NYILATKOZAT

Név: Halász Kristóf

ELTE Természettudományi Kar, szak: Matematika BSc

NEPTUN azonosító: CJ49CH

Szakdolgozat címe:
Kaliforniai erdőtűz-kockázat modellezése

A **szakdolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2023.06.05.

Halász Kristóf

a hallgató aláírása

Eötvös Loránd Tudományegyetem

Természettudományi kar

Halász Kristóf

Kaliforniai erdőtűz-kockázat modellezése

Szakdolgozat

Matematika BSc

alkalmazott matematikus specializáció

Témavezető:

dr. Zempléni András

egyetemi docens

Valószínűségelméleti és Statisztika Tanszék

ELTE

EÖTVÖS LORÁND

TUDOMÁNYEGYETEM

Budapest, 2023

Köszönetnyilvánítás

Ezúton szeretném megköszönni a témavezetőmnek, Zempléni Andrásnak a szakdolgozat írása során rámfordított időt és a sok hasznos tanácsot és ötletet.

Köszönöm a családomnak a bíztatást és támogatást, és külön köszönöm testvéremnek aki a modellek futtatásához biztosított számomra nagyteljesítményű hardvert.

Tartalomjegyzék

1. Bevezetés

A klímaváltozás hatásai napjainkban a Föld legtöbb országában érezhetőek. Az ezzel kapcsolatos kockázatok a tevékenységeink jelentős részét befolyásolják, legyen szó az iparról, mezőgazdaságról vagy akár a mindennapi életről.

A bolygó éghajlatának változásából eredő potenciális negatív hatásokat a természetre, az emberi közösségekre vagy rendszerekre klímakockázatnak nevezzük. A klímakockázatoknak komoly gazdasági, társadalmi és környezeti következményei lehetnek, ezért egyre növekvő aggodalomra ad okot a kormányzati, vállalati vagy az egyéni döntésekben egyaránt. E kockázatok közé tartoznak többek közt az extrém időjárási jelenségek, kánikulák, aszályok, árvízek, erdő- és bozóttüzek gyakoriságának növekedése. Az előzőek hirtelenségével ellentétben olyan kockázatokat is figyelembe kell vennünk amelyek csak fokozatosan fejtik ki hatásukat, például az átlaghőmérséklet növekedése, a tengerszint emelkedése vagy a csapadék mennyiségében és mintázatában bekövetkező változások.[1]

A pénzügyi szektorban is egyre nagyobb hangsúlyt kap a klímaváltozás. A világ pénzügyi rendszerei rendkívül összetett és többrétegű adósság- és hitelkapcsolati hálót alkotnak. Ezen kölcsönhatások révén, az éghajlatváltozásból eredő veszteségek még jobban felerősödhetnek, ezáltal globális pénzügyi instabilitást előidézve. [2].

Gazdasági tekintetben három fő csoportba sorolják a korábban említett kockázatokból eredő lehetséges veszteségeket[1, 2]:

1. **Átmeneti kockázatok** (transition risks): az alacsony üvegházgáz-kibocsátású gazdaságra való átállással kapcsolatos kockázatok

2. **Fizikai kockázatok** (physical risks): a hosszútávú változásokból (pl.: tengerszint tartós emelkedése), valamint a megnövekedő frekvenciájú és súlyosságú extrém időjárási jelenségek, kánikulák, aszályok, árvízek, erdő- és bozóttüzek, földrengések és további természeti katasztrófák miatt előforduló költségek és veszteségek

3. **Felelősség kockázatok** (liability risks): az alacsony kibocsátásra való átállás költségeinek fedezéséből, finanszírozásából felmerülő problémák, a fent említett szisztematikus veszteségek elkerülése érdekében

Mindhárom kockázati típusnál a becslésekre és előrejelzésekre a legelterjedtebb metodológia a szcenárióelemzés. A pénzügyi intézetek, vállalatok és kormányzatok a klímakutatók modelljei által meghatározott és széles körben elfogadott forgatókönyvek (SSP) alapján stressztesztelik portfólióikat és kitettségeiket[1, 3].

Ebben a dolgozatban a *fizikai kockázatok*kal foglalkozom, az erdő- és bozóttűz előrejelzésekhez használt matematikai modelleket mutatom be. Vizsgált területnek az amerikai egyesült államokbeli **Kalifornián**t választottam, mert itt világviszonylatban is gyakoriak a nagy kiterjedésű tűzesetek, amelyek egyre több problémát okoznak a térségben. Emiatt nagy mennyiségű erdőtüzekkel kapcsolatos adat publikusan is elérhető.

A keretrendszer és a lépések több másik jelenségből eredő kockázat és veszélyességi eloszlás becslésénél is hasonlóak, ezért a következő fejezetekben bemutatott technikák általánosabban is használhatóak a *fizikai* kockázatok elemzésében.

2. Szakirodalom bemutatása

2.1. Modellezési módszerek

Az extrém időjárási események által okozott kockázat modellezésekor a kockázatelemzés standard folyamata használható. Első körben a lehetséges kockázati tényezőket kell meghatározni, melyekről feltételezzük, hogy hatással vannak a vizsgált esemény bekövetkezésére. Ezután valamilyen regressziós vagy egyéb modell segítségével megpróbáljuk feltérképezni a tényezők és a vizsgált esemény közti kapcsolatot historikus megfigyelések alapján. Miután megalkottuk a modellt, elkészítjük az esemény terület alapú valószínűség vagy veszélyességi eloszlását az aktuális (vagy teszt) adatokból.

Ahhoz, hogy a kockázatot mérni tudjuk az adott területre a valószínűségek mellett a kitettségeket is szükséges meghatározni. Extrém időjárási jelenségek esetén ezek a kitettségek általában a társadalom számára értékes területek, vagyontárgyak vagy ingatlanok melyek sérülhetnek vagy akár meg is semmisülhetnek az esemény bekövetkezésének hatására.

Ezen fizikai kitettségek közvetve is kihatnak a gazdaság más szereplőire, melyből jelentős szisztematikus kockázatok is következhetnek, például okozhatnak ellátási lánc problémákat vagy megnövekedett biztosítói árakat. Ezért a bekövetkezési valószínűség területi eloszlásának minél pontosabb meghatározása kritikus pontja a fizikai kockázatok modellezésének.

Erdő- és bozóttűz kialakulási valószínűségének modellezése során a következő kérdésekre keressük a választ [4]:

1. Milyen tényezők mellett alakulnak ki a tűzesetek?
2. Mikor alakulnak ki a tűzesetek?
3. Hol alakulnak ki a tűzesetek?
4. Ha kialakult egy tűz, hogyan viselkedik?

Az 1-es, 2-es és 3-as kérdésekre regressziós modellekkel is választ adhatunk, ezeknek a modelleknek az előnye, hogy könnyen értelmezőek az eredmények és skálázhatóságuk miatt nagy adathalmazokon is lehet effektíven őket alkalmazni.

A 4-es kérdésnél az adott terület környezetén nagy felbontású területi tényezőinek kis időintervallumok közti változásaira keressük a választ a már bekövetkezett tűzeset esetén.

Lehetséges terjedési szcenáriókat létrehozva, szimulációs modellekkel lehet előrejelezni a kialakult tűz viselkedését. Ebben a dolgozatban csak az 1-3 kérdésekről lesz szó, a 4-esről nem.

A tűzveszélyesség modellezésekor a cikkekben gyakran használt modell a logisztikus regresszió, amely feltételes valószínűségeket becsül bináris célváltozóra [5]. A magyarázó változók és a tűzesetek közti komplex kapcsolatot az elmúlt évtizedekben gépi tanulás - többek között a *random forest* és *tree boosting* modellek - segítségével is próbálták megmagyarázni cikkekben. [6].

Általánosságban elmondható a fent idézett cikkekről, hogy a vizsgált terület felbontása ($1\ km^2$) és a magyarázó változók jellege is megegyezik. Az antropogén hatásokat általában népsűrűség, lakott területtől vagy utaktól vett távolsággal építik bele a modellekbe. Topografikus és klíma- vagy időjárásváltozókat is minden fenti modellben használnak, mint magyarázó változó. Többek között átlaghőmérséklet, csapadékmennyiség, szélsebesség, különféle tűz- és szárazságindexek.

Ebben a dolgozatban a valószínűségek meghatározása mellett megpróbálok még a historikus tűzesetekből és hozzájuk tartozó káradatokból becslést adni a jövőbeli tűzkárokra, klímaszcenáriók által meghatározott előrejelzések mentén.

3. Elméleti háttér

A regressziók elméleti hátterét és a logisztikus regresszió jellemzőit a *G. James, D. Witten, T. Hastie, R. Tibshirani: An Introduction to Statistical Learning* [7] könyvben leírtak szerint mutatom be.

3.1. Regressziók általában

Az erdő- és bozóttüzekből adódó károk modellezésénél először az előfordulás valószínűségét akarjuk valamilyen módon meghatározni. Általában ez a valószínűség valamilyen paraméteres regressziós modell segítségével adható meg. A regresszió célja a magyarázó változók és a célváltozó közti összefüggések minél jobb becslése a megfigyelt adataink alapján. Tehát az egyes megfigyeléseink (X_i, Y_i) párok, ahol X_i m-dimenziós vektor, ha m darab magyarázó változónk van, Y_i pedig a magyarázott vagy célváltozó amely általában 1-dimenziós.

A paraméteres regressziós modellek felírhatóak úgy is mint függvények:

$$Y_i = f(X_i, \beta) + \epsilon_i$$

ahol $f : \mathbb{R}^{m+dim(\beta)} \to \mathbb{R}$ függvény, β a modell paramétervektora, ϵ_i pedig a független hibatag vagy zaj.

Természetesen f a gyakorlatban nem lehet akármilyen alakú többváltozós függvény. Legtöbbször az adathalmazunkból vagy más előzetes tudás alapján már vannak feltételezéseink f bizonyos tulajdonságaira. Korlátozó tényező lehet még f-re a modellezésünk célja is (pl.: folytonos vagy kategorikus a célváltozó).

3.1.1. Definíció. *Az $\hat{y}_i = f(X_i, \beta)$ számot a β paraméterű modell X_i-re adott* ***predikció****jának nevezzük.*

Célunk a modellel olyan β-t találni ami a legjobban illeszkedik az adathalmazunkra. Ennek az illeszkedésnek a mértékét valamilyen csak β-tól függő $c_{(X,Y)}(\beta)$ függvény optimalizálásával kaphatjuk meg.

Az egyik gyakori választás c-re a legkisebb négyzetek módszere

$$c_{(X,Y)}(\beta) = \sum_i (Y_i - \hat{y}_i)^2,$$

vagyis a reziduálisok négyzetösszegének minimalizálása.

3.1.2. Definíció. *Az $Y_i - \hat{y}_i$ számot a modell i-edik* ***reziduálisá****nak nevezzük.*

Az erdő- és bozóttüzek előrejelzésénél a megfigyeléseink területi alapúak. A vizsgált területet rácspontokra osztjuk fel, így az i-edik rácsponthoz tartozó X_i magyarázó változókból álló vektor és egy $Y_i \in \{0, 1\}$ pár lesz egy megfigyelés. Az $Y_i \in \{0, 1\}$ választás a historikus adatokból adódik, mivel minden rácspontra meg tudjuk mondani hogy volt-e ott tűz vagy sem a megfigyelés időintervallumában. Bináris célváltozó esetén az egyes kategóriákba esés feltételes valószínűségének becslésére gyakran használt statisztikai módszer a logisztikus regresszió.

3.2. Logisztikus regresszió

A logisztikus regressziónál azt feltételezzük az adathalmazunkról, hogy a magyarázott változó Y_i feltételesen Bernoulli- vagy indikátor-eloszlású.

$$\mathbb{E}\left[Y_i | X_i\right] \sim \mathbb{I}(p_i)$$

A p_i paraméter nem általános az adathalmazunkra hanem a megfigyelésünktől, vagyis X_i-től függ. Továbbá a Bernoulli-eloszlás a binomiális eloszlás speciális esete:

$$\mathbb{I}(p) \sim Binom(1, p),$$

emiatt a feltételes eloszlás felírható az alábbi módon:

$$\mathbb{P}\left(Y_i = y | X_i\right) = p_i^y (1 - p_i)^{1-y}, \quad y \in \{0, 1\} \tag{3.1}$$

A logisztikus regresszió egy általános lineáris modell, amely az egyszerű lineáris regresszió kiterjesztése valamilyen kapcsolati függvény segítségével. Lineáris regressziónál a predikciók értékkészlete nem korlátos, ezért a feltételes valószínűségek becslése során, olyan kapcsolati függvény szükséges amelynek értékészlete a $[0, 1]$ intervallum. A logisztikus regressziónál ez a kapcsolati függvény a *logisztikus* vagy *szigmoid* függvény:

$$f(x) = \frac{1}{1 + e^{-x}}$$

A változók közti linearitás feltétele miatt logisztikus regressziónál a valószínűségeket a következő alakban keressük:

$$\mathbb{P}\left(Y_i = y|X_i\right) = \frac{1}{1 + e^{-\beta X_i}} \tag{3.3}$$

A 3.3 alak transzformálásával a következőkhöz jutunk:

$$\frac{\mathbb{P}\left(Y_i = y|X_i\right)}{1 - \mathbb{P}\left(Y_i = y|X_i\right)} = e^{\beta X_i} \tag{3.4}$$

Bal oldalon megjelent a keresett feltételes valószínűség *odds* formájában. Valószínűségekkel *odds*-ként gyakran lehet találkozni a sportfogadásoknál, mert ez a reprezentálás jobban illeszkedik a természetes fogadási stratégiákhoz. Például *európai odds* esetén az $\left(1 + \frac{1}{odds}\right) \cdot tét$ formulát használják a kifizetések meghatározásához, amelyre teljesül, hogy minél kisebb a számunkra nyerő esemény valószínűsége, annál nagyobb a kifizetésünk aránya a *tét*hez.

Logaritmust véve mindkét oldalon a jobb oldalt megjelenik a lineáris regresszióban használt alak.

$$\log\left(\frac{\mathbb{P}\left(Y_i = y|X_i\right)}{1 - \mathbb{P}\left(Y_i = y|X_i\right)}\right) = \beta X_i \tag{3.5}$$

Az egyenlet bal oldalát *log-odds*-nak vagy *logit*-nak nevezzük. Mivel 3.5 jobb oldala X_i-ben lineáris, ezért a logisztikus regresszió a *logit*-ban lineáris X_i szerint. A β_i paraméter azt mutatja meg, hogy egy egységgel megváltoztatva az i-edik magyarázó változót mennyivel változik a *logit*.

3.2.1. Paraméterek meghatározása

A regressziós együtthatók becslésének egyik leggyakoribb módja a *maximum likelihood* módszer, vagy rövidebben MLE (*maximum likelihood estimator*). Az optimális paramétereket a *likelihood* függvény maximalizálásával kapjuk meg.

