Bírálat
Honti Márk: Bayesi módszerek a vízi környezet modellezésében

MTA doktori disszertációjáról

Mint a szerző a mottóban is írja, nem komplex modellekkel foglalkozik, hanem a valós adatokat, feltételeket és elvárásokat mindinkább kielégítő eszközöket használ a vízi környezet modellezéséhez. Ehhez a Bayes módszertant választja, ami ugyan Thomas Bayes XVIII. századi eredményéhez nyúlik vissza, de mapjainkban is újabb és újabb továbbfejlesztett változataival találkozunk. Maga a Bayes tétel egy nagyon egyszerű megfigyelésen alapul a feltételes valószínűség kiszámítására a feltétel és következményét jelentő esemény megfordítására vonatkozóan. Ez akkor érdekes, ha ún. teljes eseményrendszerünk van, és az eseményrendszer tagjainak valószínűségére van egy ún. prior eloszlásunk. Ez az 'a priori'-nak is nevezett eloszlás tapasztalás előtti, azaz a kísérlet elvégzése előtt is rendelkezésünkre áll (előzetes megfigyelések, információk vagy szubjektív elképzelések alapján). A kísérlet elvégzése után a tapasztalatot figyelembe véve számítható ki az 'a posteriori' (tapasztalat utáni) eloszlás. A prior eloszlás szubjektív választása körül filozófikus viták bontakoztak ki, ui. Bayes eredeti elképzelése egyenletes eloszlású priorra vonatkozott. Később Laplace ezt tetszőleges diszkrét eloszlásra cserélte. Ennél is később, a XX. század első felében, mikor a modern statisztikát lényegében megalapította az angolszász iskola R. A. Fisher vezetésével (amihez a svéd H. Cramér és az indiai C. R. Rao is hozzájárultak), a Bayes tétel alapján a Bayes módszert, mint becslési módszert vezették be. Ennél már az eloszlás paraméterének prior eloszlásából határozzák meg a paraméter posterior eloszlását a konkrét megfigyelés (tapasztalat) alapján. A posterior valószínűségek maximalizálása a paraméterben fontos feladat a disszertációban, azonaban a jelölt nem tárgyalja a Bayes becslések négyzetes rizikót minimalizáló tulajdonságát, amikoris a paraméter feltételes várható értékét kell venni a feltételben álló tapasztalati megfigyelésre, és így a becslés az adott megfigyelés függvényévé, azaz statisztikává válik. Formálisan ez a feltételes várható érték vevés csak speciális elméleti háttéreloszlások esetén végezhető el, azonban a gyakorlatban is vannak numerikus statisztikai algoritmusok a feltételes várható érték számolására mintából (pl. Breiman-Fiedler ACE algoritmusa, Györfi László nemparaméteres regressziója vagy Dempster–Laird–Rubin EM algoritmusa). Kicsit hiányoltam ezeket a disszertációból. Ugyanakkor Jelölt bizonyítja jártasságát abban, hogy a vízügyi adatok vonatkozásában hogyan lehet optimális stratégiákat kidolgozni. Visszatérve a Bayes módszer aktualitására, azt sikeresen alkalmazták a XXI. század elején az Air France Rio de Janeiróból

Párizsba tartó járatának lezuhanása után a fekete doboz megtalálására az Atlanti Óceánban. Néhány év sikertelen keresés elteltével ui. a francia hatóságok az US Navy egy (hölgy) szakértőjéhez fordultak, aki historikus időjárási és áramlástani adatokat bevive a priorba, posterior valószínűségeket tudott adni a fekete doboz megtalálására az óceánban behatárolt (de azért elég nagy) terület egyes régióiban való megtalálhatóságra. Kisebb módosítások után a keresés sikerre is vezetett a legnagyobb posterior valószínűségű régióban.

A dolgozat négy fejezetre oszlik, melyek alapján hat tézist fogalmaz meg a jelölt.

