Technická univerzita v Košiciach

Fakulta elektrotechniky a informatiky

Využitie metód strojového učenia pre vesmírny výskum

Bakalárska práca

2020 **Matej Varga**

Technická univerzita v Košiciach

Fakulta elektrotechniky a informatiky

Využitie metód strojového učenia pre vesmírny výskum

Bakalárska práca

Študijný program:	Inteligentné systémy
Študijný odbor:	Informatika
Školiace pracovisko:	Katedra kybernetiky a umelej inteligencie (KKUI)
Školiteľ:	doc. Ing. Peter Butka, PhD.
Konzultant:	RNDr. Šimon Mackovjak, PhD.

Košice 2020 **Matej Varga**

Abstrakt v SJ

Vo vrchných vrstvách zemskej atmosféry sa emituje elektromagnetické žiarenie známe ako airglow. Cieľom tejto práce je návrh a realizácia procesu tvorby modelu na predikciu intenzity airglow-u. Poskytuje prehľad základných charakteristík airglow-u, opisuje kroky potrebné pre získanie relevantných dát a proces ich pochopenia a spracovania. Následne sa zaoberá návrhom experimentov a ich realizáciou. Použité algoritmy strojového učenia sú popísané po teoretickej stránke – funkcionalita a princíp učenia, ale aj po praktickej stránke - voľba vhodnej podmnožiny dát a parametrov. Nakoniec je na základe zvolenej metriky určený najúspešnejší model.

Kľúčové slová

strojové učenie, predikcia, airglow, vesmírny výskum

Abstrakt v AJ

The Earth's upper atmosphere is a region where electromagnetic radiation known as airglow is emitted. The aim of this thesis is the design and implementation of the process of creating a model for predicting the intensity of airglow. It provides an overview of the basic characteristics of airglow, describes the steps needed to obtain relevant data and the process of data understanding and preparation. Subsequently, it deals with the design of experiments and their implementation. The machine learning algorithms are described from both the theoretical point of view – functionality and principle of learning, and also from the practical point of view – selection of a suitable subset of data and parameters. Finally, the most successful model is determined based on the selected metric.

Kľúčové slová v AJ

machine learning, prediction, airglow, space research

58836

TECHNICKÁ UNIVERZITA V KOŠICIACH
FAKULTA ELEKTROTECHNIKY A INFORMATIKY
Katedra kybernetiky a umelej inteligencie

ZADANIE BAKALÁRSKEJ PRÁCE

Študijný odbor: **Informatika**
Študijný program: **Inteligentné systémy**

Názov práce:

Využitie metód strojového učenia pre vesmírny výskum

Application of machine learning techniques for space science

Študent: **Matej Varga**
Školiteľ: **doc. Ing. Peter Butka, PhD.**
Školiace pracovisko: **Katedra kybernetiky a umelej inteligencie**
Konzultant práce: **RNDr. Šimon Mackovjak, PhD.**
Pracovisko konzultanta: **Ústav experimentálnej fyziky SAV**

Pokyny na vypracovanie bakalárskej práce:

1. Podať teoretický prehľad problematiky žiarenia atmosféry (tzv. airglow) a jeho predikcie, ako aj základov potrebných pre proces objavovania znalostí a vybraných metód strojového učenia použitých pre predikciu takýchto javov.
2. Vybrať vhodné dátové množiny meraní a predspracovať dáta pre riešenie zvolenej úlohy predikcie, navrhnúť postup realizácie experimentov a ich vyhodnotenia.
3. Realizovať experimenty a vyhodnotiť vytvorené modely predikcie na zvolenej množine dát.
4. Vypracovať dokumentáciu podľa pokynov katedry a vedúceho práce (hlavná časť 30-40 strán, prílohy - používateľská a systémová príručka, DVD s textami, obrázkami a softvérovými výstupmi aplikácie; tlačená forma v nerozoberateľnej väzbe).

Jazyk, v ktorom sa práca vypracuje: slovenský
Termín pre odovzdanie práce: 29.05.2020
Dátum zadania bakalárskej práce: 31.10.2019

prof. Ing. Liberios Vokorokos, PhD.
dekan fakulty

Čestné vyhlásenie

Vyhlasujem, že som bakalársku prácu vypracoval(a) samostatne s použitím uvedenej odbornej literatúry.

Košice 29. 5. 2020 .....

Vlastnoručný podpis

Poďakovanie

Týmto by som sa chcel úprimne poďakovať doc. Ing. Petrovi Butkovi, PhD. a RNDr. Šimonovi Mackovjakovi, Phd. za možnosť pracovať na tejto bakalárskej práci, za ich odborné rady, ktoré ma mnohému naučili a predovšetkým za ich čas, ktorý mi venovali pri vypracovávaní tejto práce.

