Funksjonalitet og programmeringsspråk i Metodebiblioteket

Statistisk sentralbyrå

November 2026

Bakgrunn

Det er et ønske om å utvide funksjonaliteten i Metodebiblioteket for bedre å møte behovene statistikkseksjonene har i overgangen til Dapla. På Dapla er både R og Python tilgjengelige programmeringsspråk, og det er opp til hver seksjon/statistikk å velge hvilket språk som skal benyttes i produksjonsløpet. Python er det mest brukte språket for statistikkproduksjon i SSB, men det er også statistikker som bruker R. Per oktober 2025, hadde 66 prosent av «stat»-repoene på SSBs GitHub Python som hovedspråk, 13 prosent hadde R som hovedspråk, og resten en blanding av Jupyter notebooks (som kan være R eller Python), SAS, og andre språk.

Dette notatet gjør rede for Seksjon for metoder sine anbefalinger om hvilke prinsipper som bør gjelde for Metodebiblioteket når det gjelder programmeringsspråk for metodefunksjoner. Notatet gir videre en oversikt over hvilket språk som per i dag er hovedspråket for ulike områder av metodefunksjoner.

Valg av programmeringsspråk

Metodebiblioteket består i dag av funksjoner som er utviklet og vedlikeholdes av SSB og av funksjoner som er hentet fra andre, f.eks. andre statistikkbyråer. Funksjoner som hentes fra andre, utløser i liten grad behov for ressursbruk i SSB knyttet til vedlikehold.

At en funksjon i Metodebiblioteket finnes i samme programmeringsspråk som resten produksjonskoden den skal brukes i, gir noen fordeler:

- Enklere integrasjon av funksjoner i produksjonsprosessene
- Større engasjement og støtte fra statistikkseksjonene i vedlikehold og videreutvikling av metodiske funksjoner
- Bedre forståelse og transparens, ettersom funksjonene er skrevet i et språk brukerne allerede beherske.

Dersom metodefunksjoner ikke finnes i samme språk som produksjonskoden kan dette føre til:

- Økt kompleksitet i integrasjon
- Færre som har kompetanse på de aktuelle funksjonene og dårligere vedlikehold og videreutvikling

Samtidig, vil det å ha duplikater av funksjoner (samme funksjon programmert i både R og Python) kunne være ressurskrevende:

- Krever ekstra ressurser til utviklingsarbeid, særlig fordi noen av pakkene er svært komplekse og krever omfattende koding

- Mer omfattende vedlikeholdsarbeid, «dobbelt opp av alt».

Både R og Python har støtte for å kjøre kode fra det andre språket i sitt eget miljø. Med R-pakken reticulate kan man integrere Python-funksjoner direkte i R, mens Python-pakken rwrapr gjør det mulig å kjøre R-funksjoner i Python. Dette gjør funksjonene i Metodebiblioteket tilgjengelige, uavhengig av hvilket språk som benyttes i produksjonsløpet.

Disse pakkene i dag har visse begrensninger, og de fungerer ennå ikke alltid optimalt – spesielt ikke i mer komplekse eksempler.

Vurderinger

Fordelene og ulempene ved valg av programmeringsspråk må veies opp mot hverandre. De ulike metodeområdene i SSB varierer i kompleksitet, og teamene som jobber med områdene har ulik erfaring med de to programmeringsspråkene. Det er også variasjoner i bruk av programmeringsspråk i statistikkavdelingene - innen økonomi og næringsliv brukes primært Python, mens helse- og personstatistikk har en mer blandet bruk.

Kostander ved utvikling og vedlikehold tilsier at det bør være gode grunner til at samme funksjon skal finnes i to språk når både R og Python har støtte for å kjøre kode fra det andre språket.

Innenfor offisiell statistikk har R tradisjonelt vært det mest utbredte programmeringsspråket internasjonalt, særlig blant statistikere og metodeeksperter. Dette henger sammen med et modent økosystem av R-pakker utviklet spesifikt for surveydata, estimeringsmetoder og standarder i offisiell statistikk. Samtidig er bruken av Python i rask vekst, spesielt innen datahåndtering, automatisering, maskinlæring og integrasjon mot moderne dataplattformer. I dag ser vi derfor et mer todelt bilde, der R fortsatt dominerer innen klassisk statistikkfaglig arbeid, mens Python blir stadig viktigere i produksjonsløp og dataingeniørrettede deler av statistikkproduksjonen.

For at SSB skal kunne ta i bruk, videreutvikle og bidra til fellesressursene som utvikles innen offisiell statistikk internasjonalt, er situasjonen i dag at vi må ha tilgang til og beherske R.

Forslag til prinsipper for programmeringsspråk i metodebiblioteket

Under beskrives et forslag til hvilken funksjonalitet i Metodebiblioteket som kan skrives i begge språk, og hva som hovedsakelig kun skal være tilgjengelig i ett av språkene. Oversikten gjelder for funksjoner som er utviklet av og vedlikeholdes av SSB.

For funksjoner som bare er tilgjengelig i ett språk, kan man bruke støtte-pakker eller batch-kjøring for å kjøre kode i sitt miljø/produksjonsløp (om språket er et annet enn det som finnes for funksjonen i Metodebiblioteket).

