sykkeltellere, være til stede.

BIBLIOGRAFI

[1] Statens Vegvesen og Oslo Kommune. «Plan for sykkelveinettet i Oslo». I: *Sykkelstrategier og dokumenter - 1: Planer og strategier* (2020). URL: https://www.oslo.kommune.no/gate-transport-og-parkering/sykkel/sykkelstrategier-og-dokumenter/.

[2] Ida Laingen. *Sykkelfeltene i Bygdøy allé skaper kø og kaos*. URL: https://norgesnytt.net/sykkelfeltene-i-bygdoy-all-skaper-ko-og-kaos/114653. (Versjon: 08.09.2021).

[3] Oslo kommune. *Oslo-budsjettet 2022: Et budsjett for arbeid, sosial utjevning og klima*. URL: https://www.oslo.kommune.no/politikk/byradet/for-pressen/pressemeldinger-fra-byradet/oslo-budsjettet-2022-et-budsjett-for-arbeid-sosial-utjevning-og-klima#gref.

[4] Mathilde Walberg Hofseth. *Koker i kommentarfelt etter ny sykkelvei i Gyldenløves gate*. URL: https://norgesnytt.net/adam-tumidajewicz-beboerparkering-debatt/koker-i-kommentarfelt-etter-ny-sykkelvei-i-gyldenloves-gate/118516. (Versjon: 20.05.2021).

[5] Statens Vegvesen og Oslo Kommune. «Sykkelregnskapet for Oslo 2013-2017». I: *Sykkelstrategier og dokumenter - 4: Statistikk, analyse og spørreundersøkelser* (2020). URL: https://www.oslo.kommune.no/gate-transport-og-parkering/sykkel/sykkelstrategier-og-dokumenter/.

[6] David N Barton, Vegard Gundersen og Zander V Venter. «Bruk av stordata i arbeidet med å tilrettelegge for fysisk aktivitet. Kunnskapsstatus og forslag til anvendelse i Norge». I: (2021).

[7] Darren Boss mfl. «Using crowdsourced data to monitor change in spatial patterns of bicycle ridership». I: *Journal of Transport & Health* 9 (2018), s. 226–233.

[8] Avipsa Roy mfl. «Correcting bias in crowdsourced data to map bicycle ridership of all bicyclists». I: *Urban Science* 3.2 (2019), s. 62.

[9] Jinhyun Hong, David Philip McArthur og Mark Livingston. «The evaluation of large cycling infrastructure investments in Glasgow using crowdsourced cycle data». I: *Transportation* 47.6 (2020), s. 2859–2872.

[10] Kristiann C Heesch og Michael Langdon. «The usefulness of GPS bicycle tracking data for evaluating the impact of infrastructure change on cycling behaviour». I: *Health promotion journal of Australia* 27.3 (2016), s. 222–229.

[11] Esri. *point-in-polygon overlay*. URL: `https://support.esri.com/en/other-resources/gis-dictionary/term/1b1219bf-eae0-499d-9660-83914fb1c545`. (Hentet: 03.06.2022).

[12] Esri. *Data classification methods: Natural breaks (Jenks)*. URL: `https://pro.arcgis.com/en/pro-app/latest/help/mapping/layer-properties/data-classification-methods.htm`.

[13] de Smith, Goodchild og Longley. *Geospatial Analysis—A Comprehensive Guide, 6th edition*. 2018. URL: `https://www.spatialanalysisonline.com/HTML/index.html`. (Versjon: 2021).

[14] Jan Ketil Rød. *Innføring i GIS og statistikk - verktøy for å beskrive verden*. Fagbokforlaget, 2017.

[15] *gjennomsnitt i Store norske leksikon*. URL: `https://snl.no/gjennomsnitt`. (Versjon: 28. mai 2018).

[16] William Mendenhall og Terry Sincich. *A second course in statistics: regression analysis*. Pearson Education Limited, 2013.

[17] Geir Sverre Braut og Sirianne Dahlum. *regresjonsanalyse i Store norske leksikon*. URL: `https://snl.no/regresjonsanalyse`. (Versjon: 22.12.2021).

[18] Kyuhyun Lee og Ipek Nese Sener. «Strava Metro data for bicycle monitoring: a literature review». I: *Transport reviews* 41.1 (2021), s. 27–47.

[19] *Strava Metro Application*. URL: `https://metroview.strava.com/application`. (Lest: 14.01.2022).

[20] Carlos Gamez. *Strava Metro - Dashboard Boundary*. URL: `https://stravametro.zendesk.com/hc/en-us/articles/360051964873-Dashboard-Boundary`. (Sist oppdatert: januar 2022).

[21] Jonathan Bennett. *OpenStreetMap*. Packt Publishing Ltd, 2010.

[22] *Strava Metro Frequently Asked Questions. utg. 23.09.2020.* URL: https://metro.strava.com/faq. (Lest: 01.02.2022).

[23] *Strava Metro Metroview.* URL: https://metroview.strava.com/map/oslo-norway/ride.

[24] Carlos Gamez. *Strava Metro - Data Export and Download.* URL: https://stravametro.zendesk.com/hc/en-us/articles/360051202734-Data-Export-and-Download. (Lest: 01.02.2022).