A *likelihood* függvény általános esetben a következő

$$\mathcal{L}(\vartheta) = g(\underline{x}, \vartheta)$$

ahol g a megfigyelések ϑ paraméter szerinti együttes feltételes sűrűségfüggvénye. Logisztikus regressziónál a feltételes eloszlása a célváltozónak Bernoulli-eloszlású, ami diszkrét. Így a *likelihood* függvény a megfigyelések függetlensége és a 3.1 alak miatt a 3.3 által becsült valószínűségek szorzata:

$$\mathcal{L}(\beta) = \prod_{y_i=1} \mathbb{P}\left(Y_i = 1|X_i\right) \prod_{y_j=0} \left(1 - \mathbb{P}\left(Y_j = 1|X_j\right)\right)$$

ha a becsült valószínűségeink közt vannak nagyon kicsi számok, a $\mathcal{L}(\beta)$ függvény maximalizálása numerikusan instabil lehet. Ezért tekintsük inkább az egyenlet logaritmusát

$$\ell(\beta) = \sum_{y_i=1} \log\left(\frac{1}{1+e^{-\beta X_i}}\right) + \sum_{y_j=0} \log\left(1 - \frac{1}{1+e^{-\beta X_i}}\right)$$

$\ell(\beta)$ az ún. *log-likelihood* függvény.

Mivel a logaritmus szigorúan monoton függvény, $\ell(\beta)$ maximalizálása ekvivalens $\mathcal{L}(\beta)$ maximalizálásával. Az optimalizációs feladat a legtöbb gyakorlati esetben numerikus módszerekkel oldható meg legegyszerűbben. Szélsőértékhelyeken a paramétertér szerinti összes parciális deriváltja 0, ezért ezek a feltételek meghatároznak $m+1$ darab egyenletet $m+1$ változóra.

$$\frac{\partial \ell}{\partial \beta_i} = 0, \quad i = 0, 1, ..., m$$

Az egyenletrendszer megoldható iterációs módszerekkel, például a Newton-Raphson módszerrel.

3.2.1. Tétel. *Newton-Raphson módszer és konvergenciája*
Legyen $F : \mathbb{R}^k \to \mathbb{R}^k$ háromszor folytonosan differenciálható függvény, $\hat{x}$ megoldása a $F(x) = 0$ egyenletnek és létezzen $J_F(\hat{x})^{-1}$, ahol J_F az F Jacobi mátrixa. Ekkor $\hat{x}$-nek létezik olyan δ környezete, hogy ha ha $|\hat{x} - x_0| < \delta$, akkor a

$$x_{n+1} = x_n - J_F(x_n)^{-1} F(x_n)$$

sorozat másodrendben konvergál $\hat{x}$-hez, vagyis

$$\lim_{n\to\infty} |x_n - \hat{x}| = 0 \quad \text{és} \quad \exists \mu > 0 : \lim_{n\to\infty} \frac{|x_{n+1} - \hat{x}|}{|x_n - \hat{x}|^2} = \mu.$$

Az $F_\ell(\beta_0, \beta_1, ..., \beta_m) = \left(\frac{\partial \ell}{\partial \beta_0}, \frac{\partial \ell}{\partial \beta_1}, ..., \frac{\partial \ell}{\partial \beta_m}\right)^T$ teljesíti a Newton-Raphson módszer feltételeit, tehát az iteráció határértékeként kapott együtthatók szélsőértékhely-jelöltek. Ellenőrizni kell, hogy tényleg szélsőértékhely-e, ha nem akkor másik kezdőpontból újra kell indítani az iterációt. Az így kapott megoldás lesz a *maximum likelihood becslés* a paraméterekre.

3.3. Krigelés

3.3.1. Motiváció

A tűzveszély-előrejelző modellekben a magyarázó változók gyakran az adott lokáció (vagy rácspont) területi jellemzőit, továbbá időjárási vagy egyéb méréseket reprezentálják. A vizsgált terület rácsfelbontásának nagyításával ahhoz, hogy teljesen pontos képet lehessen alkotni egyre több helyen kell mérést végezni, ami anyagi és egyéb okokból is legtöbbször kivitelezhetetlen. Ezért a területen megfelelően sok helyen végzett mérésből próbáljuk megbecsülni a többi rácspontra a változó értékét valamilyen interpolációs módszer segítségével.

Az időjárási változókat a modellekhez *krigelés* segítségével határoztam meg, amely a geostatisztikában gyakran használt területi interpolációs módszer. A krigelés részleteit *Steiner Ferenc: A geostatisztika alapjai*[8] egyetemi tankönyvben leírtak alapján mutatom be.

A módszert Daniel G. Krige dél-afrikai statisztikus fejlesztette ki bányászatban felmerülő interpolációs problémák megoldására. A *krigelés* egyik alapgondolata, hogy egy ismeretlen ponthoz tartozó értéket, az ismert mérések súlyozott átlagaként állítja elő. Ezt a súlyozást viszont úgy állítjuk be, hogy a becslésünk szórása minimális legyen.

Könnyen belátható, hogy ha a pontokon mért értékek páronként függetlenek egymástól, akkor ez a súlyozás a számtani középpel egyezik meg. Legyen a vizsgált mennyiség, mint valószínűségi változó szórása σ, és tegyük fel, hogy van $n+1$ darab pontunk melyeknél az értékek páronként függetlenek. Továbbá legyen P az $n+1$ pont közül az egyetlen olyan pont amelyre nincsen mérésünk, a súlyozás miatt az X_P értéket a következő alakban keressük:

$$\hat{X_P} = s_1X_1 + s_2X_2 + \cdots + s_nX_n, \text{ ahol } \sum_{i=1}^{n} s_i = 1$$

amelyből a becslés szórásnégyzete a függetlenség miatt

$$\sigma_{X_P}^2 = \sigma^2 \sum_{i=1}^{n} s_i^2$$

innen σ_{X_P} minimális, ha $\sum_{i=1}^{n} s_i^2$ is minimális, amely pedig a számtani és négyzetes közép közti egyenlőtlenségből akkor lesz az, ha minden s_i egyenlő.

A valóságban viszont az egymástól kis távolságra lévő pontok értékeire irreális lenne

ez a feltevés. A krigelésnél ezért feltesszük, hogy a mérési értékek nem függetlenek egymástól és az interpolált pontok értékei jobban korrelálnak a közelebbi pontokra, mint a távolabbiakra.

3.3.2. Variogram

Legyen Z_P a pontokban vizsgált (vagy mért) valószínűségi változó, amelyet becsülni akarunk a mérési pontoktól eltérő helyeken.

3.3.1. Definíció. *(Szemi)variogram: Az a görbe, amely a távolság függvényében adja meg a mérési értékkülönbségek négyzetösszegének felét.*

$$\gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} \left[Z_{P_i} - Z_{P_{i+h}} \right]^2$$

ahol h a távolság, P_i az összes olyan pont amelytől h távolságra még van mért érték, $n(h)$ pedig az egymástól h távolságra lévő pontpárok száma.

A variogramban szereplő $Z_{P_i} - Z_{P_{i+h}}$ különbségek átlaga a szimmetria miatt 0 egy adott h-ra, ugyanis az összegben szerepel minden h távolságra lévő pár és azok -1-szerese is. Tehát a $Z_{P_i} - Z_{P_{i+h}}$-kra tekinthetünk úgy is, mint az átlagtól való eltérésekre, vagyis a variogram a h függvényében a h távolságra lévő pontok értékkülönbség-szórásnégyzetének a felét becsüli a tapasztalati szórásnégyzet segítségével.

$$\gamma(h) \approx \frac{1}{2}\mathbb{D}^2 \left(Z_x - Z_{x+h} \right)$$

A variogrammal a célunk a h-tól függő kovarianciák meghatározása. A szórásnégyzet és a kovariancia közti kapcsolatot kihasználva

$$\mathbb{D}^2 \left(\mu - \xi \right) = \mathbb{D}^2 \left(\mu \right) + \mathbb{D}^2 \left(\xi \right) - 2\mathrm{Cov} \left(\mu, \xi \right)$$

$\mu = Z_x$, $\xi = Z_{x+h}$ helyettesítéssel a következő egyenletet kapjuk:

$$\gamma(h) \approx \frac{1}{2}\mathbb{D}^2 \left(Z_x - Z_{x+h} \right) = \frac{1}{2} \left[\mathbb{D}^2 \left(Z_x \right) + \mathbb{D}^2 \left(Z_{x+h} \right) \right] - \mathrm{Cov} \left(Z_x, Z_{x+h} \right)$$

melyet átrendezve és kihasználva, hogy $\mathbb{D}^2 \left(Z_x \right) = \mathbb{D}^2 \left(Z_{x+h} \right)$, mert ugyanazokat a pontokat nézzük meg adott h-ra, becslést kaphatunk a kovarianciára:

$$\widehat{\mathrm{Cov}} \left(Z_x, Z_{x+h} \right) = \mathbb{D}^2 \left(Z_x \right) - \gamma(h) \tag{3.6}$$

A korreláció a legtöbb gyakorlatban előforduló esetben csak bizonyos $h = H$ távolságig áll fent, azon túl már a kovariancia nulla értékű lesz. Ezt a H távolságot *hatástávolság*nak nevezzük, $|h| \geq H$ esetben 3.6 miatt $\gamma(h) = \mathbb{D}^2(Z_x)$, amelyből

$$\widehat{\mathrm{Cov}}(Z_x, Z_{x+h}) = \gamma(H) - \gamma(h) \tag{3.7}$$

alakban is kifejezhető a becsült kovariancia érték bármely h értékre.

A variogram számításakor a gyakorlatban csak véges sok távolságra számítjuk ki a görbe értéket. Az így kiszámított értékek némi ingadozással adják vissza az elméleti valós kovarianciaértékeket, ezért valamilyen variogram-modellt illesztünk az adatokból számolt véges sok pontra a variogramon.

Ha h tart nullába a legtöbb esetben feltételezhetjük, hogy $\gamma(h)$ is tart a nullába, e feltételt indokolt esetekben elhagyhatjuk, de ekkor egy plusz *nugget* vagy rög paramétert is bele kell vennünk az illesztésbe.

A geostatisztikában gyakori választott variogram-modell a *szférikus* modell, amelyben expliciten szerepel a *hatástávolság*:

$$\gamma(h) = \begin{cases} C\left(\frac{3}{2}\frac{h}{H} - \frac{1}{2}\left(\frac{h}{H}\right)^3\right) & \text{ha } 0 \leq h \leq H \\ C & \text{ha } h > H. \end{cases} \tag{3.8}$$

További variogram-modellek az exponenciális-

$$\gamma(h) = C\left[1 - exp\left(-\frac{h}{A}\right)\right] \tag{3.9}$$

és a Gauss-modell:

$$\gamma(h) = C\left[1 - exp\left(-\left(\frac{h}{A}\right)^2\right)\right] \tag{3.10}$$

A keresett C-t a legkisebb négyzetek elve szerint találjuk meg. Mindhárom modellnél a $C = \mathbb{D}^2(Z)$-hez tartanak a görbék, így biztosítva a *hatástávolság*ot. Az exponenciális és Gauss-modellnél előbb kell meghatározni a $\gamma(H)$ és a C kapcsolatát, melyre egy lehetséges példa a $\gamma(H) = 0.9 \cdot C$. Ezután H-val kifejezve A-t elvégezhető a modell-illesztés.

Így a variogramból számolható bármely h-ra a kovariancia becslése:

$$\widetilde{\mathrm{Cov}}(Z_x, Z_{x+h}) = C - \gamma(h). \tag{3.11}$$

3.3.3. Interpoláció mérési értékkel nem rendelkező pontokra

Egy mérési értékkel nem rendelkező P_0 pontra az eddigi eredményeket használva akarunk minimális szórásnégyzetű becslést adni a mért értékek megfelelő súlyozásával. A becslés szórásnégyzetén a valódi Z_{P_0} és a súlyozott átlaggal kapott becslés eltérésének szórásnégyzetét értjük, vagyis a $\mathbb{D}^2\left(Z(P_0) - \sum_{i=1}^{n} s_i Z_{P_i}\right)$ értéket akarjuk minimalizálni.

Kihasználva, hogy a krigelésnél feltételezzük hogy minden pontnál Z-nek ugyanaz a várható értéke és az s_i-k összege 1, ki tudjuk fejezni a becslés eltérésének szórásnégyzetét csak a kovarianciák segítségével:

$$\begin{aligned}
\mathbb{D}^2\left(Z_{P_0} - \sum_{i=1}^{n} s_i Z_{P_i}\right) &= \mathbb{D}^2\left[\left(Z_{P_0} - \mathbb{E}[Z]\right) - \sum_{i=1}^{n} s_i \left(Z_{P_i} - \mathbb{E}[Z]\right)\right] = \\
&= \mathbb{E}\left[\left(Z_{P_0} - \mathbb{E}[Z]\right)^2 - 2\sum_{i=1}^{n} s_i \left(Z_{P_0} - \mathbb{E}[Z]\right) \cdot \left(Z_{P_i} - \mathbb{E}[Z]\right) + \right. \\
&\left. + \sum_{i=1}^{n}\sum_{j=1}^{n} s_i s_j \left(Z_{P_i} - \mathbb{E}[Z]\right) \cdot \left(Z_{P_j} - \mathbb{E}[Z]\right)\right] = \\
&= \mathrm{Cov}\left(Z_{P_0}, Z_{P_0}\right) - 2\sum_{i=1}^{n} s_i \mathrm{Cov}\left(Z_{P_0}, Z_{P_i}\right) \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{n} s_i s_j \mathrm{Cov}\left(Z_{P_i}, Z_{P_j}\right) \qquad (3.12)
\end{aligned}$$

jelöléseket megkönnyítve legyen $c_{ij} = \mathrm{Cov}\left(Z_{P_i}, Z_{P_j}\right)$, amely P_i és P_j távolságát ismerve megkapható a variogramból mint $C - \gamma(h)$.

