

Notat til programrådet: bytte av statistikksoftware i SOSGEO1120

Fra: Torbjørn Skardhamar

Med dette notatet vil jeg løfte problemstillingen om å bytte statistikksoftware for SOSGEO1120: Jeg ønsker å bytte til R fra og med våren 2021.

Ved ISS brukte vi SPSS frem til 2012 da vi byttet til STATA. Det var gode pedagogiske, praktiske og strategiske grunner for dette. Ved samfunnsgeografi har de brukt SPSS frem til 2018, men metodekursene er nå slått sammen til SOSGEO1120 fra og med 2019.

En viktig grunn for byttet fra SPSS til STATA i 2012 var at STATA åpenbart var langt mer i front i utvikling på flere måter. STATA er fremdeles et godt verktøy og dekker dagens behov, men R matcher STATA på all funksjonalitet samtidig som det er betydelig mer i front på en rekke områder.¹ Jeg mener også at R har en del pedagogiske fordeler.

Jeg synes det er rimelig starte en diskusjon om å bytte til R, og at det først og fremst gjøres på SOSGEO1120, men at det etter hvert også bør vurderes på SOS4020 etter å ha høstet erfaringer med SOSGEO1120.

Nåværende software: STATA

Da ISS byttet fra SPSS til STATA i 2012 var dette helt klart regnet som beste valg for samfunnsvitere fordi STATA har en mindre kronglete syntax enn SPSS, men også fordi funksjonaliteten var bedre på mange områder. Dette gjelder særlig avansert regresjonsmodeller, som f.eks. DiD, IV, og kvantilregresjon. STATA har slik sett svært godt utbygd funksjonalitet for det som nå er standard regresjonsteknikker, og særlig teknikker som forbindes med økonometri.

STATA regnes som å være relativt brukervennlig, ved at det både er menystyring og mulighet for å skrive script (såkalte do-filer). Ved ISS har vi primært benyttet script i undervisningen, og siste fire årene utelukkende script. Det er alltid krevende å lære et programmeringsspråk, men det er mitt inntrykk at det har fungert tilfredsstillende i undervisningen.

Jeg mener imidlertid at STATA ikke har en spesielt enkel syntax for nybegynnere, og brukervennligheten er noe overvurdert. I tillegg har STATA har følgende svakheter:

- begrenset funksjonalitet og kronglete syntax for grafikk
- kronglete å lage selv enkle tabeller og eksport til Word (selv med funksjoner som tabout, estout og asdoc)
- svært begrenset GIS-funksjonalitet (herunder å lage kart)
- mindre god støtte for regresjon som ikke er standard innen økonometri (f.eks. additive modeller og geografisk modellering)
- svært begrenset støtte for maskinlæring
- brukes nesten utelukkende i akademisk og ikke i andre yrker
- koster penger å bruke (lisens)

¹ Det finnes også alternativer, derav Python. R er nok ennå lengre fremme på statistikk og Python er i overkant avansert. Det vil trolig være aktuelt å vurdere andre systemer på senere tidspunkt.

Til tross for disse punktene dekker STATA nåværende behov for undervisningen på SOSGEO1120 og SOS4020. Men nåværende pensum setter uansett ikke spesifikke krav til hvilken software.

Hvorfor vurdere R?

Utbredelse og yrkesrelevans

R har lenge vært det dominerende i matnat-fag og statistikk. Ved UiO har statsvitenskap allerede undervist i R en god stund, og økonomi har begynt en overgang til R fra våren 2019. Jeg har forstått det slik at R også oppe til diskusjon på psykologi selv om de fremdeles bruker SPSS/AMOS. Om det er tilsvarende utvikling på andre universiteter vet jeg ikke, men det er neppe andre sosiologiprogrammer som bruker R i dag.

R har vokst særlig kraftig de siste ti år på tvers av nær sagt alle akademiske fag, men også i produksjons- og analysemiljøer i næringslivet.² Dette gjør at R har en bred yrkesrelevans.

R er open source og freeware, så det er gratis å bruke. Dermed kan studentene fritt bruke R i jobbsammenheng senere uavhengig av at arbeidsgiver vil kjøpe lisens.

Brukervennlighet

R har hatt et rykte på seg for å ha en høy brukerterskel. Dette har endret seg radikalt med utviklingen av et subsystem som kalles «tidyverse». (R har flere subsystemer for å dekke ulike behov). Å konsistent bruke «tidyverse» gjør det svært brukervennlig, blant annet fordi syntaxen er optimalisert i forhold til logikk og konsistens, og er bygget på eksplisitte prinsipper også for dataorganisering.³ Denne logikken gjelder også for grafikk via grafikksystemet ggplot2, som også er basert på eksplisitt grafisk teori.⁴

R har masse funksjonalitet for grunnleggende og avansert tabulering, samt eksportmuligheter til alle slags formater som er lette å bruke. Alt dette matcher (og i noen tilfeller overgår) STATA i brukervennlighet.

Jeg har noe erfaring fra en håndfull studenter på både MA og BA nivå som har lært seg R for å skrive sine oppgaver. Dette har i hovedsak vært svært vellykket, og det har ikke vært spesielle problemer med å tilegne seg dette programmeringsspråket.

Funksjonalitet

R matcher STATA på all funksjonalitet knyttet til de teknikkene vi underviser i i dag, herunder avanserte regresjonsteknikker.⁵ R er i tillegg et svært kraftig verktøy også for andre teknikker,

² Et eksempel er at Microsoft har integrert R sin tjeneste for cloud computing, Azure. Andre velkjente firmaer som bruker R inkluderer Google, Facebook og Twitter, men også analysemiljøer i AirBnB, Uber og New York Times bruker R.

³ H. Wickham (2013) «Tidy data», *Journal of Statistical software*, URL: <https://www.jstatsoft.org/index.php/jss/article/view/v059i10/v59i10.pdf>

⁴ Se «The grammar of graphics» av Wilkinson (2005) <https://www.springer.com/gp/book/9780387245447>

⁵ For en oversikt, se Task View: <https://cran.r-project.org/web/views/Econometrics.html> og f.eks. online bøker som denne: <https://www.econometrics-with-r.org/index.html>

herunder kluster- og sekvensanalyse, tidsserieanalyse og bayesiansk statistikk, SEM, maskinlæring, kvantitativ tekstanalyse, webscraping⁶, og kobling mot API'er.

R også er en fullt utbygd GIS software egnet for avansert geografisk analyse, både for datahåndtering, grafikk og geografisk modellering. Slik sett vil R langt på vei matche spesialisert software som ArchGIS, og vil være særlig relevant for samfunnsgeografene og andre som jobber med romlige problemstillinger (eller av andre grunner bruker kart).

Pedagogiske ressurser

Det finnes gode lærebøker i kvantitative metoder på alle nivåer som bruker R/Tidyverse, men på SOSGEO1120 bruker vi ikke software-spesifikk lærebok for tiden. Men god støttelitteratur er altså lett tilgjengelig.

Det er høsten 2019 et pedagogisk samarbeid mellom metodeundervisere ved STV, ØI og PSY med bruk av R. Det lages interaktive kurs basert på R-pakken Swirl som gir studentene automatiske tilbakemeldinger på bl.a. programmeringsoppgaver.⁷ Dette skal implementeres i SOS2900 fra våren 2020 og kommer også til benyttes på kurs ved SVT, ØI og PSY. Det skal være helt overkommelig for emneansvarlig å bygge ut med tilsvarende oppgavemoduler for SOSGEO1120 ved behov.

I en overgangsperiode vil det kunne være noe utfordringer med å sikre gode seminarledere som behersker R. Men vi har flere både faste og midlertidige i staben som håndterer R godt. Med et godt forberedt opplegg burde det ikke være vanskelig for andre å lære seg tilstrekkelig R til å undervise på dette nivået. Utfordringene vil neppe være større enn det som kan oppstå av andre grunner i dag.

Konsekvenser for SOS4020

Det er nok en fordel om det brukes samme software på SOSGEO1120 som på SOS4020, men det er også mulig å ha forskjellig software på disse kursene. I dag har vi studenter som har bachelor fra andre steder som har brukt SPSS eller i liten grad software overhodet. Denne utfordringen vil vi altså fortsatt ha. Å ha ulik software er gjennomførbart, men ikke optimalt.

Det tar noe tid før de som tar SOSGEO1120 i 2021 begynner på MA slik at det behøver i så fall ikke skje før i 2022 eller 2023. Det vil være viktig å høste erfaring med bruk av R på SOSGEO1120 før det innføres på SOS4020.

Jeg vil foretrekke å innføre R på SOSGEO1120 våren 2021, og evaluere både det året og våren 2022. Da kan R innføres på SOS4020 høsten 2022.

Et alternativ kan være å gjøre som statsvitenskap har gjort over lang tid: dele inn kurset i SPSS og R, som hos dem innebar en implisitt nivåinndeling. For SOSGEO1120 er det mulig å kjøre dobbelt løp i en overgangsperiode med STATA og R. Da må det organiseres på seminargruppenivå.

⁶ Siste versjon av STATA har dette via integrasjon mot Python fremfor egen STATA-funksjonalitet.

⁷ <https://swirlstats.com/>