Følgende prinsipper foreslås:

- Alle funksjoner i Metodebiblioteket, enten de er R-funksjoner eller Python-funksjoner, skal på sikt kunne kjøres både i R-produksjonsmiljø og Python-produksjonsmiljø, for eksempel ved hjelp av reticulate/rwrapr.

- Metodebiblioteket skal som hovedregel ikke inneholde «dubletter/kloner», dvs. at identiske funksjoner finnes i begge kodespråk, med mindre de er utviklet eksternt.
- De fleste metodeområdene i biblioteket vil basere seg på ett hovedspråk. Det betyr at de fleste funksjoner innenfor et område vil være kodet i hovedspråket. Det utelukker ikke at det kan finnes funksjoner (interne eller eksterne) i det andre språket. Valg av språk skal basere seg på dagens tilgjengelighet av funksjoner, fremtidige planer og preferanser i statistikkavdelingene.
- Videreutvikling av funksjonalitet, vil basere seg på hovedspråket til metodeområdet.
- Metodefunksjoner innenfor editering og kontroll av data bør som hovedregel finnes tilgjengelig i begge språk. Dette fordi slike funksjoner ofte er tett integrert i produksjonsprosessene, og det per i dag finnes mye god funksjonalitet tilgjengelig i begge språk.

Forutsetninger

Forutsetninger for at prinsippene skal fungere godt er:

- God generell veiledning for integrering av funksjoner mellom språkene skal utvikles. Dette inkluderer både skriftlig veiledning og instruksjonsvideoer. Begge retninger - fra Python til R og fra R til Python - skal dekkes.
- Infrastrukturen på Dapla støtter kjøring av begge programmeringsspråkene. I praksis betyr dette at både R og Python skal være installert og tilgjengelige i de ulike utviklingsmiljøene (Jupyter, VSCode og RStudio). Det anbefales å inkludere R i SSB-project for å forenkle kjøring av funksjoner i andre språk.

Bruk av prinsippene

Tabellen under gir en oversikt over hva prinsippene betyr i praksis for de ulike metodeområdene. I tabellen er ulike metodeområder knyttet til faser i produksjonsmodellen.

Prosessmodellen	**Temaområde**	**Forslag til hoved programmerings-språk**	**Begrunnelse**
2.4 Planlegge ramme og utvalg	Utvikling av utvalgsplan (strata, størrelse og allokering)	R	Kode som i hovedsak vil skrives og vedlikeholdes av Metodeseksjonen. Det finnes flere gode pakker som vi har erfaring med i R fra ISTAT.
4.1 Etablere ramme og trekke utvalg	Trekking av utvalg	R	God implementering av utvalgsallokeringsfunksjoner i R. AKU trekkes i R.
5.2 Klassifisere og kode	Klassifisering fra tekstfelt + andre variabler	Python	God støtte for ulike maskinlæringsmodeller i Python. Mest kode fra før i Python.
5.3 Kontrollere og validere	Kontrollere data med statistiske metoder	Python & R	Kontroll av data er en integrert prosess i klargjøring og kan være

			vanskelig å skille ut i spesifikke deler. Det er en fordel å ha det i samme språk som resten av produksjonsløpet. Pakker i både R og Python finnes i dag på ulike områder.
	Validere data og input type (legges i fag-biblioteket ikke Metodebiblioteket)	Python & R	Kontroll av data er en integrert prosess i klargjøring av data og kan være vanskelig å skille ut spesifikke deler. Det er en fordel å ha det i samme språk som resten av produksjonsløpet. Pakker i både R og Python finnes i dag på ulike områder.
5.4 Editere og imputere	Manuell editering (legges i fag-biblioteket ikke i Metodebiblioteket?)	Python	Utvikling av en generell ramme for dash-app for Dapla er i Python
	Regelbasert imputering	Python & R	Behov for begge. Det finnes (eksterne) pakker i begge språk.
	Modell & donorbasert imputering	Python & R	Behov for begge. Det finnes (eksterne) pakker i begge språk.
5.6 Beregne vekter	Estimering - person	R	R-pakken ReGenesees som er utviklet av ISTAT er godt utviklet med kalibreringsmetoder for vektberegning som vi mener er vanligvis best for vektberegning i personundersøkelser.
	Estimering – bedrift	Python	Vi bruker ofte modellbaserte metoder for beregning av vekter i bedriftsundersøkelser. Noen av disse er programmert i Python (av SSB). Mange av statistikkseksjonene som jobber med bedriftsundersøkelser, programmerer løpet i Python.
	Indeksberegning	R	Det finnes eksempler i både R og Python, Seksjon for metoder har jobbet med en R-pakke for indeksberegning som inkludere

			variansberegning som ikke finnes i Python.
6.1 Utarbeid produktutkast	Sesongjustering	R	Modellene som brukes aller mest, er programmert i R. Det er mye funksjonalitet som er utviklet (og vedlikeholdt) internasjonalt i R.
6.4 Gjennomføre avsløringskontroll	Tabell undertrykking	R	God funksjonalitet er utviklet internt i SSB som passer til formålet. Pakkene brukes internasjonalt av andre byråer. Svært omfattende og kompleks kode som det ikke er ressursmessig fornuftig å vedlikeholde i to språk.
	Avrunding	R	