A 3.12 egyenlettel megadott szórásnégyzetet minimalizáló s_i súlyokat akarunk meghatározni, de még hozzá kell adni a $\sum_{i=1}^{n} s_i = 1$ feltételt is. Így a Lagrange-multiplikátor módszert alkalmazva a minimalizálandó $L(\underline{s}, \lambda)$ többváltozós függvény a következő:

$$L(\underline{s}, \lambda) = c_{00} - 2\sum_{i=1}^{n} s_i c_{0i} + \sum_{i=1}^{n}\sum_{j=1}^{n} s_i s_j c_{ij} + 2\lambda\left(\sum_{i=1}^{n} s_i - 1\right)$$

A minimumhely meghatározásához a változók szerinti parciális deriváltaknak 0-nak kell lennie. Deriválás és átrendezés után a következő egyenletrendszerhez jutunk:

$$
\begin{aligned}
c_{11}s_1 + c_{12}s_2 + \cdots + c_{1n}s_n + \lambda &= c_{01} \\
c_{21}s_1 + c_{22}s_2 + \cdots + c_{2n}s_n + \lambda &= c_{02} \\
&\vdots \\
c_{n1}s_1 + c_{n2}s_2 + \cdots + c_{nn}s_n + \lambda &= c_{0n} \\
s_1 + s_2 + \cdots + s_n &= 1
\end{aligned}
$$

Az egyenletrendszer lineáris és $n+1$ változóból, valamint $n+1$ egyenletből áll, így mátrixinvertálással könnyen megoldható és a megoldás egyértelmű:

Legyen

$$
S_0 = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \\ \lambda \end{bmatrix} \quad \text{és} \quad C_0 = \begin{bmatrix} c_{01} \\ c_{02} \\ \vdots \\ c_{0n} \\ 1 \end{bmatrix},
$$

továbbá az egyenletrendszer mátrixa K, amelyet *Krige-mátrix*nak is neveznek

$$
K = \begin{bmatrix}
c_{11} & c_{22} & \dots & c_{1n} & 1 \\
c_{21} & c_{22} & \dots & c_{2n} & 1 \\
\dots & & & & \dots \\
c_{n1} & c_{n2} & \dots & c_{nn} & 1 \\
1 & 1 & 1 & 1 & 1
\end{bmatrix}
$$

ekkor a keresett súlyok megadhatóak a következő alakban:

$$
S_0 = K^{-1} \cdot C_0
$$

A *Krige-mátrix*ot és inverzét elég egyszer kiszámolni az adatainkból és utána tetszőlegesen sok ismeretlen pontra a súlyokat meghatározhatjuk egyetlen mátrixszorzás segítségével.

3.4. Döntési fa alapú modellek

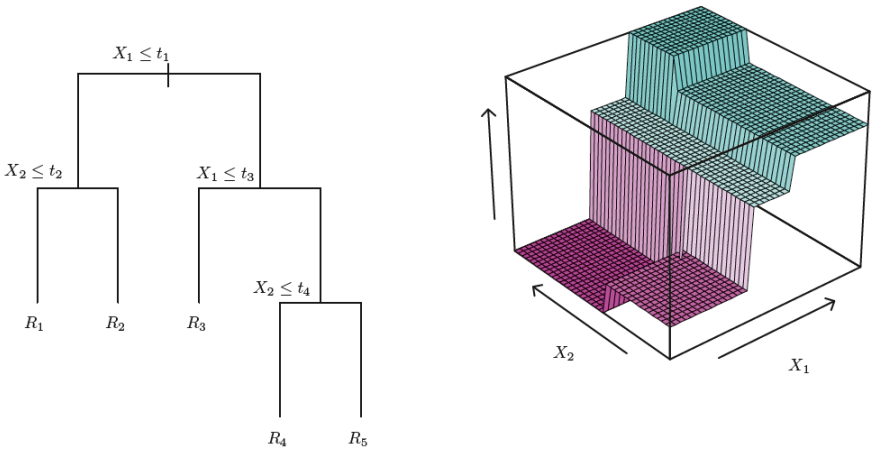
A döntési fa, *random forest* és *tree boosting* modelleket a *G. James, D. Witten, T. Hastie, R. Tibshirani: An Introduction to Statistical Learning* [7] könyvben leírtak alapján mutatom be.

3.4.1. Döntési fa

A döntési fa (vagy angolul *decision tree*) alapötlete, hogy a magyarázó változók terét kisebb részekre osztjuk a változók szerint valamilyen logika alapján. Ezeket a változók szerinti vágások sorozatát egy fával is reprezentálhatjuk, innen az algoritmus neve. Az általános lineáris modellekkel szemben, a döntési fánál nem teszünk fel lineáris kapcsolatot a magyarázó változók közt, így a regressziónál használt f függvény ennél a modellnél a következő módon írható fel:

$$f(X) = \sum_{k=1}^{K} c_k \mathbb{I}(X \in R_k) \tag{3.13}$$

ahol K a változók terének felosztásával kapott csoportok száma, R_k a k-adik csoport és c_k az k-adik csoporthoz tartozó predikciós érték, melyet az R_k csoportba tartozó megfigyelésekből kapunk valamilyen statisztika alapján (leggyakrabban az átlag, medián vagy módusz). Tehát az egy csoportba tartozó összes megfigyelésre ugyanazt a predikciót kapjuk és döntési fák teljesítménye csak a csoportok megválasztásától függ. Elméletben ezeknek a csoportoknak bármilyen alakja lehetne, de az algoritmus hatékonysága és az értelmezhetőség miatt egyszerűen $\mathbb{R}^m$-beli téglákra osztjuk a változók terét.



1. ábra. Döntési fa vizualizáció két dimenziós adatokra[7]

A célunk az algoritmus során olyan $R_1, R_2, \ldots, R_K$ többdimenziós téglalapok megtalálása az adataink alapján, amely minimalizálja a reziduálisok négyzetösszegét (vagy rövidítve *RSS*-t): $\sum_{k=1}^{K} \sum_{i \in R_k} (y_i - \hat{y}_{R_k})^2$.

Minden lehetséges téglalap-felbontást megnézni nem lenne hatékony, ezért a döntési fában *rekurzív bináris vágás*okat hajtunk végre a változók szerint mohón. Kezdetben minden megfigyelés egy csoportba tartozik és innen minden lépésben valamelyik csoportot kettéosztjuk. Az algoritmus a következő módon írható le:

1. Legyen R_k az k-adik levélhez tartozó megfigyelések halmaza és n_k az elemszáma.

2. Egy lehetséges $\theta = (i, t)$ vágás R_k-t két részre osztja:

$$R_k^{bal}(\theta) = \{(x, y) \in R_k | x_i \leq t\}$$
$$R_k^{jobb}(\theta) = R_k \backslash R_k^{bal}(\theta)$$

3. Megkeressük azt a θ^* vágást amely a legnagyobb *RSS* redukcióhoz vezet (komplexitásban ez megoldható $\mathcal{O}\left(m \cdot |R_k| log(|R_k|) \cdot |R_k|\right)$ időben)

4. Ha még nem értünk el a megállási kritériumba, akkor rekurzívan megismételjük a θ^* vágás után keletkező $R_k^{bal}(\theta^*)$ és $R_k^{jobb}(\theta^*)$ csoportokra

A 4. pontban a megállási kritériumot többféleképpen is lehet definiálni, egyik lehetséges módszer, hogy legfeljebb l megfigyelés eshet egy csoportba. Itt ha l-et az adathalmaz méretéhez képest kicsinek választjuk meg *túltanulás*hoz vezethet a modellben. Másik módszer lehet még a megállási kritériumra, hogy előre megadjuk milyen mély legyen a fa, ezáltal lecsökken a modell torzítása. Mivel az algoritmus mohón választja a vágást, általában nem olyan jó teljesítményű modell más regressziókhoz képest. Előfordulhat, hogy egy nagyon jó vágáshoz nem tudunk eljutni mohó lépésekkel, továbbá ha véletlenszerűen kettéválasztjuk az eredeti adathalmazunkat és ezeken külön-külön futattunk döntési fát akár teljesen más eredményeket is kaphatunk, mint az eredetinél.

A döntési fa ezek miatt önmagában nem tudja felvenni a versenyt szofisztikáltabb regressziós modellekkel, viszont több fát összekombinálva és a fák konstruálásánál trükköket alkalmazva lényegesen jobb teljesítmény is elérhető. A több döntési fát összekombináló vagy trükköket használó modellek többek közt a *random forest* és a *tree boosting*.

3.4.2. Random forest

A *random forest* (vagy magyarul véletlen erdő) egyik ötlete, hogy a döntési fa magas szórásnégyzetét megpróbáljuk lecsökkenteni. A szórásnégyzet tulajdonságai alapján ha van n darab független adathalmazunk $A_1, A_2, \ldots, A_n$, és egy f modellünk σ^2 szórásnégyzettel, akkor az $\hat{f} = \frac{1}{n}\sum_{i=1}^{n} f_{A_i}$ modellnek a szórásnégyzete $\frac{\sigma^2}{n}$. Tehát n különböző adathalmazra illesztett döntési fát átlagolva le lehet csökkenteni a szórásnégyzetet. A probléma ezzel az ötlettel, hogy a gyakorlatban legtöbbször nincsen n darab megfelelő méretű adathalmazunk.

Egy lehetséges megoldás az eredeti adathalmazból több adathalmaz konstruálására a *bootstrap* módszer. *Bootstrap* során az eredeti adathalmazból véletlen mintavételezéssel választunk ki k darab megfigyelést visszatevéssel. Ezt tetszőlegesen sokszor megismételve az eredetitől nagy valószínűséggel különböző $B_1, B_2, \ldots$ adathalmazokat kapunk. Abban az esetben ha $k = n$ annak a valószínűsége, hogy egy megfigyelés nem kerül be egy *bootstrap* adathalmazba

$$\mathbb{P}(X_i \notin B_j) = \left(1 - \frac{1}{n}\right)^n \to \frac{1}{e} \approx 0.3679, \text{ ha } n \to \infty$$

emiatt egy megfigyelés átlagosan a *bootstrap*-elt adathalmazok 63.21%-ban van benne, így az eredeti adathalmazunkat tanulási-teszt adathalmazokra sem kell szétvágni, ugyanis minden adathalmaznál tekinthetjük azokat a megfigyeléseket amelyek nem kerültek be a *bootstrap* tanulási adathalmazba.

Random forest-nél előre meghatározott paraméterként B darab *bootstrap* segítségével előállított adathalmazra illesztünk döntési fákat. A *random*itás az algoritmus során ott szerepel, hogy az egyes döntési fák nem az összes magyarázó változót használják hanem csak egy véletlen részhalmazát. Erre azért van szükség, mert ha az adathalmazban egyes változók eleinte nagy *RSS* csökkenést okoznak, a legtöbb fában ezek szerepelni fognak mint vágás, vagyis az így kapott fák a *bootstrap* ellenére is korrelálni fognak a predikcióknál. A véletlen részhalmazok miatt lesznek olyan fák amelyek egyáltalán nem is használják ezeket a változókat a vágásoknál, ezért az így kapott fák nem lesznek annyira korreláltak. Továbbá olyan vágássorozatok is előfordulnak egyes fáknál, amelyeket a teljes adathalmaz használatával a mohó algoritmus során nem kaphatnánk.

A *random forest* regressziós f függvénye a következő alakban írható fel:

$$f(X) = \frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} c_k^b \mathbb{I}(X \in R_k^b)$$

Több paramétert is be kell állítani még a *random forest* illesztése előtt, mint például milyen mélyek legyen az egyes fák, hány darab fa legyen, a változók hanyadészét használjuk a véletlen részhalmazoknál és *bootstrap* során mekkorák legyenek az új adathalmazok. Ezeket a paramétereket a modell *hiperparaméter*einek nevezzük és az optimális paraméterekkel rendelkező modell megkeresését pedig *hiperparaméter-optimalizálás*nak.

3.4.3. Tree boosting

Boosting során nem több döntési fa eredményeit átlagoljuk ki, hanem sorozatszerűen haladva minden új fa az illesztésnél felhasználja a korábbi döntési fák által gyűjtött információt, majd az új fát hozzáadja az előzőekhez. A *random forest*-el ellentétben itt nem használjuk a *bootstrap* eljárást, a következő fánál az adathalmazban csak a célváltozót változtatjuk, amely az előző fa λ tanulási rátával összehúzott reziduálisaiból álló vektor. A *tree boosting* predikcióit a következőképpen lehet felírni:

$$\begin{aligned}
\hat{y}_i^0 &= 0 \\
\hat{y}_i^1 &= \lambda f_1(X_i) = \hat{y}_i^0 + \lambda f_1(X_i) \\
\hat{y}_i^2 &= \lambda\left(f_1(X_i) + f_2(X_i)\right) = \hat{y}_i^1 + \lambda f_2(X_i) \\
&\dots \\
\hat{y}_i^k &= \lambda \sum_{j=1}^{k} f_j(X_i) = \hat{y}_i^{k-1} + \lambda f_k(X_i)
\end{aligned}$$

Az algoritmus elején $f_0(X) \equiv 0$ és a célváltozó $r_i^0 = y_i$ minden i-re az adathalmazban. A modellben inputparaméterként meg kell adni a döntési fáknak a számát, legyen ez B. Ekkor az algoritmus a j-edik ($j = 1, 2, \dots, B$) fa illesztésénél a következő lépéseket teszi:

1. Döntési fát illeszt az (X, r^{j-1}) adathalmazra
2. Az így kapott f_j fát hozzáadja az eddigi fákhoz a λ tanulási ráta paraméterrel leskálázva

$$f(X) = f(X) + \lambda f_j(X)$$

3. Célváltozó frissítése az új fa hibájával

$$r_i^j = r_i^{j-1} - \lambda f_j(X_i)$$

A célváltozó frissítésével a következő fa csökkenteni fogja az előző fák által elkövetett hibát a negatív előjel miatt. A tanulási ráta - amely általában 1-nél kisebb pozitív szám - a folyamat lassításáért felelős, ezáltal több különböző alakú fával frissítjük a célváltozót.

A modell regressziós f függvénye ezek alapján a következőképpen írható fel:

$$f(X) = \lambda \sum_{j=1}^{B} f_j(X_i)$$

ahol f_j a 3.13 alapján felírt döntési fa regressziós függvénye.

A *tree boosting* esetén is a modellillesztés előtt inputként meg kell adni több *hiperparamétert* (pl.: hány fa legyen, milyen mélyek legyenek a fák, tanulási ráta). Így *hiperparaméter-optimalizálás*sal itt is megkereshető az optimális modell.

4. Modell

4.1. Vizsgált terület

A 423 970 km^2 összterületű Kalifornia az USA egyik legdiverzebb állama domborzati és éghajlati szempontból egyaránt, területének 4.7%-át víz, 45%-át pedig erdő borítja. Megtalálható egyaránt sivatagi, mediterrán és hegyvidéki éghajlati zóna is. Az északi területek jellemzően hűvösebbek és csapadékosabbak. Legmagasabb pontja a Sierra Nevada hegységben található Mount Whitney 4421 m magasságával, legalacsonyabb pontja a Mojave-sivatagban −85 m-en fekvő Death-Valley, amely egyben nyáron a Föld legmelegebb térsége.

Több, mint 39 millió lakosával az USA legnépesebb állama, gazdasági szerepe is jelentős. Nagyvárosai és nemzeti parkjai világviszonylatban is sok turistát vonzanak, ezért az emberi tevékenység jelentős szinte az állam egész területén.

Kaliforniában régebben is természetes módon előfordultak erdő- és bozóttüzek, viszont a klímaváltozás okozta egyre gyakoribb extrém szárazságok miatt gyakoriságuk és kiterjedésük is jelentősen megnőtt. A 2000-es évek óta az erdőtüzek éves összmérete emelkedő tendenciát mutat, 2020-ben ez eddigi rekordot jelentő $\approx 17\,800$ km^2 méretű volt, amely Kalifornia teljes területének több, mint 4%-a.

Az államnak évente körülbelül 3 milliárd dollárba kerül az erdőtüzek elleni küzdelem, egyes extrém méretű tüzek ezen felül még több milliárd dollárnyi károkat okozhatnak.