[25] Anne Eilertsen. *linjestykke i Store norske leksikon.* URL: https://snl.no/linjestykke. (Hentet: 10.mai 2022).

[26] Erik Sunde. *Strava Metro - Glossary & Data Dictionary.* URL: https://stravametro.zendesk.com/hc/en-us/articles/1500001573281-Glossary-Data-Dictionary. (Lest: 04.02.2022).

[27] *Statens vegvesen - Om trafikkdata.* URL: https://www.vegvesen.no/trafikkdata/start/om-trafikkdata. (Sist oppdatert: 24.02.2021).

[28] Terje Giæver. «Utprøving av utstyr for å registrere sykkeltrafikk». I: (2009).

[29] Statens vegvesen. «Veileder i trafikkdata». I: *Nr. V714 i Statens vegvesens håndbokserie* (2014).

[30] Sissel Fantoft. *Oslofolk er verdensmestre i sykkelbruk.* URL: https://www.klimaoslo.no/2017/09/21/oslofolk-er-verdensmestre-i-sykkelbruk/. (Publisert: 17.09.2017).

[31] *Google Maps.* URL: https://www.google.com/maps.

[32] Aanderaa Data Instruments AS. *Inductive Loops - Technical note.* URL: https://www.mercell.com/m/file/GetFile.ashx?id=148928524&version=0. (Publisert: 28.01.2020).

[33] Statens vegvesen. *Trafikkdata.* URL: https://www.vegvesen.no/trafikkdata.

[34] Eirik Rossen. *API i Store norske leksikon.* URL: https://snl.no/API. (Sist oppdatert: 31.06.2020).

[35] *Sykkeltellere i Oslo kommune.* URL: https://data.eco-counter.com/ParcPublic/?id=3936.

[36] Mark Livingston mfl. «Predicting cycling volumes using crowdsourced activity data». I: *Environment and Planning B: Urban Analytics and City Science* 48.5 (2021), s. 1228–1244.

[37] Knut Are Tvedt. *Oslo - byutvikling og areal i Store norske leksikon.* URL: https://snl.no/Oslo_-_byutvikling_og_areal. (Sist oppdatert: 07.02.2022).

[38] Statistisk sentralbyrå. *Statistikk - Befolkning*. URL: https://www.ssb.no/befolkning/folketall.

[39] Meteorlogisk institutt. *Oslo (Blindern) - Historikk*. URL: https://www.yr.no/nb/historikk/graf/5-18700/Norge/Oslo/Oslo/Oslo.

[40] Opinion AS. «Holdningsundersøkelse om sykling i Oslo». I: *Sykkelstrategier og dokumenter - 4: Statistikk, analyse og spørreundersøkelser* (2020). URL: https://www.oslo.kommune.no/gate-transport-og-parkering/sykkel/sykkelstrategier-og-dokumenter/.

[41] Esri mfl. *Human Geography Map*. URL: https://www.arcgis.com/home/item.html?id=3582b744bba84668b52a16b0b6942544. (Versjon: 28. april 2022).

[42] Norkart. *Kommunekart*. URL: https://kommunekart.com.

[43] Aslak Fyhri mfl. *Cycle pilot study: Effect of uni-directional cycleway in Oslo*. Tekn. rapp. 2020.

[44] Bymiljøetaten. «Oslostandarden for sykkeltilrettelegging». I: *Sykkelstrategier og dokumenter - 2: Veiledere* (2020). URL: https://www.oslo.kommune.no/gate-transport-og-parkering/sykkel/sykkelstrategier-og-dokumenter/.

[45] Oslo Kommune og Spacescape. «Oslo sykkelstrategi 2015-2025». I: *Sykkelstrategier og dokumenter - 1: Planer og strategier* (2020). URL: https://www.oslo.kommune.no/gate-transport-og-parkering/sykkel/sykkelstrategier-og-dokumenter/.

[46] Shawn Turner og Philip Lasley. «Quality Counts for Pedestrians and Bicyclists: Quality Assurance Procedures for Nonmotorized Traffic Count Data». I: *Transportation research record* 2339.1 (2013), s. 57–67.

[47] Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. "O'Reilly Media, Inc.", 2012.

[48] NumPy Developers. *NumPy documentation*. URL: https://numpy.org/doc/stable/.

[49] John Hunter mfl. *matplotlib.pyplot documentation*. URL: https://matplotlib.org/3.5.0/api/_as_gen/matplotlib.pyplot.html.

[50] Esri. *Summary Statistics (Analysis)*. URL: https://pro.arcgis.com/en/pro-app/latest/tool-reference/analysis/summary-statistics.htm.

[51] The pandas development team. *pandas.Dataframe.corr*. URL: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html. (Versjon: 1.4.2).

[52] Rune Brekke - Geodata. *Kartlag: Bydeler Oslo*. URL: https://www.arcgis.com/home/item.html?id=931cf5beacd04418a74b2005ee85df4e#. (Versjon: 19.10.2021).

[53] Esri. *Join Field (Data Mangement)*. URL: https://pro.arcgis.com/en/pro-app/2.8/tool-reference/data-management/join-field.htm.