A Bevezetésben Jelölt irodalmi áttekintést ad a hagyományos (nem bayesi módszereken alapuló) modellezési gyakorlatról. A felismerhetetlenség és dekompozició által előállított műtermékeket saját esettanulmányokon mutatja be (Honti és Stumm, 2010; Honti és Istvánovics, 2019). A klasszikus modellkalibráció mellett, ahol a referencia-adatok $L(\theta, \mathbf{Y}_O)$ likelihoodját maximalizálják θ-ban a paramétertéren, a log-likelihood függvényen keresztül, Jelölt belátja, hogy Gauss eloszlás esetén a likelihood maximalizálása ekvivalens a hibatag legkisebb négyzetes minimalizálásával. (Itt a hiba az Y_O referencia adatok és az $Y_M(\theta)$ determinisztikus modell kimenet közt értendő.) Ismerteti a GLUE (Generalized Likelihood Uncertainty Estimation) eljárást (Beven és Binley, 1992), ami a Hornberger–Sper-féle érzékenységvizsgálat továbbfejlesztésének tekinthető. Ez azon alapul, hogy a kalibráció célfüggvény-értéke szempontjából egyenértékű (behavioural) paraméterkészletek nem feltétlenül korlátozódnak a paramétertér egy részére. Ezért a teljes paraméterteret lefedő Monte–Carlo mintázást végeznek. A vízi ökoszisztéma anyagcseréjét a limnológia kezdetei óta általában az oxigénben kifejezett nettó elsődleges termeléssel (Net Ecosystem Production, NEP) jellemzik. Ez a GPP (Gross Primary Production) és R (a teljes ökoszisztéma légzése) különbségeként időben kifejezhető. Ha a NEP pozitív, akkor az ökoszisztéma autotróf, ha negatív, akkor hetero- vagy disztróf. Jelölt állítása szerint a GPP és R napi (rövid távú) értékei közti lineáris korreláció nagy része a hibaterjedés műterméke, ami elfedi a napi kapcsolat valószinűsíthetően hiszteretikus jellegét (Honti és Istvánovics, 2019). A paraméterek és modellszerkezet felismerhetetlensége így a modellek alkalmazhatóságát hátráltatja.

Ezzel szemben a bayesi paraméterbecslés külső információk bevonásával javítja a paraméterek felismerhetőségét, ha az információ a kísérlettől független forrásból származik. A hagyományos kalibráció feltételezi, hogy a kalibrációban csak a modell hibája véletlen, a paraméterek nem valószínűségi változók; itt a tradicionális statisztikai következtetés csak a hibára vonatkozhat. Ezzel ellentétben, a bayesi statisztikán alapuló kalibráció a modell paramétereit és strukturális bizonytalanságát is bevonja a vizsgálatokba. A bayesi paraméterbecslés külső információk bevonásával javítja a paraméterek felismerhetőségét, de nem tudja az

összes felismerhetetlenségi problámát megoldani. A különböző modellszerkezetek strukturális egyenértékűsége csak akkor küszöbölhető ki, ha a kérdéses szerkezeti elemeket befolyásoló paraméterekre nagyon pontos információ áll rendelkezésre vagy több olyan rendszert tudnak egyszerre kalibrálni, ahol a paraméterek kifejeződése eltérő mértékű. Strukturális egyenértékűségre Jelölt példaként hozza fel az OECD 308-as kísérletet, melyben három modellváltozatot egyenértékűnek találtak. Az OECD 309-es kísérlet bevonása a kalibrációba viszont segített, mert pontosította az aerob lebomlás sebességét, és ennek alapján a 308-asbeli anaerob bomlás sebessége is kevésbé bizonytalanná vált (Honti és Fenner, 2015; Honti és mtsai, 2016).
A Bayes becslés alapegyenlete

$$\mathbb{P}(\theta|Y_O) = \frac{L(\theta, Y_O)\mathbb{P}(\theta)}{\int L(\theta, Y_O)\mathbb{P}(\theta)\, d\theta}$$

ahol $\mathbb{P}(\theta)$ a θ paraméter megfigyelési adatoktól független prior (azaz tapasztalás előtti), $\mathbb{P}(\theta|Y_O)$ pedig a megfigyelés (tapasztalás) utáni valószínűsége; $L(\theta, Y_O)$ a likelihood függvény, mely nem más, mint $f(Y_O|Y_M(\theta))$, azaz a mintaelemek adott modell és paraméter melletti sűrűségfüggvénye (folytonos esetben, a kevésbé tipikus diszkrét esetben pedig súlyfüggvényt kell használni). A számláló egy adott paraméterértékre vonatkozik, a nevezőben viszont a teljes paramétertéren kell integrálni (diszkrét esetben összegezni). Mivel a nevező már nem függ a paramétertől, a poszterior valószínűségek nagyságrendjét a számláló jellemzi. A kalibráció ezek után a posterior valószínűség maximalizálása θ-ban, amelyben a normáló tényező nem játszik szerepet, csak a prior és a likelihood egyensúlya. Ha a prior eloszlás egy pont körül koncentrálódik, akkor a posterior is egyetlen optimális paraméterkombinációt jelöl ki a többdimenziós paramétertérben. Viszont a prior dominanciája esetén a kalibrációnak nem sok értelme van, mert az adatokból a modell keveset tanul. Másfelől, egy nem informatív prior eloszlás (pl. egyenletes a paramétertéren) a hagyományos likelihood maximalizását eredményezi, ami a szokásos kalibrációhoz vezet.
A szisztematikus hibák bayesi leírása (Kennedy–O'Hagan hibamodell, 2001) nem alkalmazható, amikor a hibafolyamatot időszakos, de jelentős lökések érik a bemenő adatok hibái miatt, mert ekkor a szisztematikus hibák eloszlása időben változik. Erre a helyzetre fejlesztette ki Jelölt a zavart Orstein–Uhlenbeck folyamaton alapuló bayesi hibamodellt. Ez csapadék-lefolyás modelleknél egy időben képes a szerkezeti, bemenő és megfigyelési adathibák kezelésére (Honti és mtársai, 2013). Az új hibamodell jelentősége, hogy vele a csapadék-lefolyás modellek teljes bayesi kalibrációja és bizonytalanságvizsgálata az idősorok transzformációja nélkül is végrehajtható. A zavart Orstein–Uhlenbeck hibamodellhez Jelölt ki-