4.2. Feladat meghatározás

4.2.1. Tűzveszélyesség modellezés

A kockázatbecslés első lépéseként a tűzveszélyesség területi eloszlását kell meghatározni, ezért a dolgozatban megpróbálkozom egy egész Kaliforniára kiterjedő tűzveszélyességi modell felállításával. A magyarázó változókból, valamint a historikus tűzesetek alapján a célváltozóból álló adathalmazok meghatározása után a 3. fejezetben bemutatott modellek teljesítményét hasonlítom össze.

A tűzveszélyesség modellezésére tekinthetünk, mint bináris klasszifikációra, tehát valamilyen döntési határ meghatározásával a pixelek a modellek által becsült feltételes valószínűségük alapján a *veszélyes* vagy *nem veszélyes* kategóriákba sorolhatóak.

A modellezés során minden pontra becslést adok egy esetleges erdő- vagy bozóttűz kialakulásának valószínűségére az adott hónapban, feltéve hogy ismerjük a megelőző hó-

napok időjárási és egyéb változóit.

A célom a tűzveszélyesség predikciók segítségével egy egész Kaliforniára kiterjedő éves összkár-becslés, amelyet klímaszcenáriók mentén a jövőbeli évekre is kiterjesztek, ezért a modellkomplexitás egyszerűsítése miatt a használt adathalmazok havi adatok 2015 és 2021, valamint a projekcióknál 2025 és 2049 között.

4.2.2. Kárbecslés

A becsült feltételes valószínűségekre a vizsgált területen, tekinthetünk klasszifikáció nélkül is, így nem azt nézzük, hogy hol vannak a veszélyes területek, hanem hogy a vizsgált területen az időintervallumban mekkora területen várható tüzek kialakulása.

4.2.1. Definíció. *Valószínűségi-profil: Adott időintervallumra a modell által becsült feltételes valószínűségek összege a vizsgált területen:* $\sum_i \hat{y}_i$.

Az adott hónapra a valószínűségekből számolt valószínűség-profilokon a historikus tűzesetekhez elérhető kárbecslések segítségével becslést számolok a várható tűzkárokra.

A *CALFIRE* éves jelentéseiben [9] megtalálhatóak a tűzesetekhez tartozó összkárbecslések, tűzméret alapján csoportosítva (A-tól G-ig, melyek kiterjedési intervallumai a 2. ábrán vannak részletezve). Ezeket az összkárbecsléseket a célváltozóhoz használt tűzadatbázisban szereplő tüzek száma alapján kiátlagoltam minden méretcsoprtra (a tüzek az adatbázisban C vagy annál nagyobb méretcsoporthoz tartoznak). Így minden tűzesethez meghatározható egy átlagos kárbecslés.

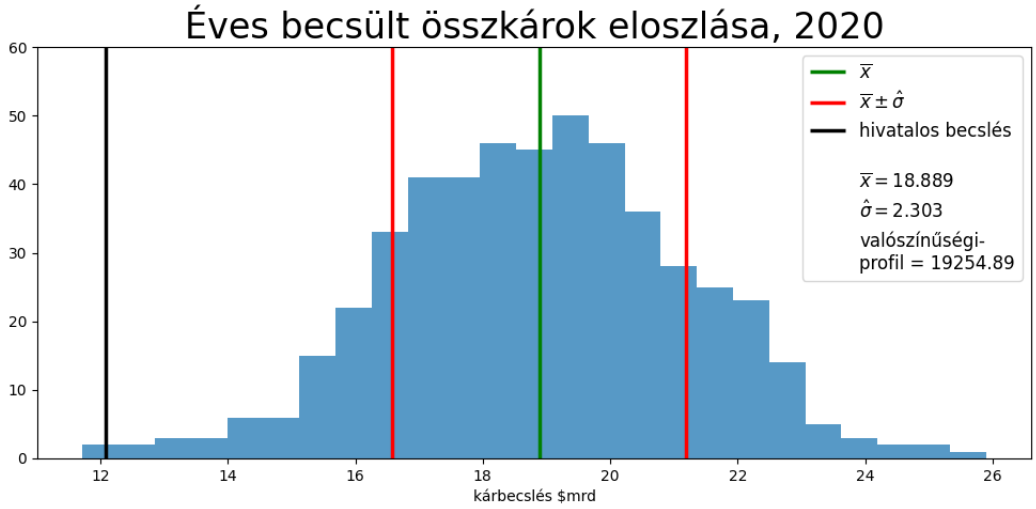
	C	D	E	F	G
kiterjedés (km^2)	0.04 - 0.4	0.4 - 1.2	1.2 - 4	4 - 20.2	20.2 <
kárbecslés / tűz ($)	71000	91000	104000	1852000	146630000

2. ábra. Átlagos kárbecslés táblázat

Egy adott hónapra a kárbecslést ezek segítségével a következőképpen számoltam:

1. Valószínűségi-profil meghatározása valamilyen regressziós modell segítségével
2. Tűzadatbázisból addig választok tüzeket visszatevéssel a *CALFIRE*-ben megtalálható méretcsoportok historikus eloszlása alapján, amíg az összterületük el nem éri a valószínűség-profilt (ha az utolsónak választott tűznél az összterület túllépné az adott valószínűség-profilt, az ahhoz tartozó kiterjedési területet a túllépés mértékével levágtam)

3. A kiválasztott tüzekhez tartozó átlagos kárbecslések összege lesz a hónaphoz tartozó összkárbecslés.



3. ábra. Szimulált tűz összkár-eloszlás, $n = 500$

Az eljárást n-szer megismételve képet kaphatunk a lehetséges összkárok eloszlásáról.

4.3. Implementáció

A modellezési folyamatot legnagyobb részét Python-ban végeztem a népszerű *data science* csomagok és modulok segítségével:

– pandas, numpy, matplotlib

Továbbá más specializált csomagok használatával:

– területi adatok feldolgozása: rasterio, geopandas, geometry, fiona, pyproj

– modellekhez: sklearn, xgboost

A krigeléshez és a magyarázó változók szignifikancia- és egyéb vizsgálatát az R programcsomag használatával valósítottam meg:

– krigelés: geoR

– regresszió: lm, glm

A szakdolgozathoz használt teljes kódbázis és a modellben véglegesen használt adathalmazok megtalálhatóak egy nyilvános *github repository*-ban:

`https://github.com/HKristof136/applied-math-thesis-2023.git`.

A nyers adatok nem kerültek fel a *repository*-ba a nagy tárhelyigényük miatt, de teljes egészében megtalálhatóak a a hivatkozásokon keresztül.

4.4. Adathalmaz konstruálás

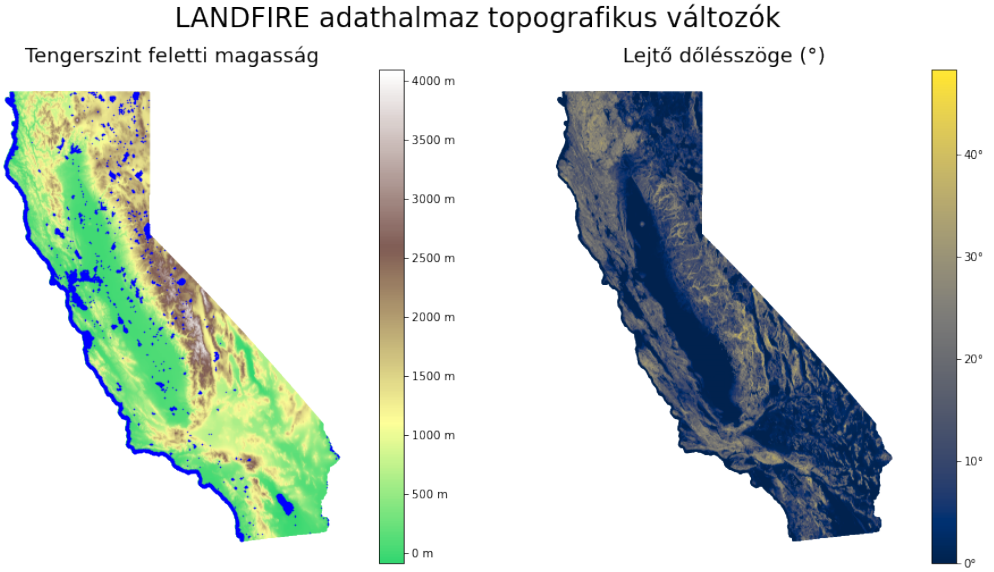
4.4.1. Topografikus változók

A modelleknél használt adathalmaz jelentős része a *U.S. Department of the Interior* és a *U.S. Department of Agriculture Forest Service* közös *LANDFIRE (LF), Landscape Fire and Resource Management Planning Tools* programjának [10] az adatait használja. A kezdeti változók a 2.1. részben bemutatott cikkek alapján kerültek kiválasztásra és ezek közül is amelyek jó minőségben publikusan elérhetőek. A topografikus adatokat a modellben konstansnak tekintjük, mivel változásuk csak hosszabb időtávon válik számottevővé.

A *LANDFIRE* programból származó topografikus adathalmazok:

Adat neve	Adat típusa	Felbontás
Vegetáció/területhasználat jellege	kategorikus	30m x 30m
Tengerszint feletti magasság	numerikus	30m x 30m
Terület lejtése (fokban)	numerikus	30m x 30m
Úthálózat	kategorikus	15m x 15m

Az adatok térbeli autokorrelációjából adódóan a közeli rácspontokon a megfigyelések csak kis mértékben térnek el, ezért célszerű a felbontáson csökkenteni. Ezáltal nem veszítünk lényegi információt és az adathalmazok mérete is jelentősen csökken.



4. ábra. LANDFIRE Topography adathalmazokból kirajzolt tengerszint feletti magasság és a terület lejtése fokban, miután a felbontást 30m x 30m -ről 990m x 990m -esre változtattam

A felbontás csökkentésénél 33×33 pixelt aggregáltam össze. így az adathalmazok mérete 1089-edszeresére csökkent. A közel 1 km^2-es rács még nem nagyon torzítja a topografikus adatokat és területhasználati mintázat is megmarad. Numerikus változó esetén az aggregációs függvény az átlag, kategorikus változónál pedig a leggyakoribb kategória a pixelek közt. Ha egy aggregálási területen több kategória is lehetne a leggyakoribb, akkor véletlenszerűen választottam egyet közülük.

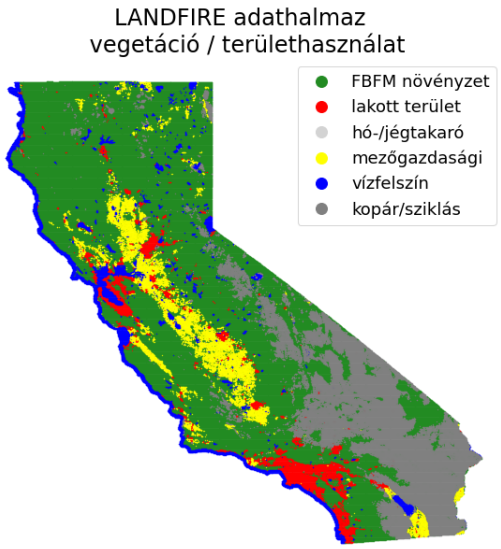
A Vegetáció / területhasználat adatok a *LANDFIRE FUEL 13 Anderson Fire Behavior Fuel Models* modell szerinti földfelszín klasszifikáció. A modell a vegetációs felszínt (a mezőgazdasági területeket kivéve) 13 csoportba bontja éghetőség és egyéb tulajdonságok alapján (FBFM típus).

Azokat a területeket ahol nem természetes vegetáció található a következő csoportokra osztja:

- kopár (Barren)
- vízfelszín (Water)
- mezőgazdasági terület (Agriculture)
- hó- vagy jégtakaró (Snow/Ice)
- lakott terület (Urban)

	99	5	2	10	1	9	93	91	98	8	4	6	11	7	3	92	12
területhasználat	Barren	FBFM type	FBFM type	FBFM type	FBFM type	FBFM type	Agriculture	Urban	Water	FBFM type	FBFM type	FBFM type	FBFM type	FBFM type	FBFM type	Snow/Ice	FBFM type
rácspontok száma	79960	73150	57607	56239	49074	29808	22572	17860	17260	15656	4647	1750	444	267	16	14	3

5. ábra. Kategóriák szerinti eloszlás



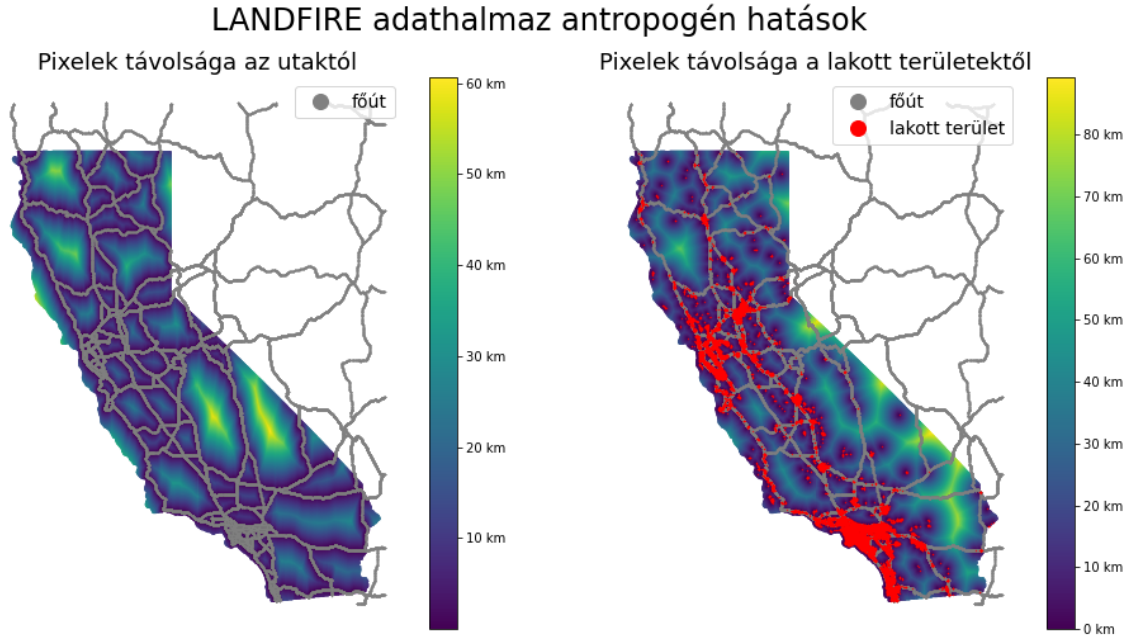
6. ábra. Terület szerinti mintázata a kategóriáknak

A linearitást feltételező regressziós modellek a kettőnél több értéket felvevő kategorikus változókat nem tudják jól kezelni, ezért a vegetáció / területhasználati adatokhoz *one-hot encoding*-ot alkalmaztam, amely minden különböző kategóriához létrehoz egy új bináris változót.
Ezek az új változók felírhatóak a következő módon:

$$X_{ij} = \mathbb{I}(X_i \in \text{Kategória}_j)$$

ahol X_i egy kategorikus változó, j pedig belőle egy egyedi kategória sorszáma.