[54] Ray Pritchard, Dominik Bucher og Yngve Frøyen. «Does new bicycle infrastructure result in new or rerouted bicyclists? A longitudinal GPS study in Oslo». I: *Journal of transport geography* 77 (2019), s. 113–125.

[55] Zijing Lin og Wei David Fan. «Modeling bicycle volume using crowdsourced data from Strava smartphone application». I: *International journal of transportation science and technology* 9.4 (2020), s. 334–343.

TILLEGG

A

RESULTATER - EKSTRA

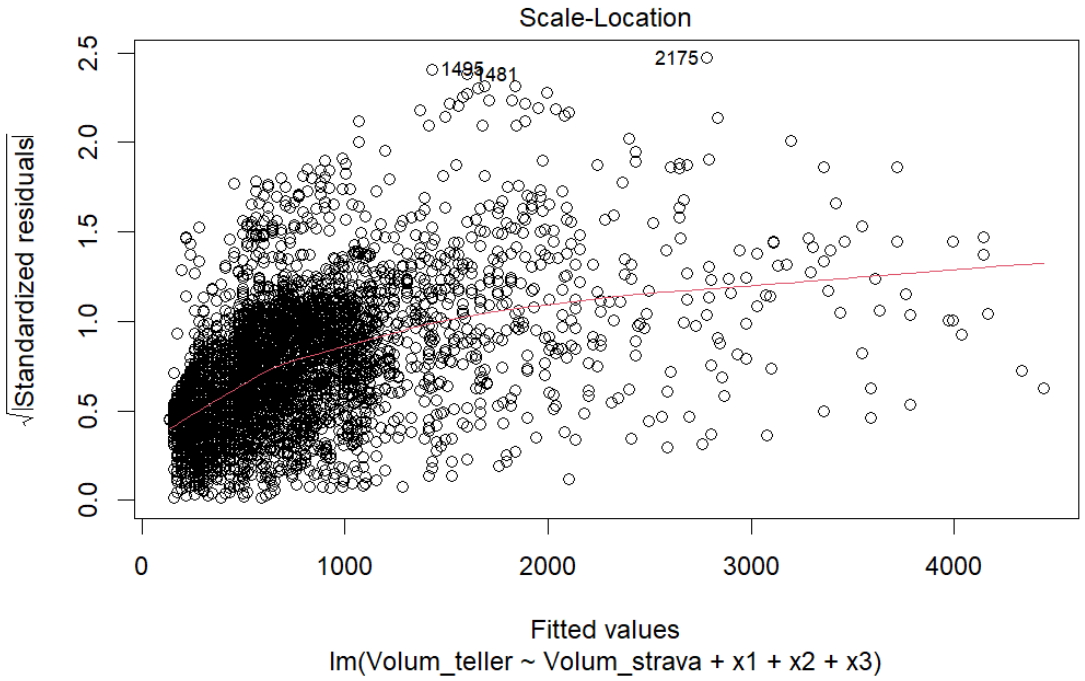
A.1 Kvalitative variabler per sykkelteller

Tabell A.1: Verdier for kvalitative variabler per sykkelteller basert på bydelstilhørighet.

Sykkelteller	x1	x2	x3
Lysaker Sykkel	0	0	0
Kongsveien 48	0	0	0
Nordstrandveien 59	0	0	0
Bryn Sykkel	0	0	0
AKERSYKE. SYKKEL 03	0	0	0
Maridalsveien 323	0	0	0
Tåsen	0	0	0
Tåsenveien 43	0	0	0
Monolittveien	0	0	0
Hoffsveien 40 Sør	0	0	0
Hoffsveien 40 Nord	0	0	0
Bærumsveien 22 Øst	0	0	0
Bærumsveien 22 Vest	0	0	0
Smestad Sykkel	0	0	0
Gaustad sykkel	0	0	0
Holmenkollveien 42	0	0	0
Jon Smestads vei 4	0	0	0
Veitvet	0	0	0
Mosseveien Oslo	0	0	0
Grorud Sykkel	0	0	0
Breivoll Sykkel	0	0	0
Østensjøvannet	0	0	0
Trasop skole	0	0	0
Bygdøy alle 13	1	0	0
FROGNERST.SYKKEL 03	1	0	0
Munkedamsveien Sykkel	1	0	0
Grenseveien GS	1	0	0
Ring 2 Majorstua Nord	1	0	0
Ring 2 Kirkevn ved Sigyns gt. Nordgående	1	0	0
Ring 2 Kirkevn ved Sigyns gt. Sydgående	1	0	0
Dr. Eufemias Gt. Sykkel East	1	0	0
Dr. Eufemias gt. Sykkel West	1	0	0
Helsfyr	1	0	0
Chr Michelsens gt sydside	0	1	0
Chr Michelsens gt nordside	0	1	0
Griffenfeldts gate 19 Vest	0	1	0
Bislettgate 1	0	0	1

