Statistisk sentralbyrå
Statistics Norway

Metodedokumentasjon om imputering av byggeår til fritidsbygg

TALL SOM FORTELLER
NOTATER / DOCUMENTS
2019 / 40

Lisa Li og Susie Jentoft

I serien Notater publiseres dokumentasjon, metodebeskrivelser, modellbeskrivelser og standarder.

Publisert 8. november 2019

ISBN 978-82-587-1015-5 (elektronisk)
ISSN 2535-7271 (elektronisk)

Standardtegn i tabeller	**Symbol**
Tall kan ikke forekomme	.
Oppgave mangler	..
Oppgave mangler foreløpig	…
Tall kan ikke offentliggjøres	:
Null	-
Mindre enn 0,5 av den brukte enheten	0
Mindre enn 0,05 av den brukte enheten	0,0
Foreløpig tall	*
Brudd i den loddrette serien	—
Brudd i den vannrette serien	|
Desimaltegn	,

Forord

Dette prosjektet startet i 2018, med hovedmål å utvikle en metode for vurdering av nye fritidsbygg uten byggeår, og er et samarbeidsprosjekt mellom Seksjon for metoder og Statistisk sentralbyrås GIS ressurssenter. Fra GIS ressurssenter har Margrete Steinnes, Trine Haagensen og Kirsten Elisabeth Holz fra Seksjon for eiendoms-, areal- og primærnæringsstatistikk bidratt med forberedelsen, kobling av data, støtte til verktøy og selve problemstillingen.

Statistisk sentralbyrå, 24. oktober 2019

Arvid Olav Lysø

Sammendrag

Prosjektet startet i 2018 for å utvikle en metode for vurdering av nye fritidsbygg uten byggeår. Statistisk sentralbyrå publiserer statistikk over nye fritidsbygg, spesielt bygg bygd de siste 5 årene. Hvert år registreres det nye fritidsbygg, hvor noen av dem er registrert uten byggeår.

I denne studien bruker vi flyfoto for å identifisere byggeåret til et utvalg av fritidsbygg med tanke på å lage en imputeringsmetode for byggeår. Ved hjelp av flyfotoene ble det notert ned hvilket år bygningsomrissene til fritidsbyggene i utvalget ble først sett med bygg, eller hvilket år bygningsomrissene ble sist sett uten bygg. Fra vår manuelle sjekk av utvalget viser det seg at flertallet, ca. 90 prosent, er etterregistrert. Ved bruk av flyfoto utviklet vi en imputeringsmetode som kan tas i bruk for estimering av hvor mange og hvilke fritidsbygg som skal inkluderes i fremtidige statistikkpubliseringer.

Imputeringsopplegget har to faser: 1) å imputere en binomisk verdi som forteller om fritidsbyggene er bygget før eller innen de siste 5 årene, og 2) å imputere byggeåret til fritidsbyggene fra de siste 5 årene.

Fase 1: Det er testet hovedsakelig tre forskjellige imputeringsmetoder: hotdeck-imputering, kalibrert imputeringsmodell med en terskelverdi og random forest-imputering. Disse testene er brukt til å bestemme om fritidsbyggene er bygd før eller innen de 5 siste årene. For alle imputeringsmetodene delte vi utvalget i test- og treningsdatasett. Resultatene til de predikerte verdiene ble sammenlignet med de observerte verdiene fra flyfotosjekket. Vi fokuserte på: 1) nøyaktighet, og 2) fordelingen på fritidsbygg bygd før eller innen de 5 siste årene, altså hvor mange som er imputert til det motsatte. Terskel-verdien, *Imp,* fungerte best med 93,84 prosent i nøyaktighet og 0,16 i forhold.

Fase 2: Fritidsbygg som er antatt å være bygd innen de siste 5 årene får imputert et byggeår. På grunn av en veldig liten datamengde var det ikke mulig å teste forskjellige imputeringsmetoder med meningsfulle resultater. Derfor bruker vi en enkel hotdeck-metode med en begrensingsandel som er lik det imputerte resultatet til det siste året.

Denne studien har vist en nyttig bruk av flyfoto i kvalitetsarbeid i SSB. Det åpner opp for andre analyseoppgaver og kvalitetssjekk. Vi anbefaler å jobbe videre med denne datakilden for å finne flere bruksmuligheter.