dolgozott egy, a modell korlátozott hosszúságú memóriáját kihasználó, kernel-alapú iteratív eljárást, mellyel hosszú idősorok likelihood számítása viszonylag kisméretű mátrixok invertálásával megoldható, szemben a régebbi gyakorlattal.
A tavi anyagforgalom modellezésében Jelölt bemutatta, hogy a rövidtávú idősor-előrejelzéseknél jó megoldást ad a bayesi tanulás és időben változó paraméterek alkalmazása (Honti és Istvánovics, 2016; Honti és mtsai, 2016; Istvánovics és Honti, 2017). Időben változó paramétereket ugyan régebben is alkalmaztak környezeti modellek kalibrációjára, de bayesi tanulás nélkül. Jelölt az alapegyenletet alkalmazza iteratív módon a mozgó időablakokban, ahol a megelőző iterációs lépés poszteriorja lesz a következő iterációs lépés priorja.
Az éghajlatváltozás hatásai hosszú távúak, ezeknél nincs értelme a bekövetkező események sorrendiségét vizsgálni, itt nem maga az idősor, hanem annak statisztikai eloszlása az előrejelzés tárgya. Jelölt kifejlesztett egy olyan közelítő likelihood függvényt, melyben az idősor-hibamodellhez hasonlóan állítható a bizonytalanság mértéke és az eloszlásra illesztés végrehajtható (Honti és mtsai, 2014). Jelölt továbbá bemutatta, hogy az eloszlásra illesztéssel a modellezett éghajlati peremfeltételek szisztematikus eltérései is sikeresen korrigálhatók.
Úgy gondolom, az értekezéssel Jelölt elérte célját: bemutatta, hogyan lehet a vízi környezet modellezésében a kalibrációt és a bizonytalanságvizsgálatot fejleszteni bayesi módszerekkel. Mivel univerzálisan érvényes (objektív) kalibráció nem létezik még egy adott témakörön belül sem, a modellek eredményei nem szükségszerűen összehasonlíthatók és általánosíthatók. Ugyanakkor a bayesi eljárások biztosítják a modellezési gyakorlatok reprodukálhatóságát, mivel a szubjektivitást okozó elemek explicit módon megjelennek benne, ezért egyértelműen dokumentálhatók. Ezzel jelölt egyfajta mesterséges intelligenciát épít, bár ezt nem mondja disszertációjában. Kicsit hiányolom a gépi tanulásra építő modern statisztikai algoritmusok (EM, ACE) említését és használatát. Erre vonatkoznak Jelöltnek feltett kérdéseim is. Ettől függetlenül a doktori mű tudományos eredményeit és a kapcsolódó publikációkat (többségükben vezető nemzetközi folyóiratokban jelentek meg, és többnyire Jelölt az első szerző) elegendőnek tartom az MTA doktora cím megszerzéséhez és a nyilvános védés kitűzését javaslom.

Budapest, 2022. november 29.

Dr. Bolla Marianna
Prof., az MTA doktora

A Jelöltnek feltenni kívánt kérdéseim:

1. A Bayes módszer szubjektivitása alkalmas-e mesterséges intelligencia építésére vízügyi adatokon?

2. A Bayes becslés feltételes várható érték képzése elvégezhető-e a gyakorlatban, ill. tervezik-e ehhez olyan modern statisztkai algoritmusok használatát, mint az EM (Expectation–Maximization hiányos adatokra) és ACE (Alternating Conditional Expectation) nemparaméteres regresszióra, kernel alapú simításokkal?