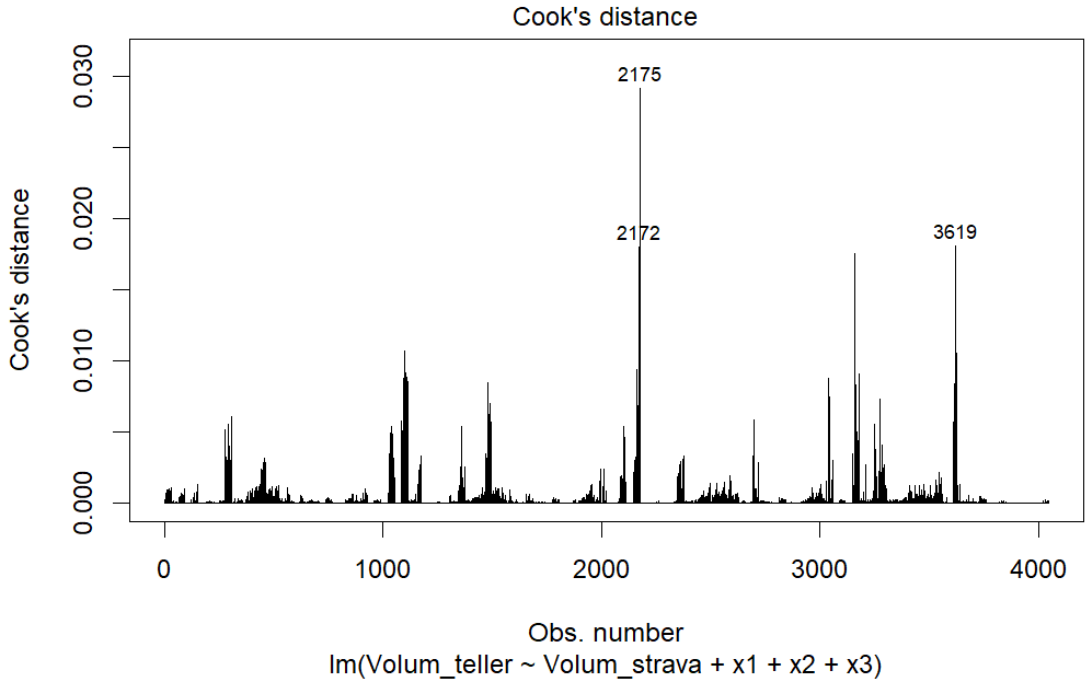
A.2 OLS-modell

A.2.1 Scale-Location plott



Figur A.1: "Scale-Location" plott viser stigende varians for ε når $\hat{y}$ øker.

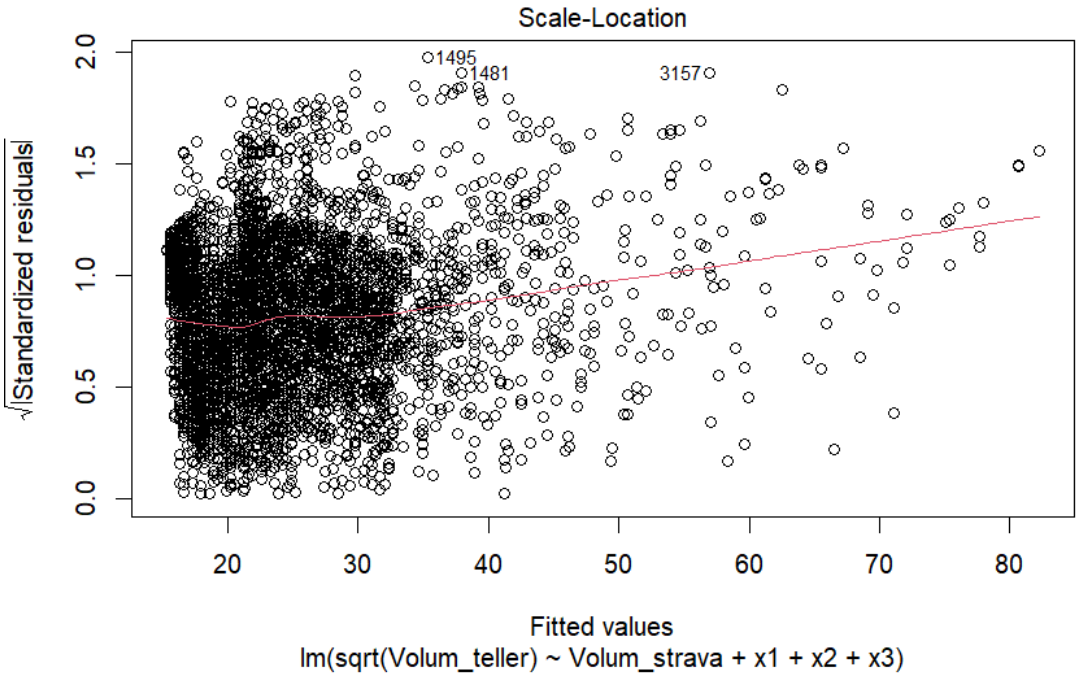
A.2.2 Cook's distanse



Figur A.2: Cook's distanse beregnet for alle observasjoner i OLS-modellen.