Obsah

Zoznam obrázkov

Zoznam tabuliek

Úvod

Vesmír od nepamäti vzbudzuje v ľuďoch zvedavosť a túžbu objavovať jeho záhady. Od pozorovaní jednoduchými ďalekohľadmi sme hlavne v minulom storočí vďaka mnohým technickým objavom prešli k systematickému a precíznemu skúmaniu.

Na Zemi je množstvo observatórií, ktoré produkujú ohromné počty dát. Nehovoriac o výskumných satelitoch, ktorých taktiež na orbite pribúda, s cieľom zistiť viac o fungovaní vesmíru. Ručné spracovanie všetkých týchto údajov je problematické, a preto je potrebné hľadať spôsoby aspoň čiastočnej automatizácie.

Vo vesmíre tiež prebieha množstvo javov, ktoré majú väčší, či menší vplyv na Zem a život na nej. Ľudstvo závisí od elektronických zariadení, ktoré však naše Slnko dokáže svojou aktivitou znefunkčniť. Preto má zmysel snažiť sa tieto javy predvídať a zistiť viac o ich vzniku.

Vhodným nástrojom na tieto úlohy sú algoritmy strojového učenia. Strojové učenie je rýchlo sa rozvíjajúce dynamické odvetvie, ktoré si už našlo uplatnenie aj v astronómií. Všetky možnosti využitia však zďaleka neboli preskúmané.

Jednou z takýchto možností je predikcia svetelného žiarenia v horných vrstvách atmosféry (tzv. airglow). Ten vzniká ako dôsledok slnečného UV žiarenia, ktoré dopadá na našu planétu počas dňa. Airglow je veľmi dobrým indikátorom procesov vrchných vrstiev atmosféry. Taktiež sa dá využiť aj na nepriame skúmanie slnečnej aktivity, štúdium vysoko-energetických častíc, ako aj pri zabezpečení priepustnosti satelitných signálov, ktorá môže byť narušená práve kvôli procesom v týchto vrstvách.

Cieľom našej práce bude pokúsiť sa vybrať vhodné dátové množiny meraní a s ich pomocou navrhnúť model intenzít airglow-u, ktorý by bol prínosom aj pri iných výskumoch, napríklad aj pri predpovedi kozmického počasia.

1 Formulácia úlohy

Cieľom tejto práce je s použitím metód strojového učenia vytvoriť model intenzít svetelného žiarenia hornej atmosféry. Pri jeho vytváraní je potrebné splniť nasledujúce požiadavky:

- predložiť základné charakteristiky svetelného žiarenia horných vrstiev atmosféry, spolu s vysvetlením jeho vzniku
- podať teoretický prehľad algoritmov strojového učenia vhodných pre takýto typ predikčného problému
- nájsť relevantné zdroje dátových množín a pripraviť dáta do požadovaného formátu
- navrhnúť priebeh experimentov spolu s vhodnými metrikami na ich hodnotenie
- zrealizovať tvorbu modelov na pripravenej množine dát
- vypracovať kompletnú dokumentáciu priebehu výskumu

2 Úvod do problematiky

2.1 Airglow - žiarenie hornej atmosféry Zeme

Väčšina ľudí predpokladá, že na dostatočne tmavom mieste ďaleko od civilizácie počas bezmesačnej noci na oblohe vidieť iba hviezdy a galaxie, ale nie je tomu tak. Okrem nich je môžné pozorovať aj málo intenzívnu žiaru v rôznych farebných odtieňoch, z ktorých prevažujú hlavne zelené a červené.

Slnko produkuje elektromagnetické žiarenie s rôznymi vlnovými dĺžkami, ktoré nepretržite bombarduje zemskú atmosféru. Jedným z nich je ultrafialové žiarenie, ktoré počas dňa rozdeľuje molekuly kyslíka a dusíka, a tak spúšťa reťazec ďalších komplexných chemických reakcií, ktoré môžu viesť až k vzniku nových molekúl ako je napríklad ozón. Po západe Slnka prestáva toto žiarenie priamo pôsobiť, avšak molekuly sa naďalej podieľajú na chemických reakciách, pri ktorých môže dochádzať k vyžiareniu svetla. Tento proces sa nazýva chemiluminiscencia a hrá dôležitú úlohu pri vzniku málo intenzívnej žiary na nočnej oblohe nazývanej airglow.

