Notkun R við stofnmat: Verklag og pakkar

Árni Magnússon

30. ágúst 2010

1 Inngangur

1.1 Tölvunotkun í stofnmati

Reiknivinna við stofnmatsvinnu byggir aðallega á þremur almennum forritum: AD Model Builder, Excel/OpenOffice og R/S-Plus. Svo eru notuð sérhæfð forrit til að vinna aldurs-lengdar lykla, keyra XSA og slíkt, auk skeljaforrita til að sjálfvirkja ferli og sækja gögn úr gagnagrunni.

1.2 Stofnmatsvinna í R

Reiknivinna í R er framkvæmd með föllum/skriftum og gögnum. Föllin koma úr þremur áttum: kjarnaföll (fylgja með R), pakkaföll (sótt í pakka) og notendaföll. Reikniferli má geyma sem skriftu, en stundum getur verið betra að færa kóða úr löngum skriftum yfir í styttri notendaföll. Notendaföll eru hreinlegri og minnka líkur á mistökum, auk þess sem vinnuferlið verður skýrara og auðveldara að endurnýta í skyldum verkefnum, t.d. ári síðar eða fyrir aðrar tegundir.

Notendaföllum má dreifa á mismunandi vegu til að samnýta þau með vinnufélögum. Pakkar eru vandaðasti dreifingarmátinn, með hjálparsíðum, sjálfvirkum villuprófum og dreifingarkerfi sem sér til þess að notendur fái nýjustu útgáfu pakkans. Með útgáfukerfi geta vinnufélagar unnið saman að pakkagerð, viðhaldi og betrumbótum.

Gögn í stofnmatsvinnu eru oftast gagnarammar, en einnig er notuð önnur gagnaform, s.s. vektorar, fylki og listar. Með því að dreifa gagnasettum á pakkaformi eru gögnin handhæg, allir með sömu útgáfu og auðvelt að endurtaka útreikninga fyrri ára.

Mikill fjöldi skráa verður oft til í stofnúttektum og svipuðum verkefnum. Flestar skrár verða til meðan á útreikningum stendur, en lokaniðurstöður í skýrslu byggja yfirleitt aðeins á einni eða nokkrum lokakeyrslum. Það getur verið gagnlegt að geyma alla útreikninga í einni yfirmöppu og allt tengt skýrsluvinnunni í annarri yfirmöppu. Annað hvort er þá afrit af lokakeyrslu(m) geymt í skýrslumöppunni, eða frágangur á útreikningamöppunni það hreinlegur að ekki fari á milli mála hvar lokakeyrslan er. Einfaldar útskýringar má geyma í readme.txt, ætlaðar þeim sem vilja endurtaka eða endurnýta útreikningana.

R nýtist oft samhliða öðrum forritum. Þannig má forvinna inntaksgögn og eftirvinna úttaksniðurstöður stofnlíkana eins og ADCAM og Gadget.

1.3 Flutningur úr S-Plus í R

R kom fyrst út 1995 og í kringum 2002 er það orðið skilvirkara og betra umhverfi en S-Plus. Þetta endurspeglast í því að áhrifamiklir notendur á borð við John Chambers (upphaflegur höfundur S) og Brian Ripley (kennslubókahöfundur) gerast innstu koppar í R þróunarhópnum og nota það í sinni rannsóknavinnu í stað S-Plus. Mesti munurinn á R og S-Plus í dag er pakkaflóran, notendaviðmótið, fjöldi notenda og að sjálfsögðu verðið. Vörumerkið S-Plus hefur farið kaupum og sölum síðustu ár, þróun hefur því sem næst stöðvast og nú vantar í það nýlegar tölfræðiaðferðir sem tengjast stofnmati.

Stofnmatsvinna á Hafró hefur smám saman færst úr S-Plus í R og talsverð umræða um þessa þróun fór fram í nóvember 2009. Þá minntist Gunnar Örvarsson á að núgildandi S-Plus leyfi rennur út 1. mars 2011, svo e.t.v. væri áhugi á að færa alla stofnmatsvinnu yfir í R fyrir þann tíma.

Í framhaldi af umræðunni um áramótin 2009/2010 hélt Árni eins dags námskeið um uppsetningu og verklag í R fyrir lengra komna og vann síðan með Gunnari Örvars að staðlaðri R uppsetningu á Hafróvélum (Tafla 1), sem tölvusvið vinnur nú eftir.

Tafla 1. Stöðluð R uppsetning á Hafróvélum.

	Windows	Linux	Umhverfisbreyta
Forritsmappa	C:/Program Files/R/R-2.11.1/bin	/usr/lib64/R/bin	
Kjarnapakkar	C:/Program Files/R/R-2.11.1/library	/usr/lib64/R/library	
Hafrópakkar	C:/Program Files/R/site	/usr/local/lib/R/site/2.11/x86_64/library	R_LIBS_SITE
Notendapakkar	C:/Program Files/R/user	~/r/x86_64/library	R_LIBS_USER

Höskuldur hefur unnið mikið verk í yfirfærslu Hafró-pakkanna 'geo', 'fjolst' og 'Logbooks' úr S-Plus í R, eins og lýst er hér að neðan.

2 Pakkar sem tengjast stofnmati

2.1 geo, fjolst, Logbooks

Þrír pakkar sem Höskuldur hefur smíðað og haldið við: 'geo' snýr að kortagerð, 'fjolst' inniheldur gögn úr stofnmælingum og 'Logbooks' inniheldur gögn úr afladagbókum. Höskuldur hefur unnið að yfirfærslu úr S-Plus í R, þar sem 'geo' er kominn í almenna notkun í R, en 'fjolst' og 'Logbooks' eru á prófunarstigi í R.

2.2 SFunctions, SData, Dypi

Þrír pakkar sem voru smíðaðir á Hafró, en hafa ekki verið fluttir úr S-Plus yfir í R. 'SFunctions' er safn fjölbreyttra falla, s.s. extract.cohort(), find.ship(), fjperstod(), get.alk(), get.station(), grisjun(), plot.stations(), ReadBayesfile(), reit.smb(), sonde.plot.station(), summary.ALK(), og update.ALK(). 'SData' er safn fjölbreyttra gagna, s.s. eyjar, fishnames, gbdypif.100, gbdypif.200, glaciers, grd.smb, grunnlina, isl.month, landedcatch, lods, rivers, skipaskra, torskur.marsrall.visit og ysa.marsrall.visit. 'Dypi' er safn falla og gagna sem tengjast dýptarlínum, s.s. barents.sea, dypi.calc(), dypi.grd, dypi.std, gbdypi.100, gbdypi.200, og plot.depth().

2.3 ROracle, ROracleUI

Tveir pakkar sem vinna saman til að sækja gögn beint úr Oracle gagnagrunni inn í R, með hraðvirkari og áreiðanlegri beintengingu heldur en fyrri aðferðir. Höfundur 'ROracle' skrifaði hann eingöngu fyrir Linux, en Gunnari Örvars tókst nýlega að þýða hann fyrir Windows. Árni smíðaði 'ROracleUI' pakkann sem skilvirkt og þægilegt viðmót til að vinna með Oracle töflur og fyrirspurnir.

2.4 FLR, scape, scapeMCMC, R-Gadget

'FLR' (Fisheries Library in R) er safn pakka fyrir stofnmat, með áherslu á aflareglurófun. Árni sótti 5 daga námskeið hjá aðalhöfundum 'FLR' í vor og hans mat er að 'FLR' sé í vinnslu og ekki mjög notadrjúgt fyrir stofnmat á Hafró að svo stöddu. Markmið 'FLR' er að útbúa gegnsætt og sveigjanlegt umhverfi (ekki svartur kassi) sem gæti flýtt fyrir ICES prófunum á aflareglum, fækkað villum og auðveldað starfsbræðrum að skilja hvern annan. Eins og staðan er í dag er erfitt fyrir ICES yfirlesara (reviewers) að dæma almennilega hvort ákveðin aflarefluprófun sé vel eða illa gerð, á þeim takmarkaða tíma sem yfirlesurum er ætlað.

Pakkarnir 'scape' og 'scapeMCMC' voru smíðaðir af Árna til að vinna með Coleraine stofnlíkanið. Raunar er 'scapeMCMC' ekki sértækur fyrir Coleraine, heldur almennur pakki til að vinna með niðurstöður úr MCMC óvissugreiningu.

R-Gadget er pakki sem Bjarki Elvarsson er að smíða, en hann er doktorsnemi hjá Gunnari Stefáns. Hann er reyndur forritari og starfar náið með Gunnari, svo ætla má að pakkinn gæti orðið vinsælt notendaviðmót fyrir Gadget. Gadget notendaumhverfið hefur náð þeim stöðugleika að nú er tímabært og raunhæft að smíða skilvirkt og þægilegt notendaviðmót í R.

2.5 gmt, PBSmapping

Pakkarnir 'gmt' (eftir Árna) og 'PBSmapping' (eftir Jon Schnute) eru fyrir kortagerð og voru smíðaðir með fiskifræði í huga, líkt og 'geo' pakkinn. Nákvæmur samanburður liggur ekki fyrir, en líklegt er að 'geo' pakkinn haldi áfram að vera vinsælasti pakkinn fyrir kortagerð á Hafró, þó ekki væri nema vegna þess að starfsmenn Hafró hafa lært á 'geo' en ekki hina tvo. Áhugasamir geta kannað hvort 'gmt' eða 'PBSmapping' henti betur fyrir ákveðna gerð korta.

3 Umræðupunktar

3.1 Verklag

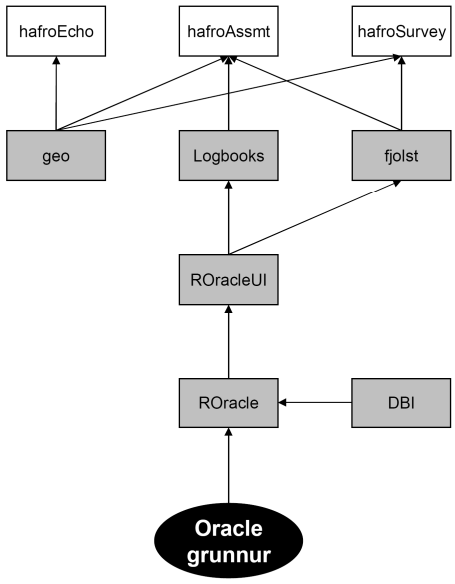
Ræða um pakka og notendaföll sem eru notuð við stofnmatsvinnu, svo allir kunni á þeim skil: geo/fjolst/Logbooks, ROracleUI, FLR/scape/scapeMCMC og gmt/PBSmapping. Í hvaða pakka er landaður afli? Notendaföll sem hafa verið notuð í stofnmatsvinnu, t.d. til að teikna myndir.

Ræða um æskilegan frágang á verkefnum, þannig að sami starfsmaður eða annar starfsmaður geti gengið að því síðar. Skriftur vs. notendaföll, skráarform falla (R vs. RData) og gagna (R vs. RData, auk csv og txt fyrir gagnaramma), readme.txt, aðskildar möppur fyrir útreikninga og skýrslu.

3.2 Pakkar

Ræða um þýðingarferli 'geo', 'fjolst' og 'Logbooks' yfir í R og almennt ástand pakkanna. Einnig hugsanlegt útgáfukerfi og dreifingarkerfi, SVN, innanhús- og utanhúsdreifingu og R-Forge.

Ræða um tengsl núverandi pakka, t.d. hvernig ROracleUI byggir á ROracle og aðrir pakkar geta notað sql() og önnur ROracleUI föll, í stað þess að sambærileg föll sé að finna í mörgum pökkum (Mynd 1).



Mynd 1. Tengsl R pakka. Gráir kassar eru núverandi pakkar, hvítir kassar eru hugsanlegir pakkar.

Ræða um samnýtingu á notendaföllum fyrir algenga útreikninga og myndagerð. Hugsanlega mætti útbúa pakka sem gætu heitið 'hafroAssmt', 'hafroSurvey', 'hafroEcho', eða slíkt. Margir gætu haft skrifaðgang til að viðhalda og betrumbæta, en gæðakröfur gætu verið minni en í undirliggjandi pökkum 'geo', 'fjolst' og 'Logbooks'.

Ræða um gömlu S-Plus pakkana 'SFunctions', 'SData' og 'Dypi', hugsanlega skörun í virkni eða úreldingu.

Ræða um uppsetningu S-Plus minjasafns, sem hægt er að leita í og endurnýta varahluti.