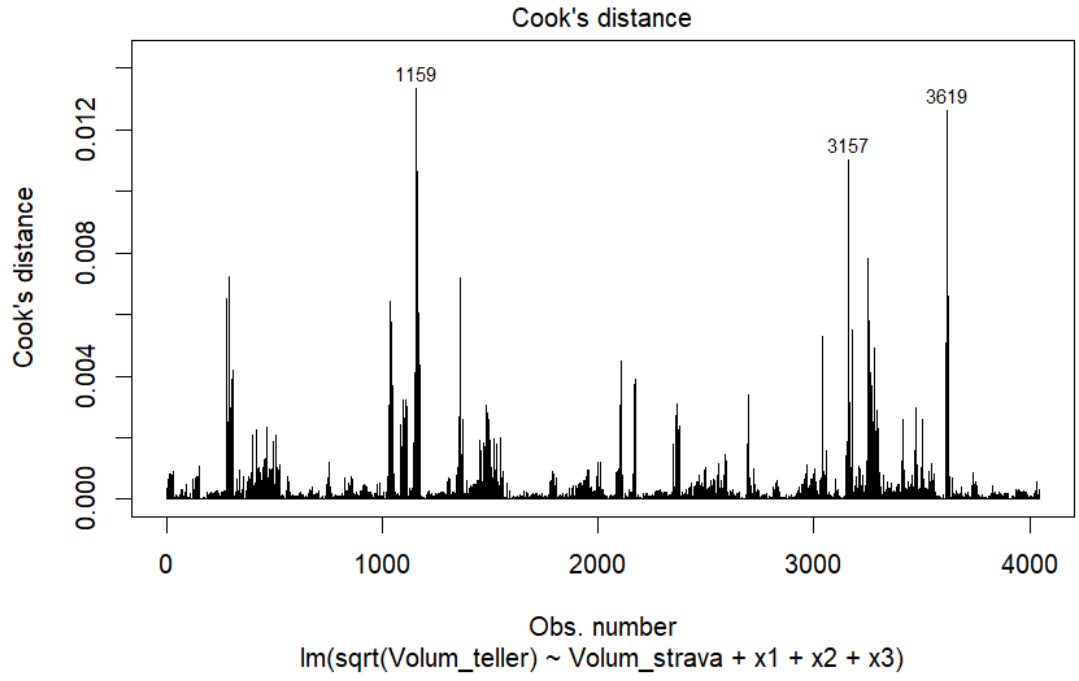
A.3 Kvadratrot-transformert modell

A.3.1 Scale-Location plott



Figur A.3: "Scale-Location" plott viser stigende varians for ε når $\sqrt{\hat{y}}$ øker.

A.3.2 Cook's distanse



Figur A.4: Cook's distanse beregnet for alle observasjoner i SQRT-modellen.

TILLEGG

B

PYTHONKODE

Data proccecing

Importing modules

In [7]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import itertools
import csv

import seaborn as sns
sns.set()
from sklearn.linear_model import LinearRegression
```

Checking the quality in dataset from Statens vegvesen

In [8]:
```
def controll_and_reformat_sv_csv(teller_df):
    """
    Used in "df_pivottable_from_csv"
    """
    # Checks that "Dekningsgraden" is 95 percent or highter for all bicycle counters.
    teller_df = teller_df[teller_df['Dekningsgrad (%)'] >= 95.0]

    teller_df = teller_df[teller_df['Antall timer ugyldig'] == 0]

    numbers = ['1', '2', '5']
    teller_df = teller_df[teller_df['Felt'].isin(numbers) == False]

    teller_df = teller_df.loc[teller_df['Felt'] == 'Totalt']
    teller_df = teller_df.loc[:,['Trafikkregistreringspunkt', 'Navn', 'Dato',
                                 'Volum']]
    teller_df.rename(columns = {'Volum':'Volum teller'}, inplace = True)

    return teller_df
```

Function called to determine which directions to include from the Strava dataset in a counting point

In [9]:
```
def define_edges_strava(strava_df, merger_codes):
```

```
    """
    Used in "df_pivottable_from_csv"
    """

    strava_df['Volum strava'] = np.nan
    for index, row in merger_codes.iterrows():
        num_id = row['edge_uid']
        index_values_strava = strava_df.index[strava_df['edge_uid'] == num_id].tolist()

        if row['Retning'] == 'Begge':
            volum_values = strava_df.loc[index_values_strava[0]:index_values_strava[-1],['forward_trip_count',
                                                                                          'reverse_trip_count']].sum(axis=1)
            strava_df.loc[strava_df['edge_uid'] == num_id, 'Volum strava'] = volum_values

        elif row['Retning'] == 'Positiv':
            volum_values = strava_df.loc[index_values_strava[0]:index_values_strava[-1],'forward_trip_count']
            strava_df.loc[strava_df['edge_uid'] == num_id, 'Volum strava'] = volum_values

        elif row['Retning'] == 'Negativ':
            volum_values = strava_df.loc[index_values_strava[0]:index_values_strava[-1],'reverse_trip_count']
            strava_df.loc[strava_df['edge_uid'] == num_id, 'Volum strava'] = volum_values
        else:
            raise ValueError('"Tellerkoder.csv" contains strings in column "Retning" which is not correct')

    return strava_df
```

This function returns a data frame with information about the proportion of men and women, as well as different age groups in a Strava data set.

In [10]:

```
def get_demographic_data(strava_csv, merger_codes=None, column_title=None):

    strava_df = pd.read_csv(strava_csv, sep=';')
    if merger_codes is not None:
        ids = merger_codes['edge_uid'].unique().tolist()
        strava_df = strava_df[strava_df['edge_uid'].isin(ids)]

    strava_df = strava_df.loc[:,['forward_male_people_count', 'reverse_male_people_count',
                                 'forward_female_people_count', 'reverse_female_people_count',
                                 'forward_13_19_people_count', 'reverse_13_19_people_count',
                                 'forward_20_34_people_count', 'reverse_20_34_people_count',
                                 'forward_35_54_people_count', 'reverse_35_54_people_count',
```

```
                                'forward_55_64_people_count', 'reverse_55_64_people_count',
                                'forward_65_plus_people_count', 'reverse_65_plus_people_count']]

    totals_strava_df = strava_df.groupby((np.arange(len(strava_df.columns)) // 2) + 1, axis=1).sum()
    totals_strava_df.rename(columns = { 1:'Menn', 2: 'Kvinner', 3: '13-19 år', 4: '20-34 år', 5: '35-54 år',
                                    6: '55-64 år', 7: '65+ år'}, inplace = True)

    sum_totals_strava_df = totals_strava_df.sum().to_frame(name=column_title)

    return sum_totals_strava_df
```

The function returns a pie chart of annual distributions between men and women based on the input dataframe.

In [11]:

```
def get_gender_pie_chart(df):
    """
    Input "df" must be a dataframe returned from "get_demographic_data"-function.
    """
    x2021all = df.iloc[4,0:2]
    x2021 = df.iloc[3,0:2]
    x2020 = df.iloc[2,0:2]
    x2019 = df.iloc[1,0:2]
    x2018 = df.iloc[0,0:2]
    cmap = plt.get_cmap("tab20")
    colors1 = cmap(np.array([19, 7]))
    colors2 = cmap(np.array([18, 6]))


    plt.figure(figsize=(13,10))
    plt.pie(x2021all, labels=['Menn', 'Kvinner'],
            startangle=90, colors=colors1, pctdistance=0.91, autopct= '%1.1f%%, \n (Oslo 2021)', radius=1.25,
            labeldistance=1.1, textprops={'fontweight':'bold', 'fontsize':15},
            wedgeprops = {'linewidth': 5, 'edgecolor': 'white'})

    plt.pie(x2021, startangle=90, colors=colors2, pctdistance=0.90, autopct= '%1.1f%%, \n (2021)', radius=1.0,
            labeldistance=1.1, textprops={'fontweight':'bold', 'fontsize':15},
            wedgeprops = {'linewidth': 5, 'edgecolor': 'white'})

    plt.pie(x2020, startangle=90, colors=colors2, pctdistance=0.87, autopct= '%1.1f%%, \n (2020)', radius=0.80,
            textprops={'fontweight':'bold', 'fontsize':15},
            wedgeprops = {'linewidth': 5, 'edgecolor': 'white'})

    plt.pie(x2019, startangle=90, colors=colors2, pctdistance=0.82, autopct= '%1.1f%%, \n (2019)', radius=0.60,
```

```
            textprops={'fontweight':'bold', 'fontsize':15},
            wedgeprops = {'linewidth': 5, 'edgecolor': 'white'})

    plt.pie(x2018, startangle=90, colors=colors2, pctdistance=0.74, autopct= '%1.1f%%, \n (2018)', radius=0.40,
            textprops={'fontweight':'bold', 'fontsize':15},
            wedgeprops = {'linewidth': 5, 'edgecolor': 'white'})

    centre_circle = plt.Circle((0,0),0.20,fc='white')
    plt.title('Strava - andel menn og kvinner', fontweight='bold', fontsize=30, pad=50)
    plt.tight_layout()
    #plt.savefig('Matplotlib_figures/Strava_andel_menn_kvinner.png')
    plt.show()
```

The function returns a bar plot with the age distribution based on the input dataframe.

In [6]:

```
def get_age_distribution_plot(df, stacked_or_not=True, title='Title'):
    """
    Input "df" must be a dataframe returned from "get_demographic_data"-function.
    """
    df_to_plot = df[['13-19 år', '20-34 år', '35-54 år', '55-64 år', '65+ år']]

    df_percentage = df_to_plot.div(df_to_plot.sum(axis=1), axis=0)
    df_percentage.reset_index(inplace=True)
    df_percentage.rename(columns = {'index':'År'}, inplace = True)

    print(df_percentage.to_string())
    if df_percentage['År'].str.contains('Hele Oslo 2021').any():
        df_percentage = df_percentage.drop([4])

    df_percentage.plot(x='År',
                       figsize=(9,10),
                       kind='bar',
                       edgecolor='black',
                       linewidth=2,
                       width=0.8,
                       stacked=stacked_or_not,
                       rot=0,
                       fontsize=18,
                       colormap=plt.get_cmap('YlOrRd'))
    plt.title(title, fontweight='bold', fontsize = 28, pad=40)
    plt.legend(fontsize=15)
```

```
    plt.ylabel('Andel i prosent', fontsize=22)
    plt.xlabel('År', fontsize=22)
    # plt.savefig('Matplotlib_figures/Aldersfordeling_Strava_versjon_3.png')
    plt.show()
```

Function that make merge countingdata and Strava data, and make a pivot table.

In [12]:

```
def df_pivottable_from_csv(num_dates, strava_csv, merger_codes_csv, sv_csv=None, ok_csv=None):

    # Reading CSV-files to pandas.DataFrame
    strava_df = pd.read_csv(strava_csv, sep=';')
    merger_codes = pd.read_csv(merger_codes_csv, sep=';')
    if sv_csv and not ok_csv:
        teller_df = pd.read_csv(sv_csv, sep=';')
    elif ok_csv and not sv_csv:
        teller_df = pd.read_csv(ok_csv, sep=';')
    else:
        raise ValueError('Choose variable input "sv_csv" or "bym_csv", not both')

    # Removing unnecessary columns
    strava_df = strava_df.loc[:,['edge_uid', 'date', 'forward_trip_count',
                                 'reverse_trip_count']]
    if sv_csv:
        teller_df = teller_df.loc[:,['Trafikkregistreringspunkt', 'Navn', 'Dato', 'Felt',
                                 'Volum', 'Dekningsgrad (%)', 'Antall timer total',
                                 'Antall timer inkludert', 'Antall timer ugyldig']]
    else:
        teller_df = teller_df.loc[:,['Navn', 'Dato', 'Totalt']]
        teller_df.rename(columns = {'Totalt':'Volum teller'}, inplace = True)


    # Quality control of data from Statens vegvesen
    if sv_csv:
        teller_df = controll_and_reformat_sv_csv(teller_df)

    # Quality controll of data from Strava
    num_uid = strava_df['edge_uid'].value_counts()
    if (num_uid > num_dates).any():
        raise ValueError('One or more "edge_uid" are duplicated in Strava CSV-file')

    # Sum up the counter values in the two directions in Strava CSV
    strava_df = define_edges_strava(strava_df, merger_codes)
```

```
strava_df.rename(columns = {'date':'Dato'}, inplace = True)

# Checking for NaN-values
print(teller_df.isnull().sum())
print(strava_df.isnull().sum())

teller_df = teller_df.dropna(subset=['Volum teller'])

# Add name of each counter to corresponding edge_uid in Strava dataframe
merger_codes.index = merger_codes['Navn']
names_and_id_df = merger_codes.drop(columns=['Navn', 'Retning', 'Gradering'])
names_and_id_dict = names_and_id_df.to_dict()['edge_uid']

teller_df['edge_uid'] = teller_df['Navn'].apply(lambda x: names_and_id_dict.get(x))
teller_df['edge_uid'] = teller_df['edge_uid'].astype(np.int64)
strava_df['edge_uid'] = strava_df['edge_uid'].astype(np.int64)

# Merging dataframes from countingstation and Strava
if sv_csv == 'Statens vegvesen/statens_vegvesen_mai21.csv':
    teller_df.Dato = pd.to_datetime(teller_df.Dato, format='%Y-%m-%d').dt.strftime('%d.%m.%Y')

merged_df = pd.merge(strava_df, teller_df, on=['edge_uid', 'Dato'], how='inner')
merged_df.reset_index(drop=True, inplace=True)
df_to_pivot = merged_df.loc[:,['edge_uid', 'Navn', 'Dato',
                 'Volum teller', 'Volum strava']]

unique_dates = df_to_pivot.Dato.unique()

print('Antall datoer:', len(unique_dates))

names = df_to_pivot['Navn'].unique().tolist()

# Make pivot table of merged dataframe
pivot_df = df_to_pivot.copy()
pivot_df = pivot_df.astype({'Volum teller': 'int'})
pivot_df = pivot_df.pivot_table(values =['Volum teller', 'Volum strava', 'edge_uid'],
                                index='Navn', columns='Dato')
pivot_df = pivot_df.stack()

pivot_df = pivot_df.astype({'edge_uid':'int'})

# Remove data from rows where "Volum strava" is higher than "Volum teller", and where "Volum teller"
# have a value below 50.
pivot_df_filtered = pivot_df.copy()
```

```
    for counter, values in pivot_df.iterrows():
        if values['Volum strava'] > values['Volum teller']:
            pivot_df_filtered.loc[counter, 'Volum strava'] = np.nan
            pivot_df_filtered.loc[counter, 'Volum teller'] = np.nan
        if values['Volum teller'] < 50.0:
            pivot_df_filtered.loc[counter, 'Volum strava'] = np.nan
            pivot_df_filtered.loc[counter, 'Volum teller'] = np.nan


    return pivot_df_filtered, names
```

Finds the correlation between counter and Strava data

In [13]:

```
def correlation(pivot_df, names, plot=False):
    """
    coff_df: Gives you the correlation coefficient for all counters in pivot_df input.
    high_correlations_counters: Return a list with the cunters having data that where
                                the correlation coefficient is above 0.75.
    """
    corr_df = pd.DataFrame(index=names)
    high_correlation_counters = []
    for counter in names:
        X = pivot_df['Volum strava'].loc[counter]
        y = pivot_df['Volum teller'].loc[counter]

        correlation_value = X.corr(y, method='pearson')
        corr_df.loc[counter, 'Correlation'] = correlation_value

        if correlation_value > 0:
            high_correlation_counters.append(counter)


    if plot == True:
        sns.set(style='whitegrid', font_scale=1.0)
        plt.figure(1, figsize=(8,12))
        sns.heatmap(corr_df,
                    cmap="YlGnBu",
                    cbar=True,
                    annot=True,
                    fmt=("2f"),
                    square=True,
                    annot_kws={'size': 15.0},
```