Az emberek által végzett tevékenységeknek az adott területen hatása van a tűzveszélyességre [4]. Ennek a hatásnak a modellbe illesztése a *LANDFIRE Operational Road dataset* [10] és területhasználati adatokból a lakott terület változó segítségével történik. Minden rácspontra a legközelebbi nagyforgalmú úttól, valamint lakott területtől vett távolsága bekerül az adathalmazba mint változó.



7. ábra. Utaktól és lakott területtől vett távolság

4.4.2. Célváltozó

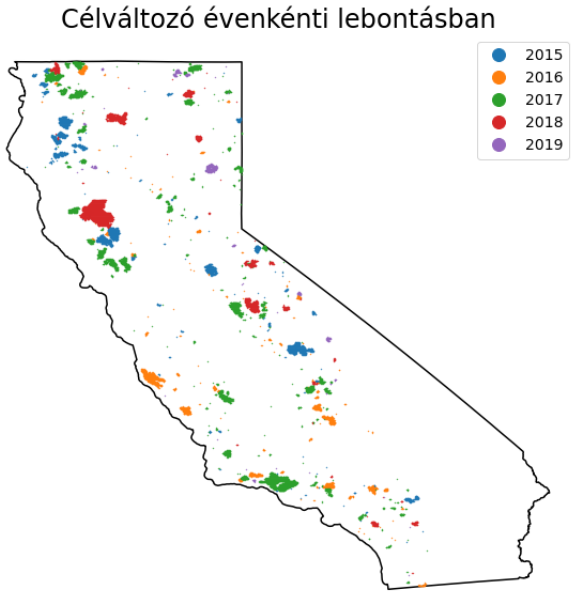
A célváltozót a *USDA Forest Service National USFS Final Fire Perimeter*[11] tűzesetek teljes kiterjedési határaiból konstruáltam meg. Az adatbázisban egy tűzeset teljes kiterjedése poligonok halmazaként van reprezentálva, amelyek unióját egy adott i tűzre F^i-vel jelölök.

Az F^i tűzpoligonoknál fontos kérdés, hogy mely pixelek érintettek a tűz által. Mivel minden X^j pixel a topografikus változók miatt egy T_j ($990m \times 990m$-es négyzet) középpontjának koordinátája, természetes választás lenne az Y^j célváltozó értékére az $Y^j = \mathbb{I}\left(F^j \cap T_j \neq \emptyset\right)$ függvény. Ekkor viszont a kisméretű tüzeknél ($< 1km^2$) a célváltozó többszörösen nagyobb területen venne fel 1-et, mint azt a kisméretű tüzek összterülete indokolná. Ez a hatás a valódi tűz összterülethez képest éves szinten akár másfélszeres is lehet, amely az éves kárbecslésnél jelentős túlbecsléshez vezetne.

Ezért egy adott i hónap adathalmazában az Y_i célváltozót a következő függvénnyel határoztam meg:

$$Y_i^j = \sum_k \mathbb{I}(X_i^j \in F_i^k),$$

ahol X_i^j az i-edik adathalmaz j-edik pontja, és F_i^k egy tűzpoligon amely az i-edik hónapban alakult ki. Ezáltal a kisebb tűzpoligonok közül néhányat nem veszek figyelembe a célváltozó konstruálásakor. Az 8. ábrán a célváltozó 2015 és 2019 között éves szinten aggregálva szerepel.



8. ábra. Tanuló adathalmazokban szereplő tűzesetek

A legtöbb tűzesetre nem érhető el napi lebontású tűzkiterjedés adat, csak teljes maximális kiterjedés, ezért a tűzesetek időbeli kiterjedés-változása emiatt nem építhető be az adathalmazokba. Így egy tűzeset függetlenül attól, hogy a teljes kiterjedését mennyi idő alatt érte el, csak a kialakulási hónap adathalmazában lesz reprezentálva a teljes maximális kiterjedésével.

4.4.3. Időjárási változók

A modellekben a magyarázó változók közül csak az időjárási változók változnak dinamikusan az idővel. Mivel az adathalmazok egy-egy hónapot fednek le a célváltozó miatt, ezért ezeknél a változóknál az adott hónap átlagát rendeljük hozzá a rácspontokhoz.

A következő időjárási adatok fognak szerepelni mint változó az adathalmazokban, és a hozzájuk tartozó havi indexelésű idősor mérőállomásonként megtalálható a *National Oceanic and Atmospheric Administration (NOAA)* honlapján [12]:

– átlaghőmérséklet,

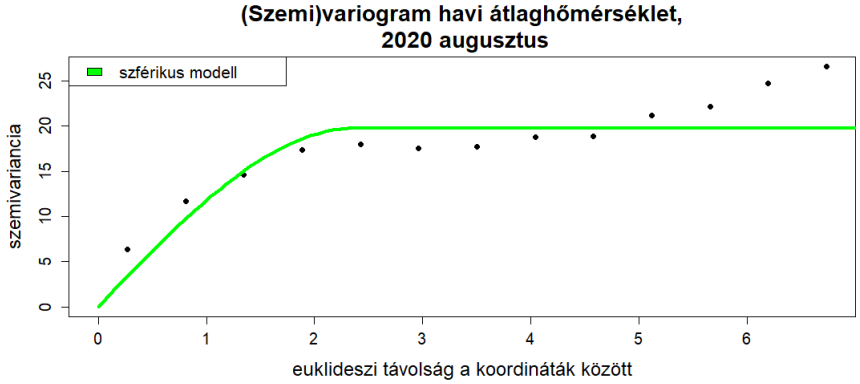
– össz csapadékmennyiség,

– átlag Palmer aszályindex (PDSI).

A Palmer aszályindex adatokat a NOAA *National Integrated Drought Information System* rendszeréből [13] 10 napos időközönként $0.25° \times 0.25°$ felbontással havi szinten kiátlagolva használom. A *PDSI* egy lokálisan sztenderdizált index, amely -4 (az adottt területen extrémnek számító szárazság) és 4 (extrém nedves körülmények) között vesz fel értéket. Az indexet egy komplex formulával számítják a lokálisan mért hőmérséklet, csapadékmennyiség és talajnedvességi szint segítségével.

Az olyan pixelekre amelyekre nincsen mérésünk a 3.3. alfejezetben bemutatott krigelés interpolációval határozom meg az adott változó értékét.

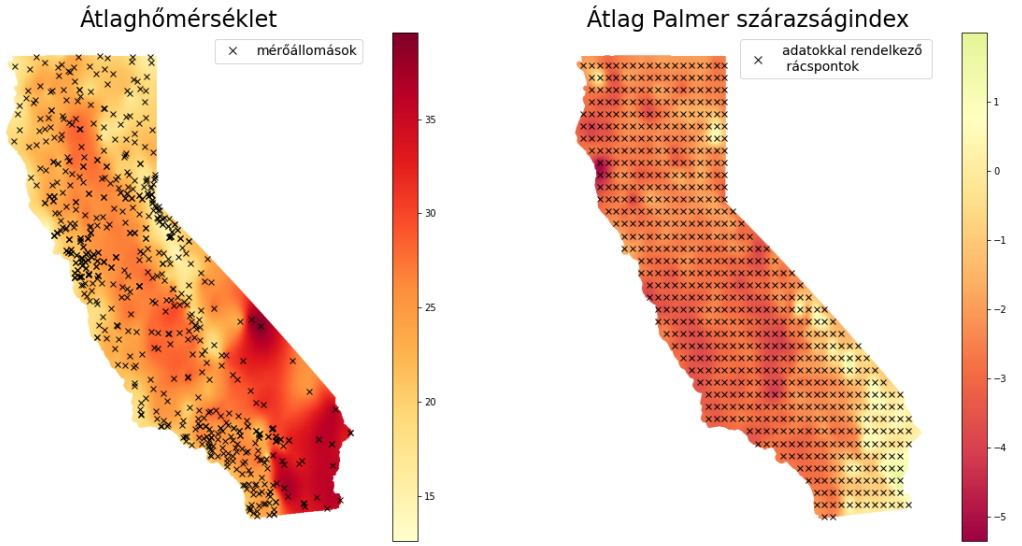
Kaliforniában 2020 nyara rendkívül meleg és száraz volt, amely nagyban hozzájárult a extrém kiterjedésű tüzek kialakulásához. A variogram a 9. ábrán a legmelegebb augusztusi hónapban jól illeszkedik a megfigyelt hőmérséklet adatokra:



9. ábra. Szemivariogram-görbe a szférikus modell illesztése után

Krigelésnél a rög- vagy *nugget*-effektus nem áll fent, ugyanis az időjárási változóknál ha két mérési pont távolsága tart a 0-hoz a megfigyelt értékek közti különbség is tartani fog a 0-hoz. Variogram modellek közül a szférikus modellt 3.8 használom az interpoláláshoz.

A 10. ábrán 2020 augusztusára a krigeléssel számolt adatokból jól kirajzolódnak Kalifornia változatos domborzati és éghajlati mintázatai.



10. ábra. Krigeléssel kiszámolt időjárási adatok vizualizálva, 2020 augusztus

4.4.4. Tanuló adathalmaz kiválasztás

A tanuló adathalmaz időintervallumának a 2015-2019-es éveket választottam, a LANDFIRE adatbázisban elérhető tűzesetek alapján a célváltozó 15 246 pixelen vesz fel 1 értéket. Minden pixel bevétele a tanuló adathalmazba nagymértékű kiegyensúlyozatlansághoz vezetne a 0 osztály felé és a modellek nem tudnának effektíven rátanulni az osztályok közötti különbségekre.

Ennek a problémának a kiküszöbölésére az *undersampling* módszert használom, amely a többségi osztálynak csak egy részhalmazát veszi be a tanuló adathalmazba. A kisebb osztályból az adatpontok kis száma miatt viszont minden megfigyelést megtartok.

A 0 osztálybeli pixelek részhalmazát véletlen választottam, olyan módon, hogy minden hónapból az időintervallumban egyenletesen kiválasztottam 5 pontot és ezek 100 legközelebbi szomszédos pixeljét a 0 osztályból. Így a 0 osztálybeli pontok mintázata egy hónapon belül hasonló a nagyobb tűzesetekével. Ha olyan pont lett sorsolva melynek környezete és az előző pontok környezetének metszete nemüres, akkor az újabb pontot újrasorsoltam.

Ezáltal a végső tanuló adathalmazokban minden hónap egyenletesen van reprezentálva és az osztályok aránya $5 \cdot 12 \cdot 500 = 30\,000$ a 15 246-hoz. Az eljárást 20-szor megismételtem, így 20 különböző tanuló adathalmazra tanítottam a modelleket.

5. Eredmények

5.1. Modell kiválasztás, hiperparaméter-optimalizálás

5.1.1. Teljesítmény metrika

Bináris klasszifikációnál népszerű metrika a *ROC* görbe (*receiver operating characteristic curve*) és az abból számolt *AUC (area under the ROC curve)*. Az *ROC* görbe a különböző döntési határokra számolt (*FPR*, *TPR*) pontokra illesztett görbe, ahol

$$FPR = \frac{\text{hamis pozitív}}{\text{hamis pozitív} + \text{valós negatív}},$$

$$TPR = \frac{\text{valós pozitív}}{\text{valós pozitív} + \text{hamis negatív}}$$

melyekről könnyen látható, hogy 0 és 1 közötti értéket vehetnek fel, vagyis a görbe alatti terület is legfejlebb 1 lehet, amely a tökéletesen klasszifikáló modellel érhető el. Az *AUC* értéke a döntési határtól független, és minél közelebb van az 1-hez annál jobbak a modell predikciói, ezért jól alkalmazható különböző modellek teljesítményének összehasonlítására.

A kárbecslés miatt a modell predikcióira fontos elvárás, hogy az adathalmazok által meghatározott intervallumon a valószínűségi-profil a valódi tűz összterülethez minél közelebb legyen. Így a hiperparaméter-optimalizálásnál, olyan paramétereket keresünk, melyekre az *AUC* maximális és emellett a valószínűség-profil eltérés minimális.

Mivel az *AUC* a skálázástól is független, ezért a maximális *AUC*-vel rendelkező modell predikcióihoz létezik olyan c szám, hogy c-vel átskálázva ($\hat{y}' = c\hat{y}$) az intervallumon belüli valószínűség-profilok és a tűz összterületek közti átlagos négyzetes eltérés minimális. Egy éves időintervallumot nézve, a havi adathalmazokhoz tartozó valószínűségi-profil és valódi tűz összterületekre, az optimális c legyen a következő:

$$c_{\text{opt}} = \arg\min_{c} \frac{1}{12} \sum_{k=1}^{12} \left(\sum_{i} y_k^i - c\hat{y}_k^i \right)^2 . \tag{5.1}$$

5.1.2. Regressziós modellek

Az adathalmazokra tanított modelleknek a 3. fejezetben bemutatott regressziós modelleket választottam: logisztikus regresszió, *Random Forest* és az XGBoost könyvtárból *tree-based XGBRegressor*.

A változók közül a logisztikus regressziónál, csak azokat hagytam meg amelyek az R statisztikai programcsomagban szereplő *glm* szerint szignifikánsak.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-5.406e+01	1.845e+00	-29.292	< 2e-16	***
lon	-4.973e-01	1.867e-02	-26.632	< 2e-16	***
lat	-2.343e-01	1.470e-02	-15.944	< 2e-16	***
ELEVATION	6.869e-04	3.752e-05	18.309	< 2e-16	***
SLOPE	9.553e-02	2.231e-03	42.819	< 2e-16	***
FBFM1	8.403e-01	1.637e-01	5.132	2.87e-07	***
FBFM2	1.727e+00	1.624e-01	10.633	< 2e-16	***
FBFM3	NA	NA	NA	NA	
FBFM4	2.536e+00	1.949e-01	13.015	< 2e-16	***
FBFM5	9.231e-01	1.608e-01	5.742	9.37e-09	***
FBFM6	1.490e+00	2.316e-01	6.434	1.24e-10	***
FBFM7	-2.977e-01	4.957e-01	-0.601	0.54807	
FBFM8	2.001e+00	1.687e-01	11.864	< 2e-16	***
FBFM9	1.693e+00	1.692e-01	10.007	< 2e-16	***
FBFM10	9.305e-01	1.671e-01	5.568	2.58e-08	***
FBFM11	1.459e+00	5.576e-01	2.616	0.00888	**
FBFM12	NA	NA	NA	NA	
FBFM13	NA	NA	NA	NA	
Agriculture	-1.048e+00	2.156e-01	-4.861	1.17e-06	***
Water	-1.468e+00	3.089e-01	-4.752	2.01e-06	***
Barren	-2.537e+00	2.187e-01	-11.597	< 2e-16	***
DISTANCE_FROM_URBAN_AREA	-2.224e-02	1.587e-03	-14.008	< 2e-16	***
DISTANCE_FROM_ROADS	-3.709e-02	3.831e-03	-9.680	< 2e-16	***
PRCP_prev1	-2.626e-02	1.168e-03	-22.486	< 2e-16	***
PRCP_prev2	-2.046e-02	8.192e-04	-24.972	< 2e-16	***
PRCP_prev3	-2.888e-03	3.784e-04	-7.631	2.33e-14	***
TAVG_prev1	2.289e-01	6.416e-03	35.677	< 2e-16	***
TAVG_prev2	-1.252e-01	9.561e-03	-13.100	< 2e-16	***
TAVG_prev3	-8.627e-02	6.823e-03	-12.644	< 2e-16	***
PDSI_prev1	2.672e-01	4.575e-02	5.841	5.20e-09	***
PDSI_prev2	-9.071e-01	7.513e-02	-12.074	< 2e-16	***
PDSI_prev3	4.136e-01	4.293e-02	9.634	< 2e-16	***

11. ábra. Változók szignifikancia-tesztje az R programcsomaggal

A 11. ábra alapján a lakott területtől és az utaktól vett távolságot reprezentáló változók (*DISTANCE_FROM_ROADS* és *DISTANCE_FROM_URBAN_AREA*) együtthatója negatív, amely az emberi tevékenység tüzekre gyakorolt pozitív hatását mutatja. A tengerszint feletti magasság (*ELEVATION*) és a terület lejtése (*SLOPE*) változóhoz tartozó együtthatók pozitívak, tehát a modell szerint a többnyire erdővel borított hegységeknél nagyobb a tűz kialakulásának valószínűsége. Néhány *FBFM* bináris változóra a modellillesztés *NA* értéket ad, mivel ezen változók nagyon ritkák és a véletlen-mintavételezés miatt 1 értékkel rendelkező pixelek nem kerültek be az adathalmazba.

A másik két modellnél a logisztikus regressziónál nem szignifikáns változók meghagyása nem okozott teljesítmény javulást, ezért mindhárom modellnél a 11. ábrán meghatározott változókat használom a tanuló adathalmazokból.

Validációs adathalmaznak a teljes 2020-as évet, teszt adathalmaznak pedig a teljes 2021-es évet választottam minden pixellel. A kárbecslések konzervatívak lesznek a teszt és más a tanulás és validáció során nem látott adathalmazokon, mert 2020 Kaliforniában erdőtűzméret szempontjából rekord év volt, ezért az átskálázáshoz az 5.1. szerint számított

c_{opt} értékek ennél az évnél lesznek a legnagyobbak a historikus évek közül.

Mindhárom modellnél végeztem hiperparaméter-optimalizálást az *sklearn* könyvtárban található *RandomSearchCV* függvény segítségével, az alábbi paramétertereken:

– Logisztikus regresszió

- 'class_weight'$= \{0 : i, 1 : j\}, \quad i, j \in \{1\} \cup \{k \cdot 50 : k = 1, ..., 20\}$

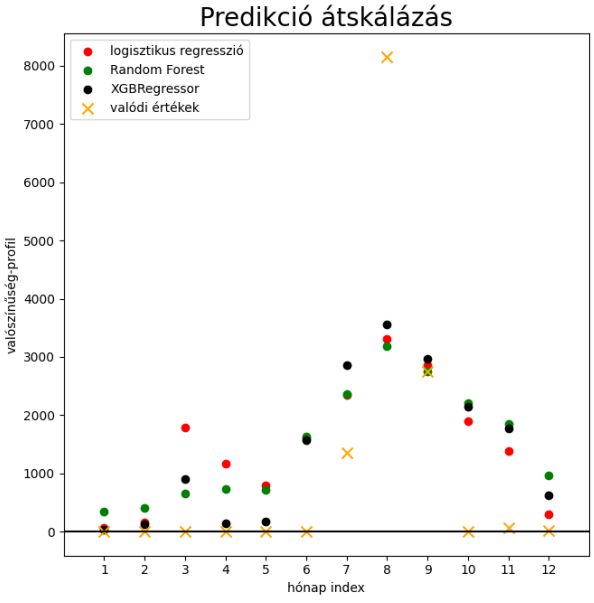
– Random Forest

- 'n_estimators'$= i, \quad i \in \{k \cdot 50 : k = 1, ..., 20\}$
- 'min_samples_leaf'$= i, \quad i \in \{k \cdot 250 : k = 1, ..., 40\}$

– XGBRegressor

- 'n_estimators'$= i, \quad i \in \{k \cdot 50 : k = 1, ..., 20\}$
- 'max_depth'$= i, \quad i \in \{2, 3, 5, 7, 9\}$
- 'learning_rate'$= i, \quad i \in \{k/10 : k = 1, 2, ..., 10\}$

A hiperparaméter-optimalizálás során a maximális AUC-vel rendelkező paraméterezéseket választottam a modellekhez. Az így kapott modellek valószínűség-profil szempontjából a $2:1$-hez arányú undersampling miatt a validációs adathalmazon nagy mértékben túlbecsülnek minden hónapra, ezért a predikciókat 5.1. alapján minden modellnél, minden tanuló adathalmazra meghatározott c_{opt}-al átskáláztam.



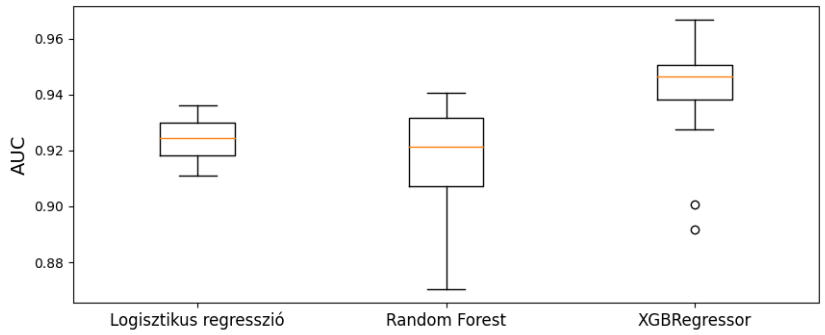
12. ábra. Modell pedikciók átskálázása az optimális c-vel

Az optimális paraméterezések a modelleknél a következők:

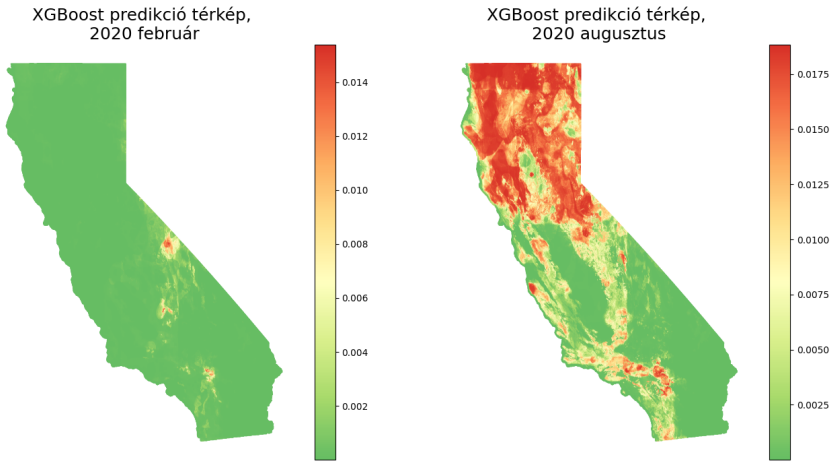
```
Logisztikus regresszió
        -  {'class_weights': {0: 1, 1: 1}, 'solver': 'newton-cholesky'}


Random Forest
        -  {'n_estimators': 250, 'min_samples_leaf': 500}


XGBRegressor
        -  {'n_estimators': 100, 'max_depth': 5, 'learning_rate': 0.1}
```

5.2. Modell AUC eredmények

A teszt adathalmazon a modellek átlagteljesítménye a logisztikus regressziónál $AUC=$ $= 0.924$, a Random Forest-nél $AUC= 0.9174$, az XGBoost-nál $AUC= 0.9422$ értéket vesz fel. Így a klasszifikációs feladatra az XGBoost a legjobb a kipróbált modellek közül a tanuló adathalmazok alapján.



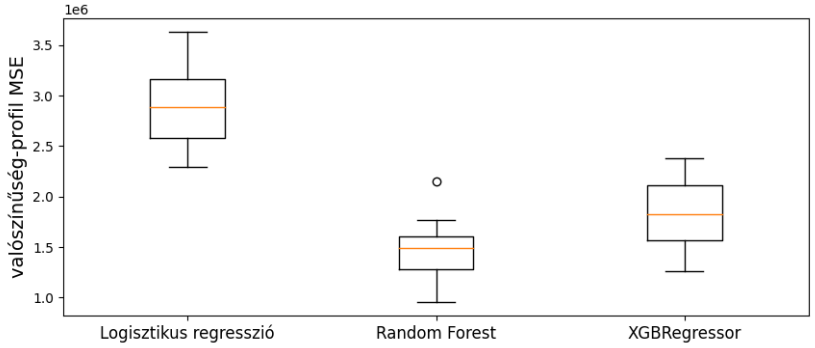
13. ábra. AUC értékek a teszt adathalmazon



14. ábra. XGBRegressor predikciós térképei, február és augusztus hónapokra

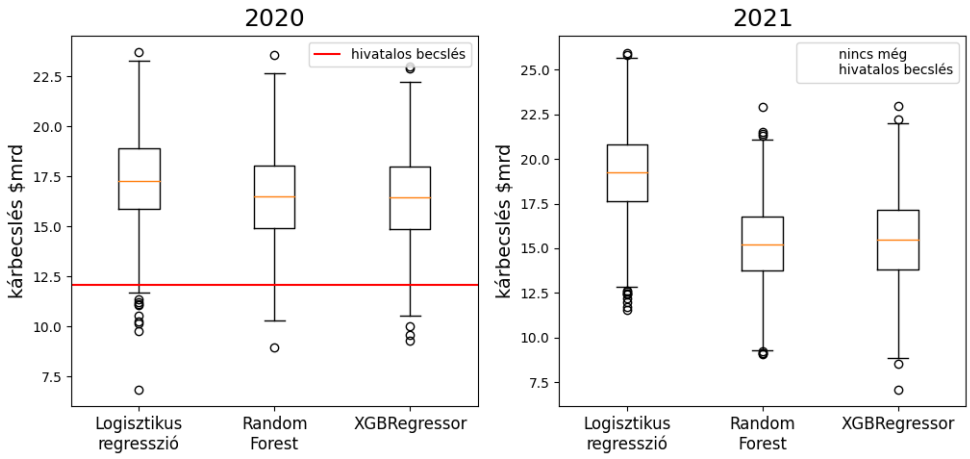
Valószínűség-profilban a validációs adathalmaz konzervatív megválasztása miatt, az eltérések a teszt adathalmaznál jelentősen nagyobbak. A 15. ábrán a teszt adathalmazra havi szinten prediktált valószínűség-profil és a havi célváltozó-összeg közötti átlagos négyzetes eltérés szerepel. A legnagyobb MSE a modellek közül a logisztikus regressziónál figyelhető meg, amelynél gyököt véve havonta átlagosan $\approx 1700\ km^2$ különbséget jelent.

A túlbecslés egyik oka lehet az, hogy 2021 szárazabb volt a 2020-as évhez képest, de 2020 rekord össz tűzterület kiterjedése miatt az éghető növényzettel borított területek még nem regenerálódtak, így nem voltak tűzveszélyesek. Ezt a hatást az alkalmazott modellek nem tudják figyelembe venni a topografikus és vegetációs jellegű változók statikussága miatt az adathalmazokban.



15. ábra. Valószínűség-profil differencia MSE metrikával a teszt adathalmazon

Kárbecslésnél az átskálázással a validációs és a teszt adathalmazon is túlbecslés figyelhető meg a modelleknél éves szinten. A 16. ábrán a különböző tanuló adathalmazra illesztett modellek közös összkár-eloszlások és a hivatalos becslés látható. Mindhárom modellnél több, mint 5 milliárd dolláros a túlbecslés a tűzméret tekintetében eddigi rekordnak számító 2020-as évre (hivatalos becsléseknél általában csak alsó becslést adnak meg).



16. ábra. Minden modellre összesített káreloszlás a validációs és a teszt adathalmozokon

A különbség részben megmagyarázható azzal, hogy a minimális négyzetes eltérésnél a nem tűzszezonbeli (nyár-ősz) hónapokra - ahol általában csak nagyon kevés tűzeset jellemző - jelentősebb mértékben túlbecsülik a valószínűség-profilt a modellek. Továbbá a tűzadatbázisban nem található meg az összes tűzeset, amely a kiátlagolásnál nagyobb kárbecslésekhez vezethet egyes tűzméretcsoportoknál.

6. Jövőbeli projekciók

A klímaprojekciókhoz az adathalmazokat a *Shared Socioeconomic Pathways* vagy rövidítve SSP klímaszcenáriókhoz tartozó *CMIP6 ensemble modell output*ok[14] alapján, havi szinten 2025 és 2049 között, a teszt és validációs adathalmazokhoz hasonlóan konstruáltam.

17. ábra. Különböző *Shared Socioeconomic Pathways* szcenáriók [3]

Az 17. ábrán található klímaszcenáriók közül az SSP2 és SSP5 szcenáriók szerinti projekciókat használtam a jövőre vonatkozó becslésekhez. Ezekben a szcenáriókban szereplő társadalmi és gazdasági változások néhány jellemzője:

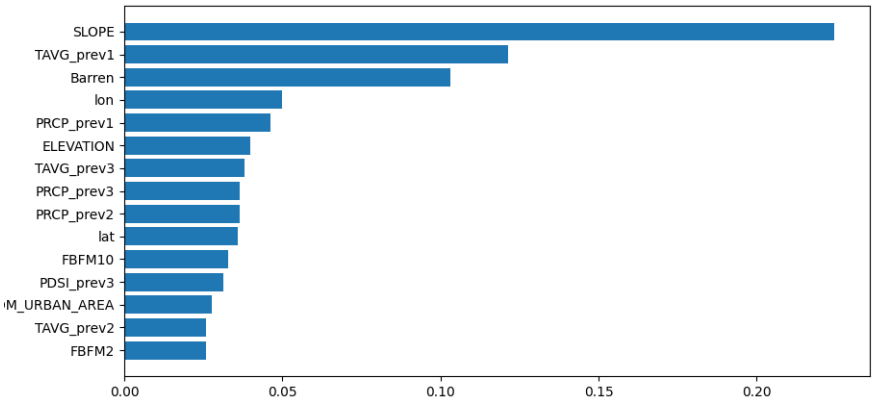
- *SSP2 - Middle of the Road*: társadalmi, gazdasági és technológiai trendek nem térnek el jelentősen a historikusan megfigyeltektől, a fejlődés- és vagyonkülönbségek továbbra is egyenőtlenül oszlanak el a világban. Az országok és vállalatok csak lassan haladnak a kitűzött klímacélok felé.
- *SSP5 - Fossil-fueled Development (Taking the Highway)*: A gyorsabb gazdasági növekedés és technológiai fejlődés érdekében az országok a nem-megújuló energiaforrások minél effektívebb kiaknázását részesítik előnyben. A magas gazdasági növekedés mellett a Föld népesége is gyors ütemben emelkedik. Klímaváltozás okozta problémák enyhítésére az országok kevés erőforrást fordítanak.

Az átlaghőmérséklet és havi csapadékmennyiség változókat a szcenáriókhoz krigelés segítségével a *Copernicus, CMIP6 Climate projections*[15] adatbázisból származó *CESM2* modell output adatokból[16, 17] határoztam meg. A Palmer aszályindexet pedig a fenti modellel konzisztens módon megalkotott PDSI SSP2 és SSP5 projekciókból krigelés segítségével határoztam meg az *NCAR RDA Global Palmer Drought Severity Index (PDSI)* adatbázis adataiból[18].

A legjobb AUC-hez tartozó paraméterekkel rendelkező, mind a 20 tanuló adathalmazra külön-külön tanított modellt futattam le a projekciókra. Majd kiszámoltam mindre a predikciókat és kárbecsléseket. A jövőbeli károk eloszlásához minden modellnél a 4.2.2.-ben szereplő n-et 25-re állítottam be.

6.1. Klímaszcenárió kárelőrejelzések

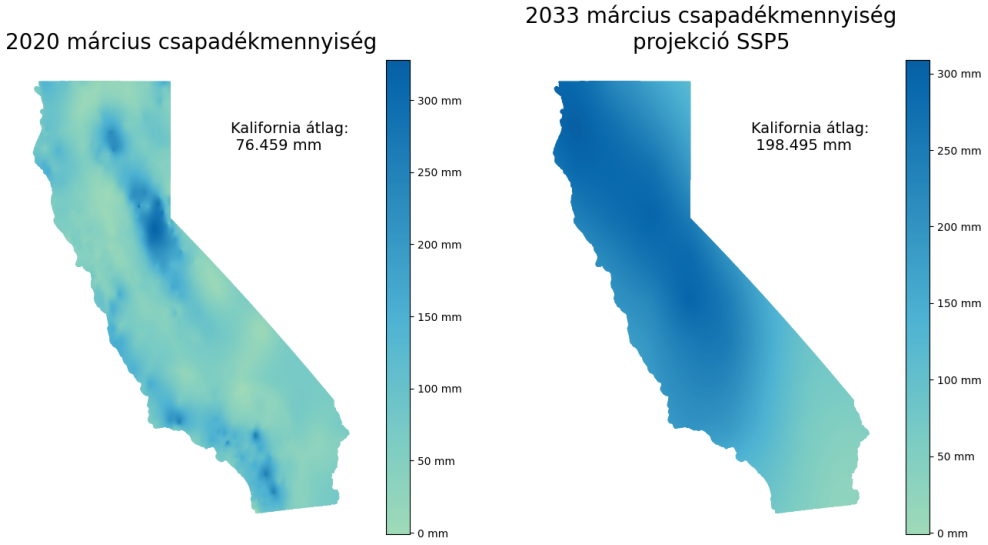
Mindhárom modellnél közel stagnálás figyelhető meg mindkét szcenárióra, amely részben megmagyarázható a statikus adatokkal rendelkező változók fontosságával a modellekben. A döntési fa alapú modellekre elérhető változó-fontossági index, amely egy adott változóhoz az azt használó vágások relatív gyakoriságát rendeli hozzá. Statikus változókra a változó-fontossági indexek összértéke a két modellnél átlagosan > 0.62. A 10 legfontosabb változó az XGBRegressor modellnél az 18. ábrán látható.



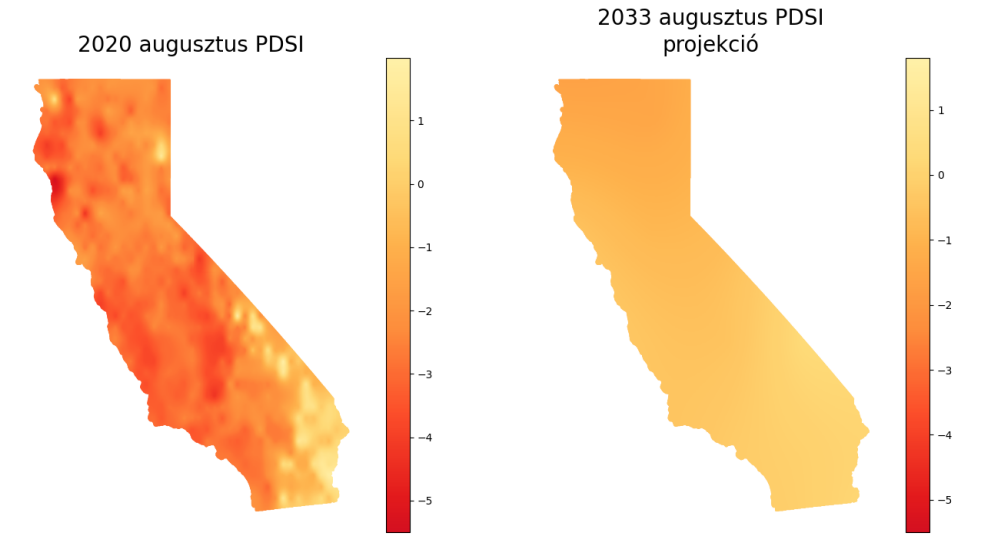
18. ábra. XGBRegressor átlagos változó-fontossági sorrendje

Továbbá a projekciós adatok lényegesen alacsonyabb felbontása miatt ($1.5° \times 1.5°$, a sűrűn lévő mérőállomásokkal szemben) a krigeléssel számolt időjárási változók nem mutatnak területileg olyan pontos mintázatokat, mint a közelmúlt valós megfigyelései. A Palmer aszályindex értéke a krigelés előtti alacsonyabb felbontás miatt lényegesen kisebb terjedelmet és lokális jellegzetességeket mutat.

A 19 és 20. ábrákon a lokális mintázatok elmosódása jól látszódik a havi csapadékmennyiségre és a Palmer aszályindexre egyaránt. A modellek ezért csak tompított mértékben tudják előrejelezni a predikciókban az időjárási változók hatását.



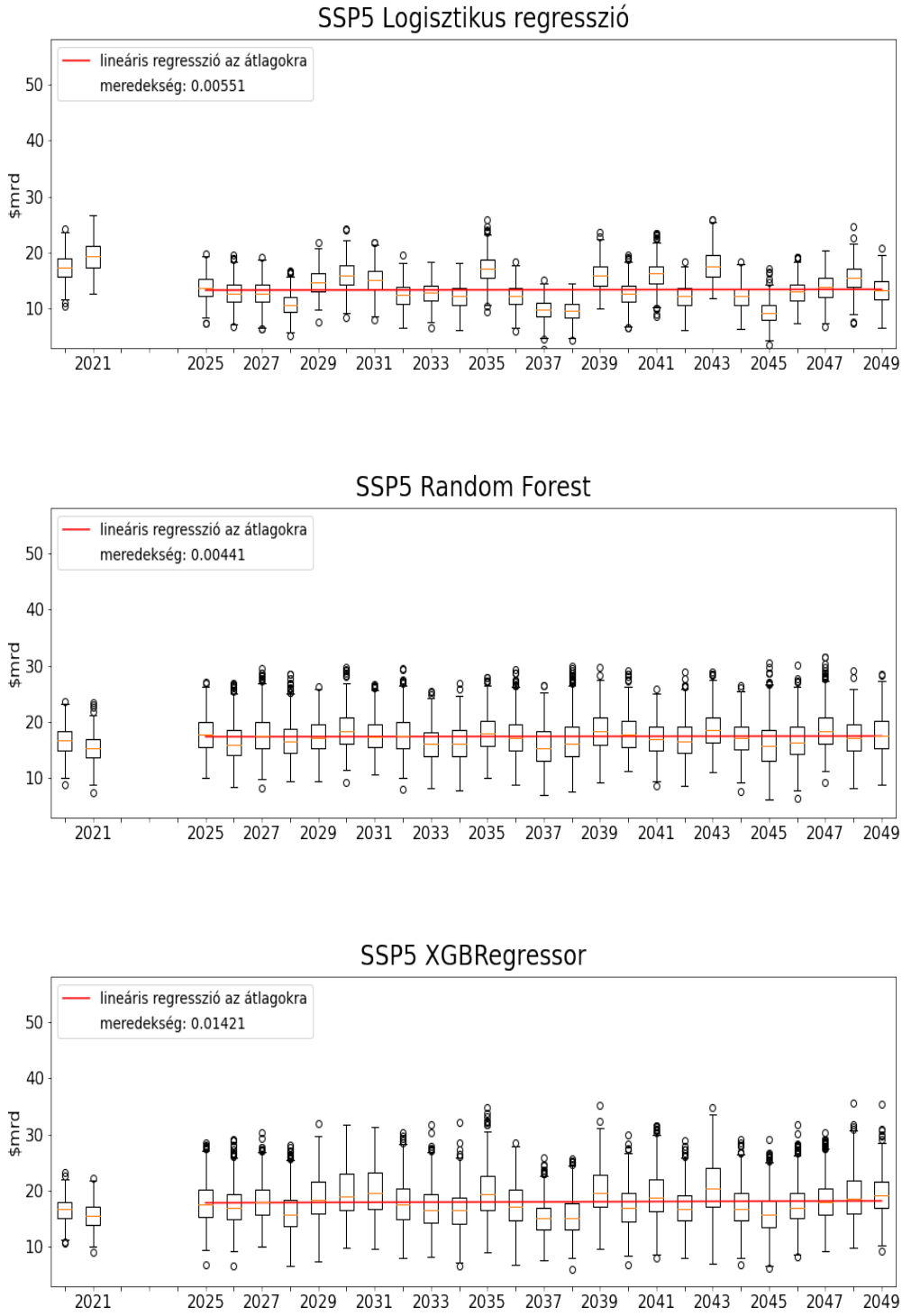
19. ábra. A projekcióknál a havi csapadékmennyiség mintázatában elmosódás figyelhető meg.



20. ábra. PDSI lokális mintázatának összehasonlítása.

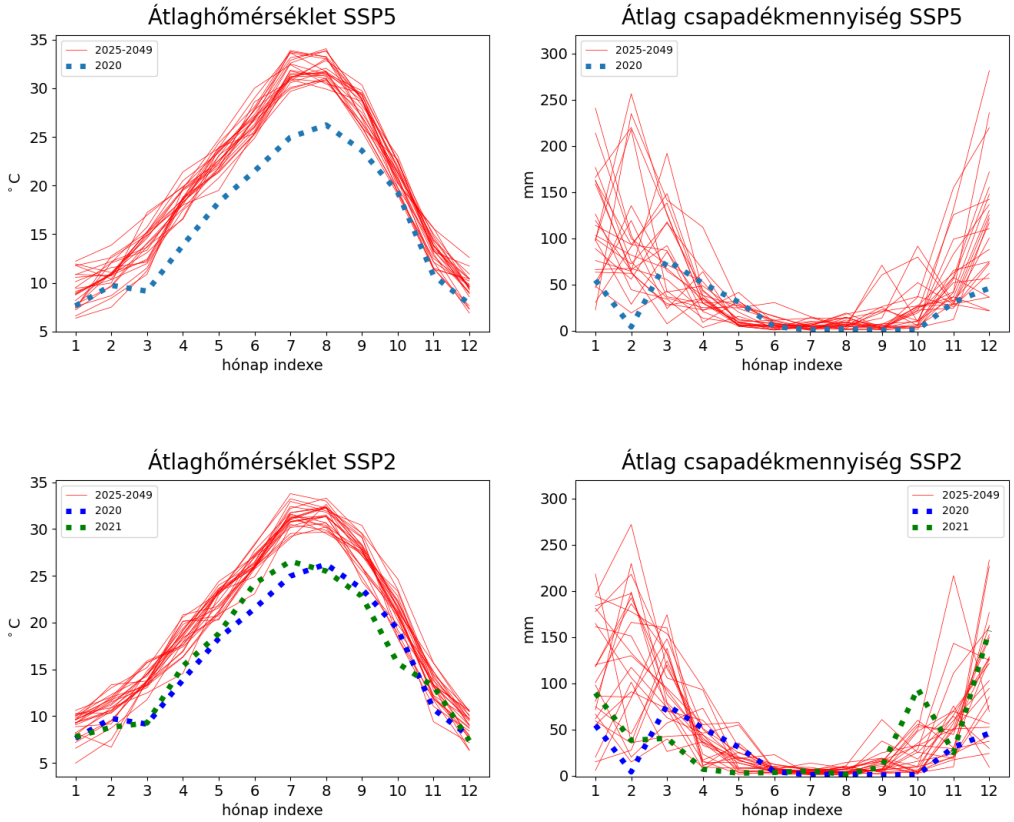
6.1.1. SSP5

A modelleknél az SSP5 klímaszcenárióra az idő előrehaladtával az összkár-becsléseknél nem figyelhető meg emelkedő tendencia. A 21. ábrán a modellek által a 2020 és 2021-es, valamint a jövőbeli évekre becsült összkár-eloszlások láthatóak.



21. ábra. SSP5 szcenárió szerint 2025-2049 időintervallumra futatott előrejelzések a modellek által

A 21. ábrán a logisztikus regressziónál a legtöbb jövőbeli évre a 2020-2021-es évekhez képest az összkár-becslés jelentősen alacsonyabb. Erre magyarázotot adhat a 11. ábrán szereplő együtthatóknál az időjárással kapcsolatos változók fontossága, és a klímaszcenárióban szereplő időjárási változók eltérése a 2020-2021-es évek megfigyeléseitől. A 22. ábrán az egész Kaliforniára vonatkozó havi átlaghőmérséklet és csapadékmennyiség szerepel 2020-2021-re, valamint a jövőbeli évekre. A legtöbb jövőbeli évre az átlaghőmérséklet emelkedésének növelő hatása a valószínűség-profilra nem tudja ellensúlyozni a csapadékmennyiség jelentős csökkentő hatását. A projekcióban a legkevesebb csapadékkal rendelkező évekre (2035, 2043) az összkár-becslés eloszlása megközelíti a 2020-2021-es évekre szimulált értékeket. A 2020 és 2021-es évek szárazsága és csapadékhiánya még a legrosszabb klímaszcenárió szerint is extrémnek számít.