Pri pozorovaní zo Zeme je airglow najlepšie viditeľný na začiatku noci. Častokrát je to len tenký farebný pás tesne nad horizontom, ktorý je ťažko pozorovateľný voľným okom. V skutočnosti tento úkaz prebieha na veľkej časti oblohy, lenže zachytiť ho dokážu iba veľmi citlivé detektory. Farebnosť a jas airglow-u sa líši v závislosti od miesta a času. Je to spôsobované mnohými rôznymi faktormi, ktoré na neho vplývajú. Aj keď väčšinu procesov stojacich za vznikom tohto úkazu dokážeme fyzikálne vysvetliť, stále sú aj také, ktoré nám nie sú úplne jasné. Keďže airglow je dôsledok ultrafialového žiarenia emitovaného zo Slnka, tak zmeny v slnečnej aktivite majú zásadný dopad na jeho intenzitu.

Snímky airglowu sa môžu zdať veľmi zaujímavé aj pre laika, avšak toto žiarenie môže taktiež spôsobovať problémy pri astronomických pozorovaniach, pri ktorých pôsobí rušivo hlavne počas vytvárania dlhých expozícií. Toto nežiaduce pozadie je

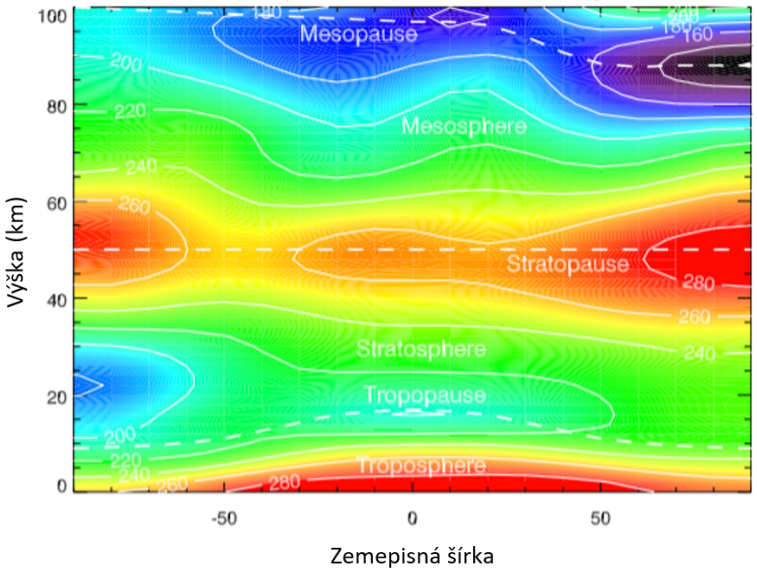
potom nutné odstrániť, čo môže byť tiež problematické kvôli tomu, že airglow vytvára pomerne dynamické štruktúry, meniace sa v rozpätí niekoľkých minút. Okrem estetického má airglow ale aj vedecké využitie. Odzrkadľujú sa v ňom procesy prebiehajúce vo vrchných vrstvách atmosféry, ako napríklad vplyv kozmického počasia na našu planétu, ale taktiež sa z neho dá nepriamo študovať aj slnečná aktivita. Naviac, môžu byť dáta o intenzite airglow-u vedľajším produktom iných astronomických pozorovaní, čo uľahčuje zbieranie takýchto dát, aj keď sú náročnejšie na čistenie a spracovanie.

2.2 Základné charakteristiky airglow-u

Žiarenie atmosféry Zeme sa delí do dvoch hlavných kategórií – termálne a netermálne žiarenie. Prvé menované môže byť vyvolané zrážkami medzi zložkami atmosféry, teda prenosom energie spojenej s tepelným pohybom molekúl a atómov. Príkladom je polárna žiara (aurora). Netermálne žiarenie má svoj pôvod v excitovaných stavoch atómov a molekúl, ktorých excitáciu nie je možné dosiahnuť len zrážkami. Je potrebný dodatočný zdroj energie. Týmto zdrojom je už vyššie spomínané elektromagnetické žiarenie zo Slnka s vlnovou dĺžkou kratšou ako 242 nm. Pojem airglow pokrýva všetky typy netermálneho žiarenia, či už má atomárny alebo molekulárny pôvod.

V závislosti od výšky Slnka nad horizontom delíme airglow na dayglow (keď je Slnko nad horizontom), twilightglow (Slnko je už pod horizontom, ale stále sú osvetľované vrchné vrstvy atmosféry) a nightglow (astronomická noc, teda Slnko je nižšie ako 18° pod obzorom). Vo všeobecnosti je dayglow najjasnejším druhom airglow-u, keďže vzniká pri priamom osvetľovaní atmosféry v celej jej výške. Avšak rozptyl slnečného svetla je o niekoľko rádov intenzívnejší ako dayglow, preto je ho náročnejšie pozorovať ako nightglow. V našej práci sa ale budeme venovať len nightglow-u (aj keď na jeho označenie budeme používať všeobecné pomenovanie airglow).