Innhold

1. Innledning

Hvert år registreres nye fritidsbygg, hvor noen av fritidsbyggene er registrert uten byggeår. Bygg uten byggeår kan enten være nye bygg med en feil i rapporteringen eller eldre bygg registrert i ettertid, noe som indikerer en underdekning i bygningsregisteret (Matrikkelen). Statistisk sentralbyrå (SSB) publiserer statistikk over nye fritidsbygg og ønsker å estimere byggeåret til denne gruppen med ukjent byggeår for å vite hvor stor andel som kan inkluderes i statistikken. Det er spesielt viktig å vite om byggene ble bygget de siste 5 årene.

Flyfoto er detaljerte bilder tatt fra luften over et område. Fotoene er justert til et geografisk koordinatsystem og sydd sammen. Opplysninger fra fotoene kan brukes til å koble til andre standardiserte datakilder, for eksempel bygningsregisteret. Koblede (integrerte) datasett kan gi mer informasjon enn datakildene hver for seg. De kan også brukes til andre formål, for eksempel kvalitetsanalyse. I denne studien bruker vi flyfoto for å identifisere byggeåret til et utvalg av fritidsbygg med tanke på å lage en imputeringsmetode for byggeår.

Studien er basert på nye fritidsbygg uten registrert byggeår. Enhetene kommer fra SSBs uttrekk av Matrikkelen, som er Norges offisielle register over bygninger, boliger og adresser. Matrikkelen inneholder opplysninger om selve bygget/eiendommen. Eksempler er eiendommens identifiseringsnummer, geografiske bygningspunkt, areal og bygningstype, i tillegg til opplysninger om hvem eieren er. Enheter i Matrikkelen med bygningstype 161, 162 og 163 viser til fritidsbygg. En kobling mellom to årganger for hvert bygningsnummer vil vise om bygningen har byttet bygningstype, en endring som indikerer at fritidsbygget ikke er nyoppført.

Enhetene fra Matrikkelen blir koblet til FKB-bygg, et norsk register for detaljert bygningsinformasjon, eid av Kartverket. FKB-bygg inkluderer beskrivelser av bygninger og bygningsomriss. Koblingen mellom Matrikkelen og datasettene i FKB-bygg skjer i en stegvis prosess basert på geografiske koblinger; først på bygningspunkt, deretter på tilbyggspunkt og til slutt på adressepunkt. Prosessen avsluttes når man finner en kobling. Denne metoden ble utviklet i 2012, og er siden blitt forbedret.

Dataene til denne studien er fra 2017 og inkluderer $N = 823$ enheter (fritidsbygg) uten byggeår. I samme år publiserte SSB opplysninger om $T = 4624$ nye fritidsbygg.

2. Metoder for analyse og imputering

2.1. Utvalg

I utvalget er det trukket $n_u = 200$ tilfeldige fritidsbygg, uten tilbakelegging. Disse danner basis for analyse og utvikling av imputeringsopplegget. Dette antallet var rimelig med tanke på ressursbruken for å håndtere manuelle flyfotosjekk i løpet av en begrenset tid. Dette er den første gangen en slik analyse er gjort, og vi hadde ingen forutsetninger for å bruke stratifiseringsvariabler eller spesielle allokeringer.

2.2. Flyfotosjekk

Bygningsomrissene til fritidsbyggene i utvalget ble koblet til flyfoto med flere årganger for å identifisere byggeåret. Hvert bygningsomriss ble notert med hvilket år bygget ble først sett på flyfoto (variabel: *først sett)*, og hvilket år bygningsomriss ble sist sett uten bygg (variabel: *sist sett*). Vi så på flyfoto fra årgangene 2004 og

senere. Når byggene ikke ble sett ved siste flyfotodato, ble de regnet som bygg fra registreringsåret, som i dette tilfelle er 2017. Fritidsbyggene blir da satt automatisk til 2017 i variabelen *først sett*.

I tillegg er det lagd en binomisk variabel som forteller om fritidsbygget er eldre eller nyere, grensen er de 5 siste årene, som i dette tilfelle er fra og med 2013. Flyfoto for fritidsbyggene ble analysert ved bruk av en WMS (Web Map Service) i ArcGIS. Flyfoto eksisterer ikke for alle årganger, og noen foto har heller ikke god kvalitet.