```
                    linewidths=(0.2),
                    xticklabels='auto',
                    yticklabels='auto')
        plt.show()

    return corr_df, high_correlation_counters
```

In [26]:

```
def csv_to_df(strava_csv, replacement_codes, direction=None):

    # Reading CSV-files to pandas.DataFrame
    strava_df = pd.read_csv(strava_csv, sep=',')

    # Removing unnecessary columns
    strava_df = strava_df.loc[:,['edge_uid', 'date', 'forward_trip_count',
                                 'reverse_trip_count']]


    # Reading "replacement_codes" to dictionary and replace edge_uid with other IDs
    new_ids = pd.read_csv(replacement_codes, sep=';', header=None, index_col=0, squeeze=True).to_dict()

    strava_df2 = strava_df.replace({'edge_uid': new_ids})
    strava_df2.rename(columns={'edge_uid': 'linjekoder'}, inplace=True)

    if direction == 'Begge':
        sum_trip_counts = strava_df2['forward_trip_count'] + strava_df2['reverse_trip_count']
        strava_df2['Syklister med Strava'] = sum_trip_counts

    unique_street_segments = strava_df2['linjekoder'].unique()

    output_df = pd.DataFrame()
    for street_segment_id in unique_street_segments:

        current_rows_df = strava_df2.loc[strava_df2['linjekoder'] == street_segment_id]

        data_to_output_df = current_rows_df.groupby(['date']).mean()
        output_df = output_df.append(data_to_output_df)

    valid_ids = list(new_ids.values())
    output_df = output_df[output_df['linjekoder'].isin(valid_ids)]

    convert_dict = {'linjekoder': int, 'forward_trip_count': int, 'reverse_trip_count': int, 'Syklister med Strava': int}
    output_df = output_df.astype(convert_dict)
```

```
    output_df = output_df.reset_index()

    return output_df
```

In [27]:
```
def predict_total_cyclist(num_strava, x1=0, x2=0, x3=0):

    model_ols = 132.652 + 4.257*num_strava + 328.574*x1 + 540.228*x2 + 795.932*x3

    model_poisson_squared = 15.33 + 0.066*num_strava + 5.643*x1 + 10.48*x2 + 14.7*x3
    model_poisson = model_poisson_squared**2

    return model_ols, model_poisson
```

In [73]:
```
def prediction_from_df(input_data_df):

    output_data_df = input_data_df.copy()
    all_ols_values = []
    all_poisson_values = []
    for index, row in input_data_df.iterrows():
        num_strava = row['Syklister med Strava']
        ols_value, poisson_value = predict_total_cyclist(num_strava, x1=1)

        all_ols_values.append(int(ols_value))
        all_poisson_values.append(int(poisson_value))

    output_data_df.insert(loc=5, column='Model OLS', value=all_ols_values)
    output_data_df.insert(loc=6, column='Model Poisson', value=all_poisson_values)

    return output_data_df
```

In [72]:
```
def comparing_correction(before_df, after_df):

    unique_street_segments = before_df['linjekoder'].unique()

    output_df = pd.DataFrame(columns=['Linjekoder', 'OLS', 'Poisson'])

    ols_before_all_sum = before_df['Modell ols (før)'].sum()
    poisson_before_all_sum = before_df['Modell poisson (før)'].sum()
```

```
    ols_after_all_sum = after_df['Modell ols (etter)'].sum()
    poisson_after_all_sum = after_df['Modell poisson (etter)'].sum()

    print(unique_street_segments)
    for street_segment_id in unique_street_segments:

        # OLS
        street_segment_before = before_df.loc[before_df['linjekoder'] == street_segment_id]
        sum_street_before = street_segment_before['Modell ols (før)'].sum()

        street_segment_after = after_df.loc[after_df['linjekoder'] == street_segment_id]
        sum_street_after = street_segment_after['Modell ols (etter)'].sum()

        adjusted_bicycle_volume_ols = (sum_street_after/ols_after_all_sum
                                       - sum_street_before/ols_before_all_sum) * ols_before_all_sum

        # Poisson
        street_segment_before = before_df.loc[before_df['linjekoder'] == street_segment_id]
        sum_street_before_poisson = street_segment_before['Modell poisson (før)'].sum()

        street_segment_after = after_df.loc[after_df['linjekoder'] == street_segment_id]
        sum_street_after_poisson = street_segment_after['Modell poisson (etter)'].sum()

        adjusted_bicycle_volume_poisson = (sum_street_after_poisson/poisson_after_all_sum
                                           - sum_street_before_poisson/poisson_before_all_sum)* poisson_before_all_sum

        data_to_df = [street_segment_id, adjusted_bicycle_volume_ols, adjusted_bicycle_volume_poisson]

        output_df.loc[len(output_df)] = data_to_df

    return output_df
```

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway