Flóðagreining með bayesískri tölfræði - Lokaskýrsla fyrir Vegargerðina

Fjalarr Páll Mánason

12. maí 2010

Verkefnið var styrkt af Vegagerðinni að upphæð 500 þúsund kr.
Verkefnastjóri: Birgir Hrafnkelsson, fræðimaður við HÍ
Nemandi sem vinnur að verkefninu: Fjalarr Páll Mánason, MS nemi í verkfræði við HÍ
Meðleiðbeinandi: Sigurður Magnús Garðarsson, prófessor við HÍ
Aðrir samstarfsmenn: Árni Snorrason, forstjóri Veðurstofu Íslands. Óðinn Þórarinsson, Veðurstofu Íslands. Guðrún Þóra Garðarsdóttir, Vegagerðinni

1 Inngangur

Þátttakendur í verkefninu er taldir hér. Fjalarr Páll Mánason meistaranemi í verkfræði við HÍ, vann að gerð forrita, fyrirlestra og veggspjalds. Birgir Hrafnkelsson fræðimaður við Raunvísindasfofnun HÍ, er aðalleiðbeinandi Fjalars fyrir meistaraverkefni hans. Sigurður Magnús Garðarsson prófessor við HÍ, er meðleiðbeinandi Fjalars. Árni Snorrason og Óðinn Þórarinsson eru samstarfsaðilar frá Veðurstofu Íslands og Guðrún Þóra Garðarsdóttir er samstarfsaðili frá Vegagerðinni.
Fyrstu vikur verkefnsins fóru í það að verða sér út um heimildir og greinar sem fjalla um bayesíska tölfræði og flóðagreiningu. Þetta fór ekki endilega saman í öllum heimildum heldur notuðu flestar heimildir sem fjalla um flóðagreiningu annars konar aðferðir heldur en bayesíska tölfræði til að leysa vandamálið. Skoðað var sérstaklega það sem hafði verið gert áður og hvaða sérstöðu verkefni myndi hafa. Kom í ljós að sú aðferð sem notast er við í þessu verkefni hefur ekki verið notuð áður á þann hátt og við nálgumst það.
Það sem við leggjum upp með er að notast við B-spline rennlislykla sem byggja á bayesísku nálguninni til að varpa vatnshæð yfir í rennsli. Það er alltaf óvissa í þeirri vörpun en bayesíska nálgunin gefur færi á að taka þá óvissu með í reikningana þegar kemur að flóðagreiningunni. Þetta er eitthvað sem ekki hefur gert áður að okkur meðvituðum og er þetta því ný aðferð við að greina flóð.
Við flóðagreiningu hingað til, hafa tímaraðir af vatnsrennslisglidum verið notaðar til að framkvæma flóðagreiningu. Þær tímaraðir hafa verið notaðar sem fastar, þ.e. ekki hefur verið gert ráð fyrir neinni óvissu í þeim. Staðreyndin er hins vegar sú að vatnsrennslisgildin eru fundin út frá vörpun frá vatnshæðargildum. Vatnshæð er mæld í sífellu með vatnshæðarmælum sem við gefum okkur að séu fullkomnlega réttir, en vörpuninni frá vatnshæð yfir í vatnsrennsli fylgir óvissa sem við tökum inni í reikningana fyrir flóðagreiningu. Það er þessi óvissa sem gerir þetta verkefni frábrugðið öðrum svipuðum verkefnum.
Flóðgreiningin sem slík er gerð með svokallaðri hágildisgreiningu (e. extreme value analysis). Tvær aðferðir hágildigreiningu eru notaðar. Annars vegar er notast við árleg hágildi (e. block extrema model) og hins vegar er notast við gildi yfir ákveðnum þröskuldi (e. threshold model).