Fra flyfotoene ser vi at antatte fritidsbygg også kan være påbygg til en eksisterende bolig, terrasse, eller et avgrenset område på størrelse med et dukkehus eller postkasse. Vi har derfor satt noen begrensninger for hva vi mener er fritidsbygg i denne analysen: 1) bygninger skal være større enn 10 m^2, 2) nærmeste avstand til vei er mer enn 2 meter, og 3) nærmeste avstand til et bygg er mer enn 4 meter. Begrensningene gjør at utvalget blir redusert til $n_u = 187$ fritidsbygg uten byggeår for 2017, og den totale populasjonen blir redusert til $N = 784$ fritidsbygg uten byggeår.

Senere har vi innsett at koblingen mellom Matrikkelen og FKB-bygg ikke alltid har vært optimal. De antatte fritidsbyggene som vi trodde var påbygg, terrasse m.m. kan skyldes avvik mellom GPS-punktene til Matrikkelen og FKB-bygg, noe som igjen vil føre til en feil i bygningsomriss. Det antatte fritidsbygget blir istedenfor koblet til et annet bygg. På grunn av feil i koblingen er det umulig å vite hvilket fritidsbygg som har ukjent byggeår, men bygget ligger ikke langt unna koblingspunktet.

2.3. Imputeringsmetode

Imputering er gjort for enheter utenfor utvalget, totalt $n_e = N - n_u = 597$ enheter. I tillegg kan imputeringsmetoden brukes for alle enheter som mangler byggeår i de neste årene. Imputeringsopplegget har to faser: 1) å imputere en binomisk verdi som forteller om fritidsbyggene er bygget før eller innen de siste 5 årene, og 2) å imputere byggeåret til fritidsbyggene fra de siste 5 årene.

Fase 1 – imputering av binomisk variabel for fritidsbygg bygd før eller innen de 5 siste årene

I første fase imputeres en binomisk variabel som forteller om fritidsbygget er bygd før eller innen de 5 siste årene. Den er definert slik:

$$y = \begin{cases} 0, & \text{fritidsbygg bygd innen de 5 siste årene,} \\ 1, & \text{fritidsbygg bygd før de 5 siste årene.} \end{cases}$$

Vi har testet hovedsakelig tre forskjellige imputeringsmetoder: hotdeck-imputering, kalibrert imputeringsmodell med en terskelverdi og random forest-imputering.

Hotdeck

Hotdeck-imputering er en donor-imputeringsmetode som trekker tilfeldig fra de observerte verdiene. Trukne verdier blir så imputert til alle enheter som mangler verdi, altså de som mangler bygningsstatus bygd før eller innen de 5 siste årene. Fordelen med metoden er at den er enkel å gjennomføre og gir en realistisk imputeringsverdi, det vil si at ingen enheter får et årstall med desimaler eller et årstall utenfor observasjonsårene. Ulempen er at de tilfeldige komponentene i metoden kan gi et resultat som er langt unna hva vi forventer å få.

Kalibrert imputeringsmodell

I en kalibrert imputeringsmodell har hvert nivå et fast antall enheter, som beregnes ut fra summen av predikerte verdier fra en modell. Denne imputeringstypen brukes også i andre statistikker i SSB, et eksempel er imputering av overtid i AKU (Jentoft

& Mevik, 2014). I denne studien ble den kalibrerte summen bestemt fra predikerte verdier i en logistisk regresjonsmodell.

Fritidsbyggene har mange forklaringsvariabler; listen over de mest interessante variablene er i Vedlegg A. Istedenfor å bruke alle variablene – noe som ikke nødvendigvis gir gode resultater – tilpasses først modellen med Akaike information criterion (AIC). Vi fjernet kommunenumrene i AIC-testen, da utvalget ikke representerte alle kommunene. I tillegg ville testen gi veldig mange variabler, en for hvert kommunenummer, som sannsynligvis ville føre til en overtilpasset modell med dårlig evne til prediksjon. Modellen predikerer tall mellom 0 og 1 basert på forklaringsvariablene. Summen av de predikerte tallene bestemmer terskelverdien for om predikerte tall skal gjøres om til 0 eller 1, og dermed også fordelingen mellom antall observerte byggeår bygget før eller innen 5-års grensen. Verdiene 1 og 0 forteller om fritidsbyggene er klassifisert som eldre bygg ($y = 1$) eller nyere bygg ($y = 0$). Antall bygg av hver type benevnes med henholdsvis n_1 og n_0.