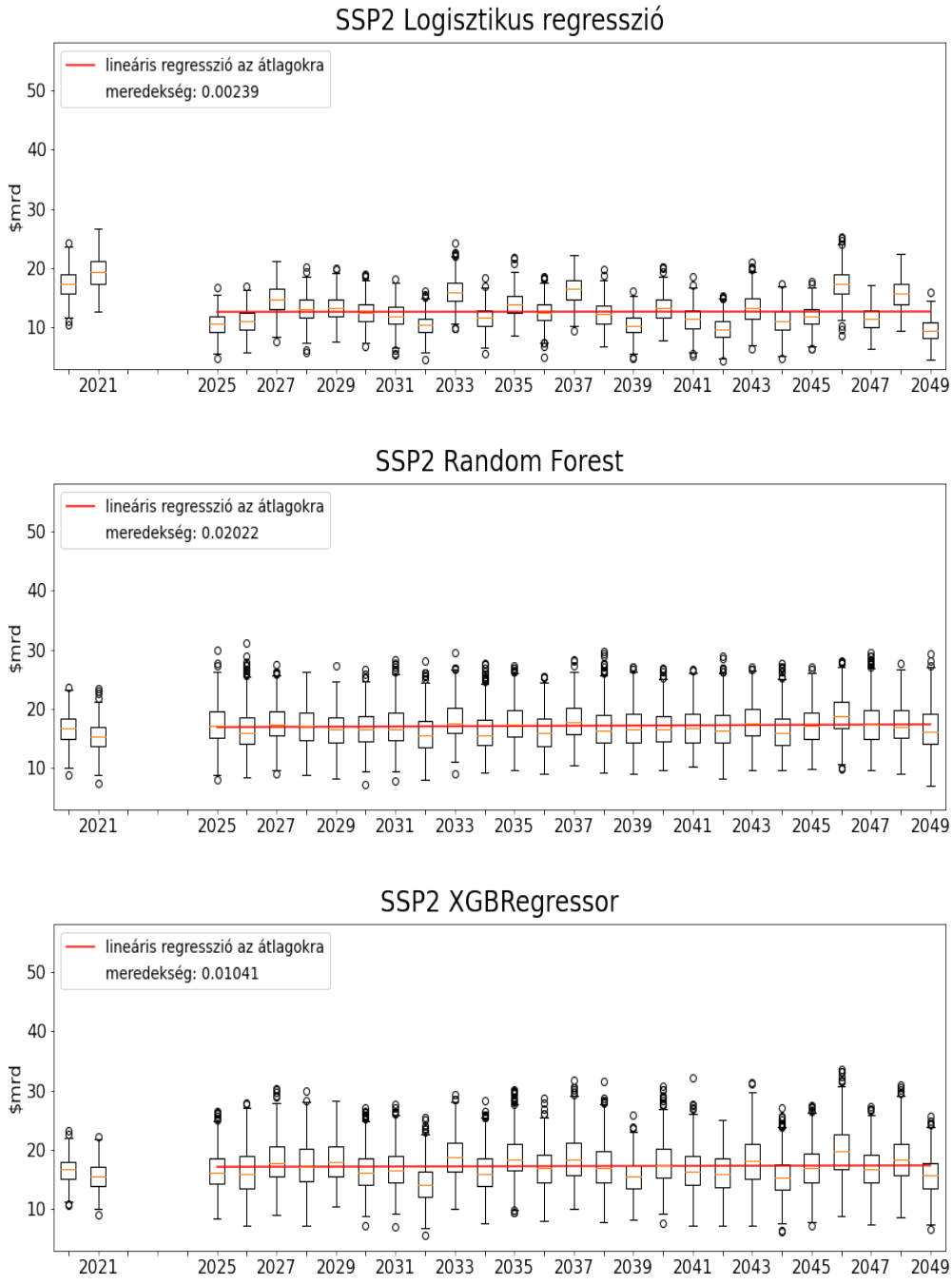
22. ábra. Időjárási változók havi átlagértékei egész Kaliforniára nézve (SSP2 és SSP5)

A döntési fa alapú modelleknél 2020 és 2021-hez képest kissé nagyobb a jövőbeli becslések mediánértéke, amely azt mutatja, hogy ezeknél a modelleknél a hőmérséklet-emelkedés és csapadéktöbblet ellentétes hatása kiegyenlítődik. Viszont itt sem figyelhető meg emelkedő tendencia, ennek egyik lehetséges oka lehet a statikus területhasználati és vegetációs változók fontossága a modellekben. Továbbá az időjárási változók területi sajátosságainak és mintázatának tompított hatása az alacsony felbontás miatt.

6.1.2. SSP2

Az SSP5 és SSP2 szcenáriók közötti időjárástrend eltérések ellenére a modellek nem mutatnak jelentős különbségeket a predikciókban, ugyanis ahogy a 22 ábrán is látszik Kaliforniára 2050-ig a két szcenárió nem tér el jelentősen.

Ezért az SSP5-nél már említett modellspecifikus okok erre a klímaszcenárióra is egyaránt érvényesek. A 23. ábrán a modellek által a 2020 és 2021-es, valamint a jövőbeli évekre becsült összkár-eloszlások láthatóak az SSP2-es klímaszcenárióra.



23. ábra. SSP2 szcenárió szerint 2025-2049 időintervallumra futatott előrejelzések a modellek által

6.1.3. Összesítő táblázat

A 24. táblázatban mindhárom modell által becsült összkár-eloszlások mediánjai szerepelnek mindkét klímaszcenárióra. A 2015-2021 közötti évekre szimulált kárbecslések valós megfigyeléseken alapulnak, ezért mindkét szcenárióra ugyanazok. A 2015-2019-es évek mediánjai lényegesen kisebbek az extrémnek számító 2020 és 2021-es évekéhez képest, viszont a 2025-től számolt összkár-eloszlások mediánjaitól nem térnek el jelentősen.

A hivatalos becslések ingadozása azzal magyarázható, hogy a károk nagy mértékben függnek a tüzek kialakulásának helyétől, így kisebb össz-tűzkiterjedésből is keletkezhetnek nagy károk és fordítva, nagyobb tüzekhez is tartozhatnak viszonylag kisebb károk. A modellek mediánja sok szimuláció középértéke, így itt természetes a kisebb ingadozás.

Szcenárió	SSP5				SSP2		
Modell	Logisztikus regresszió	Random Forest	XGBoost	Hivatalos becslés	Logisztikus regresszió	Random Forest	XGBoost
2015	13.40	14.66	16.82	4.71	13.40	14.66	16.82
2016	11.55	14.38	16.54	0.48	11.55	14.38	16.54
2017	13.93	14.63	17.42	18.01	13.93	14.63	17.42
2018	15.46	15.32	16.43	26.35	15.46	15.32	16.43
2019	12.17	13.60	14.15	0.16	12.17	13.60	14.15
2020	17.44	16.68	16.69	12.08	17.44	16.68	16.69
2021	19.36	15.30	15.65	-	19.36	15.30	15.65
2025	13.75	17.73	17.55	-	10.68	17.12	16.25
2026	12.72	16.03	17.01	-	11.08	16.04	15.96
2027	12.77	17.56	17.97	-	14.71	17.44	17.82
2028	10.68	16.52	15.73	-	13.03	17.08	17.07
2029	14.71	17.13	18.48	-	13.35	16.56	17.90
2030	16.04	18.41	18.95	-	12.45	16.53	16.11
2031	15.08	17.36	19.57	-	11.99	16.64	16.55
2032	12.58	17.50	17.63	-	10.41	15.64	14.06
2033	12.84	16.15	16.54	-	16.04	17.66	18.82
2034	12.30	16.09	16.47	-	11.74	15.49	15.88
2035	17.13	18.03	19.41	-	13.91	17.31	18.40
2036	12.31	17.12	17.14	-	12.59	15.97	16.93
2037	9.93	15.24	15.07	-	16.56	17.71	18.36
2038	9.66	16.08	15.11	-	12.29	16.43	16.88
2039	15.86	18.40	19.53	-	10.19	16.52	15.62
2040	12.77	17.74	16.94	-	13.32	16.55	17.47
2041	16.31	16.91	18.77	-	11.48	16.78	16.29
2042	12.21	16.57	16.73	-	9.77	16.41	16.01
2043	17.58	18.60	20.38	-	13.27	17.61	18.13
2044	12.40	17.12	16.85	-	11.05	15.94	15.36
2045	9.30	15.67	15.66	-	11.97	17.08	16.89
2046	13.06	16.30	17.05	-	17.31	18.73	19.77
2047	13.89	18.39	17.99	-	11.44	17.35	16.68
2048	15.52	17.20	18.65	-	15.69	17.02	18.33
2049	13.32	17.60	19.28	-	9.45	16.10	15.69
2025-2049 mediánja	12.84	17.13	17.55		12.29	16.64	16.88

24. ábra. Összesítő táblázat az összkár eloszlások mediánjairól (milliárd dollár)

7. Összefoglalás

A több adatforrásból származó komplex adatokat sikerült könnyen értelmezhetőre, valamint a modellek által is használható formára hozni. A modellezési folyamat összes tervezett része létrejött és a dolgozat során megírt kódbázis is hatékonyan használható, továbbá akár más vizsgált területre és tetszőleges klímaszcenáriókra könnyen általánosítható.

A modellek a tűzveszélyesség előrejelzésében az AUC (> 0.9) szerint éves szinten jó teljesítményt értek el, a veszélyesség mintázata nagyon hasonló a *CALFIRE* hivatalos tűzveszélyességi előrejelzésével. Ezért a modellek tűzveszélyesség klasszifikálásra jól használhatóak.

A kárbecslések értéke a valószínűség-profilok túlbecslése miatt lényegesen nagyobb értéket mutat a validációs és a teszt adathalmazon egyaránt, mint a hivatalos becslések. Ez a többletérték a nem teljes tűzadatbázis és az egy tűzre jutó átlagos kárbecslés pontatlanságával is magyarázható.

Klímaszcenáriók szerinti projekciók éves összkárbecslésénél az időjárás-trend változások hatása nem olvasható ki a modellek predikcióiból, amely az időjárás változók alacsony felbontása és a topografikus és antropogenikus változók statikus jellegéből és modellbeli fontosságából következik.

A végső eredmények nem mutatják az előzetesen várt látványosan emelkedő trendeket, ennek lehetséges magyarázata, hogy a össz-tűzkiterjedés nem növekedhet korlátlanul, és Kaliforniában már a jelenkori helyzet is nagyon súlyosnak mondható.

7.1. Továbbfejlesztési lehetőségek

– A célváltozó konstuálásakor nem vettem figyelembe az adott tűzeset időbeli változását, amely a területi autokorreláció miatt a modelleknél valószínűség-profil túlbecslést eredményezhetnek. Ezt a problémát műholdképek alapján, tűz- vagy füstklasszifikációs gépi tanulási technikákkal kezelni lehet. Ezáltal az egyes adathalmazok időlépésköze is csökkenthető akár heti vagy napi szintre.

– Adott tűzeset vége után a területen a növényzet regenerálásig egyáltalán nem éghető, ez szintén műholdkép klasszifikáció segítségével monitorozható és ezzel a vegetációs változók dinamikussá tehetőek.

– A modellek a teljes vizsgált területről vett véletlen mintavételezett tanuló adathalmazokon tanultak, emiatt a lokális hatásokat kevésbe tudják figyelembe venni. Kisebb területekre lebontva lokálisan tanított modellek súlyozott átlagaként (krigeléshez hasonlóan) számolt predikciók lehetséges, hogy lokálisan és ezáltal globálisan is pontosabb mintázatokhoz vezetnek.

Hivatkozások

[1] Basel Committee on Banking Supervision. *Climate-related financial risks – measurement methodologies*. 2021. URL: https://www.bis.org/bcbs/publ/d518.pdf.

[2] Emanuele Campiglio, Louis Daumas, Pierre Monnin és Adrian von Jagow. "Climate-related risks in financial assets". *Journal of Economic Surveys* (2022). DOI: https://doi.org/10.1111/joes.12525.

[3] Keywan Riahi, Detlef P. van Vuuren, Elmar Kriegler és Brian O'Neill. *The Shared Socio-Economic Pathways (SSPs): An Overview*. URL: https://unfccc.int/sites/default/files/part1_iiasa_rogelj_ssp_poster.pdf.

[4] Sandra Oliveira, Jorge Rocha és Ana Sá. "Wildfire risk modeling". *Current Opinion in Environmental Science & Health* 23 (2021), 100274. old. ISSN: 2468-5844. DOI: https://doi.org/10.1016/j.coesh.2021.100274. URL: https://www.sciencedirect.com/science/article/pii/S2468584421000465.

[5] K.L Pew és C.P.S Larsen. "GIS analysis of spatial and temporal patterns of human-caused wildfires in the temperate rain forest of Vancouver Island, Canada". *Forest Ecology and Management* 140.1 (2001), 1–18. old. ISSN: 0378-1127. DOI: https://doi.org/10.1016/S0378-1127(00)00271-1.

[6] Marj Tonini, Mirko D'Andrea, Guido Biondi, Silvia Degli Esposti, Andrea Trucchia és Paolo Fiorucci. "A Machine Learning-Based Approach for Wildfire Susceptibility Mapping. The Case Study of the Liguria Region in Italy". *Geosciences* 10.3 (2020). ISSN: 2076-3263. DOI: 10.3390/geosciences10030105.

[7] Gareth James, Daniela Witten, Trevor Hastie és Robert Tibshirani. *An Introduction to Statistical Learning*. New York: Springer, 2013.

[8] Steiner Ferenc. *A geostatisztika alapjai*. 1990.

[9] *California Wildfire Statistics*. URL: https://www.fire.ca.gov/our-impact/statistics.

[10] *LANDFIRE. (2023, April) Homepage of the LANDFIRE Project, U.S. Department of Agriculture, Forest Service; U.S. Department of Interior. Available:* URL: http://www.landfire.gov/index.php.

[11] USDA Forest Service National Forest System Lands GIS és Fire personnel. *National USFS Final Fire Perimeter*. URL: https://data.fs.usda.gov/geodata/edw/datasets.php?xmlKeyword=S_USA.FinalFirePerimeter.

[12] Jay H. Lawrimore, Ron Ray, Scott Applequist, Bryant Korzeniewski és Matthew J. Menne. *Global Summary of the Month (GSOM)*. URL: https://doi.org/10.7289/V5QV3JJ5.

[13] University of Idaho Dr. John Abatzoglou. *U.S. Gridded Palmer Drought Severity Index (PDSI) from gridMET*. URL: https://www.drought.gov/data-maps-tools/us-gridded-palmer-drought-severity-index-pdsi-gridmet.

[14] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer és K. E. Taylor. "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization". *Geoscientific Model Development* 9.5 (2016), 1937–1958. old. DOI: 10.5194/gmd-9-1937-2016. URL: https://gmd.copernicus.org/articles/9/1937/2016/.

[15] *Copernicus Climate Change Service, Climate Data Store, (2021): CMIP6 climate projections*. 20230515. verzió. 2021. DOI: 10.24381/cds.c866074c.

[16] Gokhan Danabasoglu. *NCAR CESM2 model output prepared for CMIP6 ScenarioMIP ssp245*. 20230515. verzió. 2019. DOI: 10.22033/ESGF/CMIP6.7748. URL: https://doi.org/10.22033/ESGF/CMIP6.7748.

[17] Gokhan Danabasoglu. *NCAR CESM2 model output prepared for CMIP6 ScenarioMIP ssp585*. 20230515. verzió. 2019. DOI: 10.22033/ESGF/CMIP6.7768. URL: https://doi.org/10.22033/ESGF/CMIP6.7768.

[18] Aiguo Dai. *Dai Global Palmer Drought Severity Index (PDSI)*. Boulder CO, 2017. URL: https://doi.org/10.5065/D6QF8R93.