2 Árleg hágildi

Aðferðin sem notast við árlegu hágildin er þannig að árlegu hágildunum er safnað saman og athugað er hvernig árlega hágildisdreifingin (*e. generalized extreme value distribution (GEV)*) passar við þau gildi. Þessi dreifing hefur þrjá stika. Staðsetningarstika, μ, skölunarstika, σ, og lögunarstika, ξ. Þéttleika almennu hágildisdreifingarinnar má sjá hér að neðan.

$$p(y_i|\mu,\sigma,\xi) = \frac{1}{\sigma}\Big[1+\xi\Big(\frac{y_i-\mu}{\sigma}\Big)\Big]^{-(\frac{1}{\xi})-1} \times exp\Big\{-\Big[1+\xi\Big(\frac{y_i-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}}\Big\} \tag{1}$$

Stikar almennu hágildisdreifingarinn hafa allir sinn tilgang. Staðsetningarstikinn (μ) gefur til kynna miðgildi drefingarinnar og skölunarstikinn (σ) segir til um dreifni hennar. Lögunarstikinn (ξ) hefur stóru hlutverki að gegna því hann segir til um þykktina á hala dreifingarinnar (sjá mynd 1). Ef lögunarstikinn er neikvæður gefur það til kynna að hali dreifingarinnar sé endanlegur. Það þýðir í okkar tilfelli að það sé takmarkað hvað flóð getur orðið stórt. Ef lögunarstikinn er stærri eða jafn núlli er hali dreifingarinnar ótakmarkaður og þykkari eftir því sem lögunarstikinn stækkar. Það þýðir í okkar tilfelli að möguleg stærð flóða er ótakmörkuð, og eftir því sem halinn þykkist verða líkur á stórflóðum meiri.
Þetta sést glögglega ef skoðuð eru endurkomutíma gröf flóða. Endurkomutíma gröfin sýna væntigildi hámarks vatnshæðar sem fall af endurkomutíma. Eftir því sem lengri tími líður eykst væntigildi hámarksflóðs á í tímabilinu. Áhugavert er að skoða endurkomutíma m.t.t. mismunandi gilda á lögunarstikanum. Ef lögunarstikinn er núll er aukning á stærð flóða línuleg m.t.t. logra af endurkomutíma. Ef lögunarstikinn er stærri en núll er aukningin veldisvaxandi m.t.t. logra af endurkomutíma. En fyrir lögunarstika minni en núll er aukningin takmarkandi þannig að hámarksflóð fara vissulega stækkandi með hækkandi endurkomutíma, en þau falla að aðfellu sem takmarkar mögulega stærð flóða.
Mynd 1 hér að neðan sýnir hvernig breyting á lögunarstikanum, ξ, hefur áhrif á þéttleikafall almennu hágildisdreifingarinnar sem og endurkomutíma gröf.

3 Þröskuldslíkan

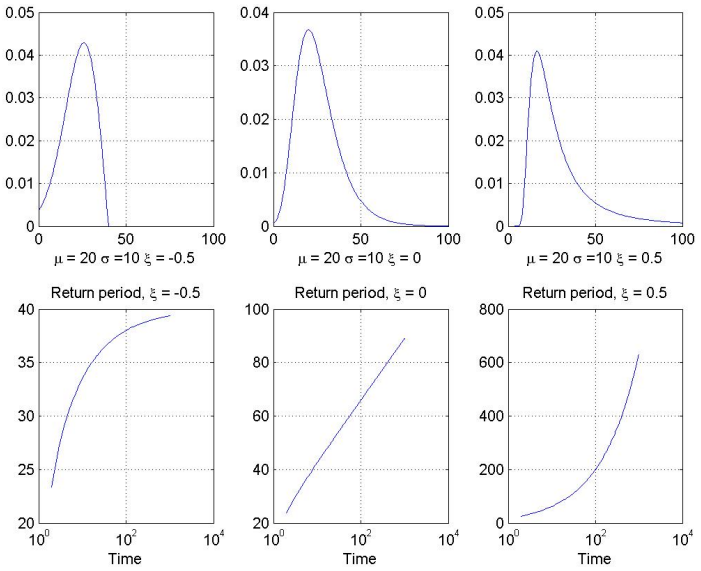
Í þröskuldslíkaninu er notast við almennu Pareto dreifinguna (*e. generelized pareto distribution (GP)*). Gildi yfir ákveðnum þröskuldi, u, eru tekin saman og þröskuldsgildið, u, er dregið frá þeim. Þá höfum við eftir safn af umframgildum ($x = (y_i - u)$, fyrir $y_i > u$) og athugað er hvernig þau passa við almennu Pareto dreifinguna. Almenna Pareto dreifingin hefur tvo stika. Skölunarstika, σ_t, og lögunarstika, ξ. Þéttleikafall almennu Pareto dreifingarinnar má sjá hér að neðan

$$p(x|\sigma_t,\xi) = \frac{1}{\sigma_t}\Big[1+\frac{\xi x}{\sigma_t}\Big]^{-\frac{1}{\xi}-1} \tag{2}$$

Sterkt samband er á milli stikanna í almennu hágildisdreifingunni og almennu Pareto dreifingunni. Lögunarstikinn er fræðilega sá sami í báðum líkönum en samband σ_t og stikanna úr GEV dreifingunni er eftirfarandi

$$\sigma_t = \sigma + \xi(u-\mu)$$

Þröskuldslíkanið er búið til á eftirfarandi hátt: Þröskuldsgildi er valið með aðferðum sem lýst verður betur hér á eftir. Öll gildi sem eru hærri en þröskuldurinn eru tekin til athugunar í gerð líkansins. Til



Mynd 1: Þéttileikaföll og endurkomutíma gröf fyrir GEV dreifingu með mismunandi gildi á lögunarstika

þess að gildin verði ekki háð hvor öðrum er farið í það ferli að sía gildin yfir þröskuldinum. Tilgangur líkansins er að nota þröskuldinn til að finna atburði sem notast má við í hágildisgreiningu. Þar sem við erum að vinna með daglegar tímaraðir getur komið upp það vandamál að mörg gildi sama atburðar fari yfir þröskuldinn og myndu því vera notað í líkanið ef ekkert væri að gert. Þetta myndi hafa í för með sér að óæskileg fylgni væri milli sumra gildanna sem notuð væri í líkaninu. Það sem við viljum er hins vegar aðeins stærsta gildi hvers atburðar. Þá kemur upp það vandamál að ákveða hvað teljist til eins atburðar. Það getur oft á tíðum verið mjög óljóst hvort flóð einn daginn teljist til sama atburðar og flóðið sem var t.d. 5 dögum áður.
Til að leysa þetta vandamál ákváðum við að búa til svokalla greiðu sem myndi skipta þeim gildum sem færu yfir þröskuldinn upp í atburði. Síðan finnum við hæsta gildi hvers atburðar og notuðumst aðeins við það í þröskuldslíkaninu. Hugmyndin að greiðunni er eftirfarandi. Ef stærð greiðunnar er 1, þá myndu t.d. tvö gildi sem eru hærri en þröskuldurinn teljast til sama atburðar ef ekki væri nema einn dagur sem liði á milli þessara tveggja gilda. Almennt ritað myndi þetta líta eftirfarandi út.
Ef stærð greiðunnar er n, þyrftu að líða $n+1$ dagar á milli tveggja gilda sem fara yfir þröskuldinn til þess að þessi gildi myndu teljast til sitthvors atburðarins.

3.1 Val á þröskuldi

Þegar þröskuldurinn er valinn eru nokkur atriði hafð að leiðarljósi.

3.1.1 1: Mean residual life plot (MRLP)

Samkvæmt hágildisfræðum gildir eftirfarandi. Ef Y hefur GP dreifingu gildir,

$$E(Y) = \frac{\sigma_t}{1-\xi}$$

Því gildir, ef X er tímaröð sem til skoðunar er og þröskuldsgildi er u_0,

$$E(X - u_0 | X > u_0) = \frac{\sigma_{u_0}}{1 - \xi}$$

Ef $(X - u_0)$ fyrir $X > u_0$ fylgir GP dreifingu hljóta hljóta öll þröskuldsgildi hærri en u_0 einnig að hafa það í för með sér að skilyrt umframgildi fylgi GP dreifingu. Þannig, ef $u > u_0$ gildir einnig

$$E(X - u_0 | X > u_0) = \frac{\sigma_u}{1 - \xi}$$

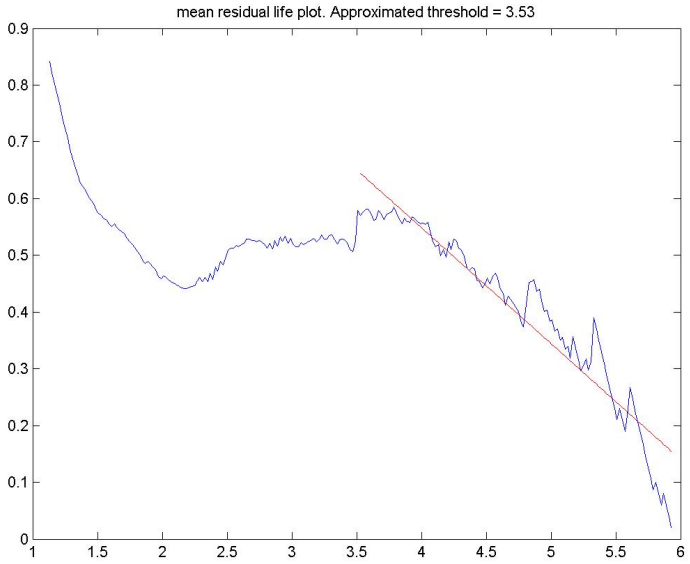
En þar sem skýrt samband er á milli skölunartika fæst

$$E(X - u | X > u) = \frac{\sigma_{u_0} + \xi u}{1 - \xi}$$

Þar með sést að ef u_0 er minnsti mögulegi þröskuldur þannig að gildin til skoðunar fylgi GP dreifingu þá er væntigildið að ofan línulegt m.t.t. u fyrir öll u stærri en minnsta mögulega þröskuldsgildið u_0. Með þessum rökum er farið í það að rýna í svokölluð *mean residual life plot (MRLP)* og ákvarða þröskuldsgildið út frá því hvenær MRLP verður línulegt m.t.t. u. Eftirfarandi er plottað upp til að fá MRLP.

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (y_(i) - u) \right) : u < y_{max} \right\} \tag{3}$$

Þar sem n_u er fjöldi gilda sem eru stærri en þröskuldsgildið.
Dæmi um MRLP má sjá á mynd 2.



Mynd 2: Mean Residual Life plot (MRLP) með viðbættri línu til að lýsa hinu línulega sambandi

Eins og sést á mynd 2 myndi hæfilegt val á þröskuldi fyrir þessa tilteknu á vera u.þ.b. $u = 3.5$.

3.1.2 2: Stikaaðferðin

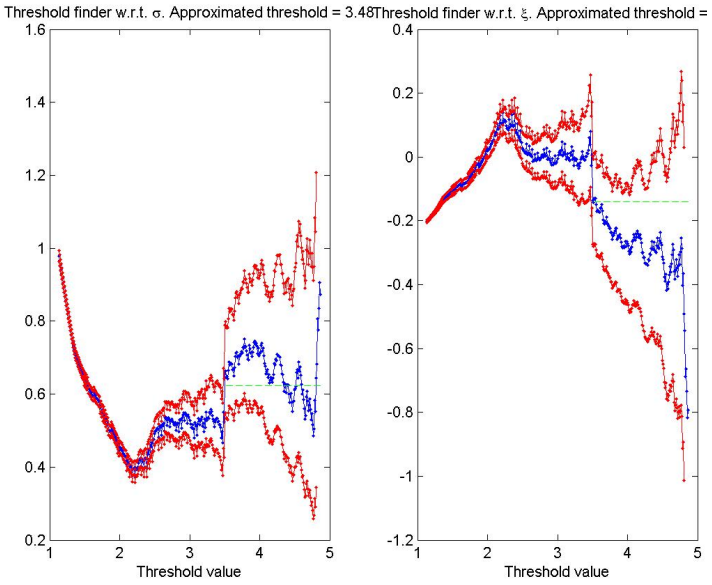
Önnur leið við val á þröskuldi er að notfæra sér aftur sambandið milli skölunar- og lögunarstikana m.v. breytilegt þröskuldsgildi. Minnsta mögulega þröskuldsgildið u_0 er minnsta gildið þar sem umframgildin fylgja almennu Pareto dreifingunni. Fræðilega séð er lögunarstikinn sá sami fyrir öll þröskuldsgildi stærri eða jöfn hinu minnsta mögulega þröskuldsgildi.
Eftirfarandi samband gildir milli stika almennu Pareto dreifingarinnar m.v. misnunandi þröskuldsgildi

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0)$$

Þetta má umrita á eftirfarandi hátt

$$\sigma^* \equiv \sigma_{u_0} - \xi u_0 = \sigma_u - \xi u$$

Jafnan á undan segir okkur að σ^* er fasti fyrir öll þröskuldgildi sem eru stærri eða jöfn minnsta mögulega þröskuldgidinu u_0. Þannig má með því að skoða graf af σ^* m.t.t. þröskuldsgildis áætla hæfilegan þröskuld með því að athuga hvenær grafið verður orðið að fasta m.t.t. u.
Það sama má gera fyrir lögunarstikann, þar sem hann á að vera fasti fyrir öll þröskuldsgildi stærri eða jöfn minnsta mögulega þröskuldsgildi.



Mynd 3: σ^* og ξ ásamt 95% öryggisbili m.v. breytilegt þröskuldsgildi

Eins og sést á mynd 3 myndi hæfilegt val á þröskuldi fyrir þessa tilteknu á vera u.þ.b. $u = 3.5$.

3.1.3 3: Almennar aðferðir

Ásamt því að notfæra sér viðmiðunarreglurnar hér að ofan, sem byggja á fræðum einum saman, skoðum við einnig nokkrar þumalputtareglur. Við reynum að haga þröskuldinum þannig að fjöld gilda sem notaður verður í líkanið verði á bilinu $(1 - 2, 5)*$árafjölda í gagnasafninu sem notað er.
Einnig skoðum við sérstaklega hvert þröskuldsgildið myndi vera ef við myndum haga því þannig að fjöldi mælinga yfir þröskuldi væri jafn árafjöldanum. Þetta er aðferð sem notast hefur verið við hjá

vatnamælingum og er byggð á reynslu og þægindum þó fræðileg gildi á þröskuldi fengin úr aðferðunum (1) og (2) gætu verið frábrugðin.

4 Mismunandi útfærslur á verkefninu

Í upphafi var ráðist í verkefnið á þann hátt að notast var við tímaraðir sem lýstu daglegu hágildi á vatnshæð fyrir ár. Hágildisgreining var gerð á þessar tímaraðir, bæði með almennu hágildisdreifingunni (hágildi innan fastra tímabila (e. block extrema)) og almennu Pareto dreifingunni (hágildi yfir þröskuldi (e. threshold)). Stikarnir fyrir almennu hágildisdreifinguna og almennu Pareto dreifinguna voru fundir með bayesískum tölfræðiaðferðum og vatnshæð var teiknuð upp sem fall af endurkomutíma. Því næst var þessum endurkomugröfum fyrir vatnhæð varpað yfir í samkonar endurkomugröf fyrir vatnsrennsli. Bayesíska tölfræðin kemur sér mjög vel í þessum aðgerðum þar sem þægilegt er að halda utan um óvissuna í rennslislyklinum sem og óvissuna sem kemur út frá hágildisgreiningunni. Vandamálið með þessa aðferð er hins vegar að árgögn eru sjaldnast geymd á formi vatnhæðar, heldur sem vatnsrennsli. Þau vatnsrennslisgögn sem geymd eru hafa verið búin til með rennslislyklum en ekki er haldið upp á óvissuna í vörpunni frá vatnhæð yfir í rennsli. Þess vegna var nauðsynlegt að nálgast vandamálið á annan hátt. Ákveðið var að taka inn vatnsrennslisgögn og varpa þeim tilbaka yfir í vatnshæð með miðgildis B-spline-rennslislykli. Þessum vatnhæðargildum er svo varpað aftur yfir í rennslisgildi en nú er allur lykilllin notaður en ekki einungis miðgildi hans. Með þessu móti fæst mat á óvissu í rennslisgögnunum. Framkvæmd verður hágildisgreining á þessum gildum með aðferðum sem annars vegar byggja á hágildum innan fastra tímabila og hins vegar aðferðum sem byggja á hágildum yfir þröskuldi.

5 Framhaldið

Gerð seinna líkansins er á góðri leið með að klárast. Þegar því líkur verður líkanið notað við að flóðagreina þær ár sem gögn fást fyrir. Eftir það tekur við úrvinnsla á niðurstöðum. Þetta mun vera tekið sama í meistararitgerð og stefnt er að því að rita upp úr henni fræðigrein. Vonast er til að sérstæða verkefnisins, varðandi óvissureikninga annars vegar og notkun B-spline rennslislykla við flóðagreiningu, hins vegar, muni gera það möglegt að fá grein um það birta í ritrýndu vísindatímariti. Unnið er að því að klára verkefnið í sumar og verja það í haust.