AIC

AIC er en av de viktigste informasjonskriteriene. Den generelle formelen er

$$\mathrm{AIC}(M) = 2\text{log-likelihood}_{\max}(M) - 2\dim(M),$$

hvor M er kandidatmodell og $\dim(M)$ er lengden av dens parametervektor. Vi ønsker ideelt sett en enkel modell med god tilpasning, og AIC fungerer som et penelarisert[1] log-likelihood kriterium som balanserer mellom modelltilpasning (hvor god log-likelihood er) og modellens kompleksitet (antall parametre). Modellen med høyest AIC-score blir så valgt.

Logistisk lineær modell

En generalisert lineær modell er en fleksibel generalisering av ordinær lineær regresjon, den tillater bl.a. at responsvariabelen ikke er normalfordelt. Et eksempel på en slik modell er logistisk regresjon, som ofte brukes når Y er en binær variabel (tar verdiene 0 og 1). Den logistiske regresjonen med n forklaringsvariabler $x_1, x_2, \ldots, x_n$ er

$$\Pr(Y = 1 \mid x) = \pi(x) = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n}}{1 + e^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n}},$$

hvor $\alpha, \beta_1, \beta_2, \ldots, \beta_n$ er estimatene til variablene som ga høyest AIC-score. Modellen kan også skrives som

$$\text{logit}\,\pi(x) = \log\frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n,$$

der log står for den naturlige logaritmen.

Deretter predikerer modellen verdier basert på den lineære modellen. Tallene som blir predikert er alle mellom 0 og 1. Vi finner så en terskelverdi som bestemmer om den predikerte verdien skal indikere 0 eller 1. Vi testet fem forskjellige terskelverdier, mean, median, Imp, Tersk og Forh. Modellene er basert på utvalget som igjen blir splittet i trenings- og testdatasett. Antall observasjoner i henholdsvis trenings- og testdatasettet benevnes n_{tr} og n_{te}.

Terskelverdien Imp er definert med denne formelen

$$\text{Imp} = 1 - \frac{1}{n_u}\sum_{i=1}^{n_u} \hat{y}_i,$$

hvor n_u er antall observasjoner fra utvalget, og $\hat{y}_i$ er deres predikerte verdier.

[1] Penelarisert kommer fra det engelske ordet penalize, som betyr å straffe.

Terskelverdien Tersk er andelen bygg bygd før 2013. Altså

$$\text{Tersk} = \frac{n_{1,tr}}{n_u} = 1 - \frac{n_{0,tr}}{n_u},$$

hvor $n_{0,tr} = 18$ er antall bygg fra de siste 5 årene i treningsdatasettet og $n_u = 187$ er totalt antall bygg i utvalget. Da blir $n_{1,tr}$ antall bygg bygd før de siste 5 årene i treningsdatasettet.

Terskelverdien Forh er den andelen treningsdatasettet utgjør av det totale utvalget.

$$\text{Forh} = \frac{n_{tr}}{n_u} = 1 - \frac{n_{te}}{n_u},$$

hvor $n_{te} = 46$ er størrelsen på testdatasettet.

ROC I tillegg til den vanlige terskelen, ble det også testet med ROC-kurve (Reciever Operating Characteristic). Denne metoden prøver å finne den optimale verdien av en test som gir flest sanne positive (høy sensitivitet) og færreste falske positive (høy spesifisitet).

Random forest Random forest er en veiledet (supervised) maskinlæringsalgoritme som kan brukes til både klassifisering og numerisk prediksjon. Hensikten er å lage mange klassifiserings- eller regresjonstrær fra dataene og deretter lage et gjennomsnitt for å redusere variansen. Metoden har blitt populær, og den har vist seg å fungere godt i mange sammenhenger (Hastie et al., 2009). Vi testet random forest for imputering av y, altså om fritidsbygg er bygd før eller innen de 5 siste årene, med variabler beskrevet i Vedlegg A.

Testmetode Vi testet forskjellige imputeringsmetoder for å se om fritidsbyggene er bygget før eller innen de 5 siste årene. For alle imputeringsmetodene delte vi utvalget i test- og treningsdatasett. Først trente vi treningsdatasettet med observerte verdier, og deretter testet vi resultatene på testdatasettet for å se hvor gode prediksjoner vi fikk til slutt. Resultatet på testsettet ble da sammenlignet med observerte resultater fra flyfotosjekket. For hver imputeringsmetode ble samme prosedyre gjentatt 100 ganger med nye trenings- og testdatasett hver gang og kryssvalidert mellom de observerte og predikerte verdiene, se konfusjonsmatrisen i Figur 2.1.

Konfusjonsmatrisen i Figur 2.1 sammenligner de observerte verdiene fra treningsdatasettet med de predikerte verdiene fra testdatasettet. Husk at antall fritidsbygg fra 2013 og senere er benevnt med n_0, mens antall fritidsbygg fra før 2013 er benevnt med n_1. Antall enheter der observerte verdier er det samme som predikerte verdier, blir da n_{00} for nyere bygg og n_{11} for eldre bygg. Der vi har en endring fra de observert til predikert verdi, vil n_{01} være antall bygg observert som nyere bygg og predikert til eldre bygg, og n_{10} antall bygg observert som eldre bygg og predikert til nyere bygg.

Figur 2.1 Eksempel på konfusjonsmatrise

		Predikert verdi: Nytt bygg (0)	Predikert verdi: Gammelt bygg (1)
Observert verdi	Nytt bygg (0)	n_{00}	n_{01}
	Gammelt bygg (1)	n_{10}	n_{11}

Før imputeringsmetoden logistisk regresjon ble brukt, gjennomførtes en AIC-test for å se hvilke variabler som var viktige, hvor kommunenummer var manuelt fjernet. Det som er viktig å se på er: 1) nøyaktighet og 2) fordelingen på fritidsbygg bygd før eller innen de 5 siste årene, altså hvor mange som er imputert som det motsatte (en feil).

Vi definerer nøyaktighet, δ, som andel enheter som er riktig predikert,

$$\delta = \frac{n_{00} + n_{11}}{n_u},$$

mens fordelingen, ρ, viser i hvilken grad det er en overvekt av feilpredikeringer den ene veien,

$$\rho = \frac{|n_{01} - n_{10}|}{n_{01} + n_{10}},$$

Størrelsen ρ gir en indikasjon på *statistisk ekvivalens* mellom observerte og predikterte tallene. Selv om to forskjellige metoder kan gi samme nøyaktighet, kan de gi grunnlag for svært forskjellige statistikker. Dette er hva ρ måler. Hvis $\rho \approx 0$, vil predikerte tall gi den samme statistikken som observerte tall, og vi oppfatter dette som en god imputeringsmetode.

Fase 2 – imputering av byggeår

I den andre fasen imputerer vi byggeåret for nye bygg bygd de siste 5 årene. På grunn av lite data var det ikke mulig å få meningsfulle resultater ved å teste ulike imputeringsmetoder. Derfor bruker vi en enkel hotdeck-metode med en beskrankningsandel som er lik det imputerte resultatet til det siste året. Kalibreringen for fritidsbygg bygd det siste året skal være lik andelen i utvalget.

3. Resultater fra analyse og imputering

3.1. Analyse av variablene til fritidsbygg bygd før eller innen de 5 siste årene

I utvalget var det $n_u = 187$ av 200 bygninger som ble klassifisert som fritidsbygg ved flyfotosjekket. Av dem var kun $n_0 = 18$ fritidsbygg lagt inn som fritidsbygg bygd de siste 5 årene, mens $n_1 = 169$ fritidsbygg er eldre enn de siste 5 årene, en etterregistrering. Utvalget er skjevt fordelt, da 8,63 prosent av fritidsbyggene er bygd fra 2013 og 91,37 prosent er bygd før 2013. En oppsummering av nøyaktighet og balansen i fordelingen fra 100 simuleringer vises i Tabell 3.1.

Tabell 3.1 Sammenligning av imputeringsmetoder med terskel

	Terskel	Nøyaktighet	Fordeling
Kalibrert logistisk modell	Median	53,78	0,90
	Mean	63,85	0,74
	Imp	93,84	0,16
	Tersk	61,41	0,78
	Forh	86,50	0,37
ROC	Median	84,09	0
	Mean	83,87	0
	Imp	93,22	0
	Tersk	83,70	0
	Forh	86,13	0
Hotdeck	-	82,68	0
Stratified hotdeck	Kvantil 2	83,58	0,21
	Kvantil 3	86,04	0,27
Random forest	Normal	91,15	0,98
	CV	90,80	0,92

Kilde: Statistisk sentralbyrå.

Resultatet fra kryssvalideringen viser at den kalibrerte logistisk modellen var den beste for vårt formål. Variablene som ga den beste modellen var *kildeGAB* (GAB/FKB – hvor objektet er registrert fra), *nerVei* (avstand til nærmest vei) og *andel* (andel av nye registrerte fritidsbygg i kommune). Disse ga høyest AIC-score med en verdi på 74,66. Vi observerte at noen kommuner hadde mange nylig registrerte fritidsbygg, noe som sannsynligvis gjenspeilet en ekstra innsats i kommunen med etterregistrering enn en reell vekst i bygging av fritidsbygg. Dette tok vi hensyn til med variabelen *andel*.

Fra utvalget er andel fritidsbygg bygd før 2013 på 90,37 prosent og andel fritidsbygg bygd fra 2013 på 8,63 prosent. Sammenligning mellom resultatene fra imputeringsmetodene og manuellsjekkingen, ser vi at terskelen som heter Imp ga best fordeling mellom de feilimputerte.

Estimater fra denne generaliserte lineære modellen gir parameterne fra Tabell 3.2, hvor to av variablene, *KildeGab* og *andel*, er signifikante.

Tabell 3.2 Estimater til den generaliserte lineære modellen

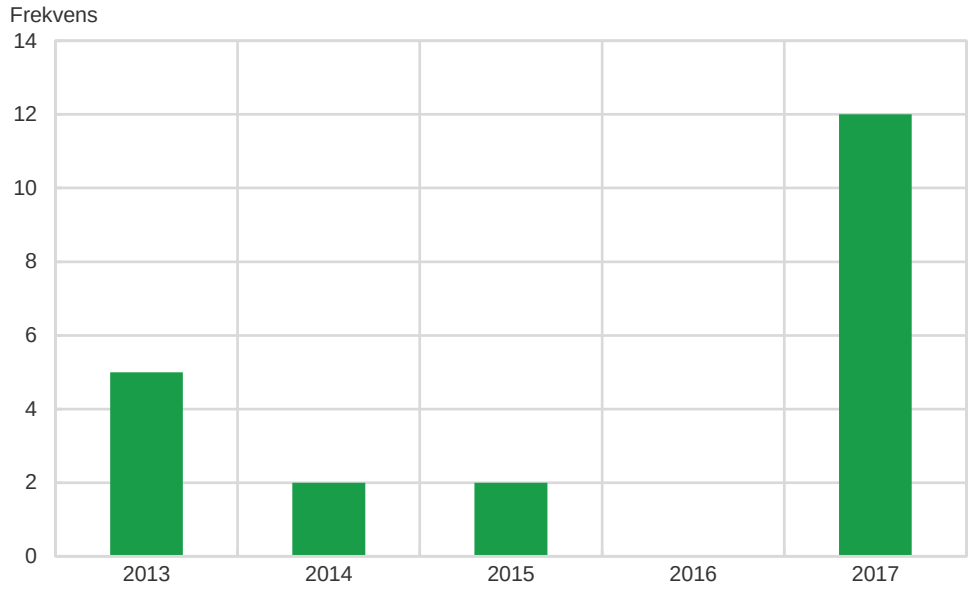
Parameter	Estimater	Forklaring
α	0,8856	Skjæringspunktet
β_1	-0,1487	KildeGAB
β_2	0,00000153	nerVei
β_3	2,123	Andel

Kilde: Statistisk sentralbyrå.

4. Oppsummering av byggeår for fritidsbygg bygd de 5 siste årene

I Figur 4.1 ser vi fordelingen til hvor mange fritidsbygg som er bygd mellom 2013 og 2017 fra utvalget. Ingen fritidsbygg ble registrert i 2016, og cirka halvparten av fritidsbygg er registrert i 2017. Flyfoto viser at det mangler kart for noen av årgangene, eller at fritidsbygget ikke er inkludert i flyfotoet.

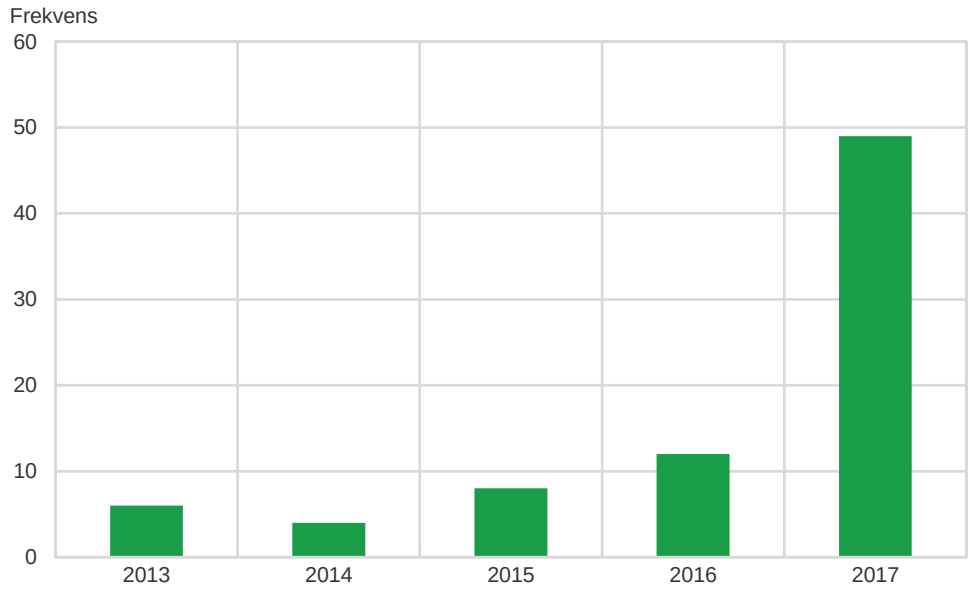
Figur 4.1 **Histogram for første året fritidsbygget ble først sett på foto for utvalget. Kun de som ble bygd innen de 5 siste årene.**



Kilde: Statistisk sentralbyrå.

Figur 4.2 viser fordelingen til fritidsbyggene som manglet byggeår og bygd fra 2013, både observerte og imputerte tall. Tilfeldige komponenter av imputeringsmetode betyr at ikke alle årene før 2017 er helt likt representert. Små tall betyr at dette ikke kommer til å påvirke statistikken særlig mye.

Figur 4.2 **Histogram for første året fritidsbygget ble først sett på foto for utvalget og imputerte tall. Kun de som ble bygd innen de 5 siste årene.**



Kilde: Statistisk sentralbyrå.

5. Konklusjoner

I denne analysen har vi sett på fritidsbygg uten byggeår, noe som enten kan være etterregistrerte bygg, eller bygg som ikke har blitt registrert på grunn av en feil i rapporteringen. Det er fritidsbygg fra de siste 5 årene som ikke har blitt registrert som vi er interessert i. Altså hvor stor andel fritidsbygg som er bygd de siste 5 årene som kan inkluderes i statistikken. Totalt er det $N = 823$ fritidsbygg som ikke har byggeår. Av disse sjekket vi manuelt $n_u = 187$ fritidsbygg ved hjelp av flyfoto. Av dem var kun $n_0 = 18$ fritidsbygg lagt inn som fritidsbygg bygd de siste 5 årene, mens $n_1 = 169$ fritidsbygg var bygd før de siste 5 årene, en etterregistrering. Det betyr at flertallet på ca. 90 prosent er etterregistrert og er blitt ekskludert fra den vanlige rapporteringen.

Fritidsbygg som er sjekket manuelt, delte vi inn i trenings- og testdatasett. Vi brukte observasjonene i treningsdatasettet til å trene imputeringsmetodene, og deretter predikere verdiene for testdatasettet. Imputeringsmetoden som ga best resultat for vårt formål er kalibrert logistisk modell med terskelverdien Imp. Denne imputeringsmetoden ga 93,84 prosent i nøyaktighet og 0,16 i forhold.

For fritidsbyggene som ikke ble sjekket manuelt brukte vi imputeringsmetoden for å imputere verdier. Resultatet ble $N_0 = 79$ fritidsbygg bygd fra de 5 siste årene som kunne vært publisert i 2017. I samme år publiserte SSB opplysninger om $T = 4624$ nye fritidsbygg. Det betyr at underestimeringen av nye fritidsbygg er i underkant av 2 prosent i tidligere statistikkpubliseringer.

Ved bruk av flyfoto har gjort oss i stand til å utvikle en imputeringsmetode som kan tas i bruk for estimering av hvor mange og hvilke fritidsbygg som skal inkluderes i fremtidige publiseringer. Metoden kan også brukes i fremtiden uten et nytt utvalg, men flere ressurskrevende manuelle sjekker behøves. Flyfoto eksisterer ikke for alle årganger, og noen av dem har heller ikke god kvalitet.

Denne studien har vist en nyttig bruk av flyfoto i kvalitetsarbeid i SSB. Det åpner opp for andre analyseoppgaver og kvalitetssjekk. Vi anbefaler å se på flere muligheter for hvordan man kan utnytte denne datakilden.

Referanser

Jentoft, Susie & Mevik, Anna-Karin (2014): *Arbeidskraftundersøkelsen. Imputering av overtid ved indirekte intervju.* Notater 2014/13, Statistisk sentralbyrå

Hastie, Trevor; Tibshirani, Robert & Friedman, Jerome (2009): *The elements of statistical learning. Data mining, inference and prediction*. Springer science, NY

Heldal, Johan (2006): *Logistisk regresjon – kurskompendium i byråskolens kurs SM507.* Notater 2006/54, Statistisk sentralbyrå

McCullagh, P. & Nelder, J.A (1989). *Generalized linear models (2nd edition).* Suffolk, Great Britain: Chapman & Hall.

Vedlegg A: Tilgjengelige variabler

Tabellen viser alle variablene som er brukt i denne analysen, de fem siste variablene er lagt til shape-filen[2] for å gjøre analysen enklere.

A. 1 Variabler og variabelbeskrivelse

Variabel	Variabelbeskrivelse
FID	FID er form for ID
areal	Arealen til bygningen
SMAT	SMAT2017 – større sannsynlighet for ny bygg
kilder	GAB/FKB – hvor objektet er tatt fra
nerDIST	Nærmeste distanse til bygg
nerVei	Nærmeste distanse til vei
byggType	Bygningstype – 161, 162, 163
areal2	Areal til nærmeste bygg
byggar2	Byggeår til nærmeste bygg
fylke	Fylke
byggType2	Bygningstype til nærmeste bygg
kommunenr	kommunenummer
andel	Andelen til nye bygg. Nye bygg/eksisterende bygg per kommune
For2013	Binomisk variabel for om fritidsbygget ble bygd før eller etter 2013.
ForstSett	Første år sett med bygg
SistSett	Sist år sette uten bygg

Kilde: Statistisk sentralbyrå.

[2] En shape-file inneholder informasjon system (GIS) verktøy.

Vedlegg B: Tabell for nye fritidsbygg for de siste fem årene

Figur B.1 Nye fritidsbygg de siste 5 årene.

Nye fritidsbygg i de siste fem år. Antall per år. Fylke

	Nye fritidsbygg i det siste året per 1. januar [1]					Antall fritidsbygg per 01.01.2017
	2013	2014	2015	2016	2017	
Østfold	73	127	77	84	96	20 203
Akershus	22	27	39	34	46	14 717
Oslo		2	3		1	2 272
Hedmark	391	409	373	515	451	36 744
Oppland	487	507	496	682	856	50 060
Buskerud	539	512	506	679	722	46 249
Vestfold	62	78	57	59	67	14 639
Telemark	379	372	316	417	431	31 103
Aust-Agder	177	178	133	162	152	19 104
Vest-Agder	269	297	255	288	221	21 450
Rogaland	219	216	178	188	142	19 801
Hordaland	310	330	298	302	239	33 009
Sogn og Fjordane	101	131	128	127	102	13 345
Møre og Romsdal	207	196	309	187	190	21 160
Sør-Trøndelag	290	323	305	345	346	33 688
Nord-Trøndelag	148	143	181	176	168	17 794
Nordland	270	234	191	201	212	34 983
Troms Romsa	134	101	102	117	102	15 369
Finnmark Finnmárku	74	75	92	75	80	12 312

[1] 'Nye fritidsbygg' inkludere kun nybygde, og er basert på igangsattdato. Eksisterende bygg som ble omgjort til fritidsbygg er ikke tatt med, heller ikke påbygg/tilbygg. Antall nye fritidsbygg gjelder for perioden 1. januar forrige år til 1. januar dette år.

Kilde: Statistisk sentralbyrå: https://www.ssb.no/331533/nye-fritidsbygg-i-de-siste-fem-ar.antall-per-ar.fylke

Figurregister

Tabellregister