Atmosféra sa delí na rôzne vrstvy v závislosti od výšky nad zemským povrchom. Na obrázku 2 – 1 je znázornená zmena telpoty v jednotlivých vrstvách v júli. Pre troposféru a mezosféru je charakteristické znižovanie teploty so zvyšujúcou sa výškou, na rozdiel od stratosféry a termosféry, v ktorých teplota stúpa spolu s výškou. Tropopauza oddeľuje troposféru od stratosféry, tak ako stratopauza rozdeľuje stratosféru a mezosféru. Medzi mezosférou a termosférou sa nachádza mezopauza. Airglow je jav vznikajúci vo vrchných vrstvách atmosféry, konkrétne v mezosfére (výšky od 50 km do 100 km) a termosfére (výšky nad 100 km). V skutočnosti významná časť našich poznatkov o chemických reakciách prebiehajúcich v týchto vrstvách pochádza práve z pozorovania airglow-u (Savigny, 2017).

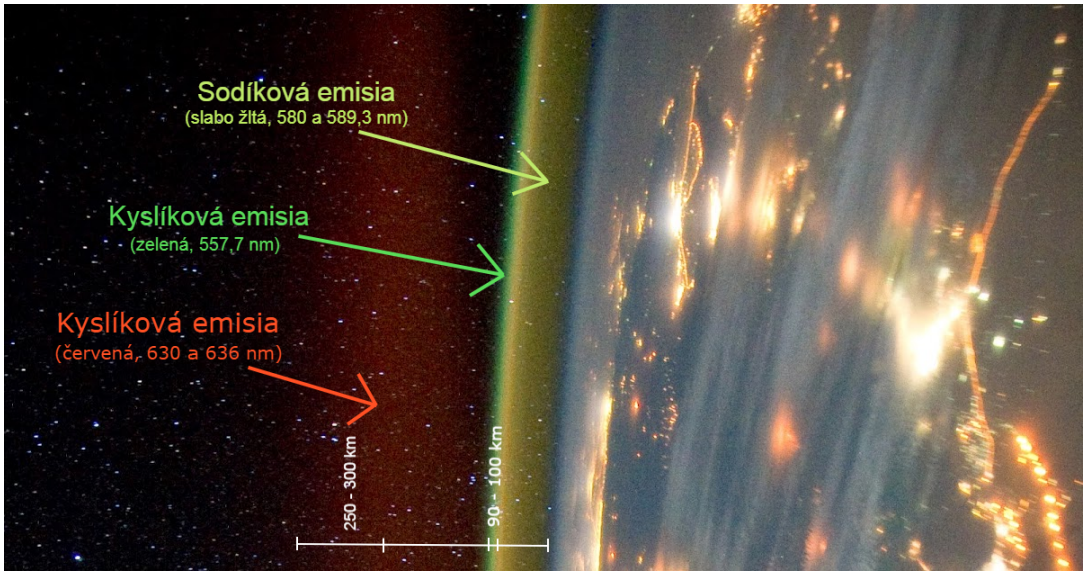


Obrázok 2 – 1 Teplota jednotlivých vrstiev atmosféry v Kelvinoch v mesiaci júl. Zdroj: (Savigny, 2017)

Airglow si netreba zamieňať s polárnou žiarou, ktorá vzniká pôsobením vysoko energetických častíc zo Slnka (prevažne elektrónov), ktoré prenikajú do spodnej termosféry a mezosféry počas geomagnetických búrok. Avšak niektoré druhy emisií sú prítomné v obidvoch úkazoch (napríklad zelená kyslíková čiara – 557.7 nm, ktorej sa budeme venovať nižšie). Airglow je globálne vyskytujúci sa, nepretržitý a väčši-

nou homogénny jav. Na rozdiel od polárnej žiary, ktorej výskyt je časovo neurčitý, typicky vytvára na oblohe charakteristické štruktúry a je pozorovateľná výhradne v polárnych oblastiach (Savigny, 2017).

Airglow je najlepšie viditeľný z ISS (viď. obrázok 2 – 2), respektíve zo zemskej orbity, pretože pri bočnom pohľade na vrstvu atmosféry človek vidí až 100-násobne dlhší úsek ako pri vertikálnom pohľade z povrchu. Ak by sa astronauti vybrali k inej planéte s atmosférou, pravdepodobne by videli veľmi podobný úkaz, pretože airglow bol objavený tiež v atmosfére Marsu aj Venuše (Slanger et al., 2001). Tento bočný pohľad z vesmíru je užitočný aj v tom, že jasne vidno jednotlivé vrstvy, z ktorých sa airglow skladá.



Obrázok 2 – 2 Pohľad na vrstvy airglow-u z ISS. Zdroj: NASA

Existujú dva významné dôvody, prečo airglow tvorí vrstvy a nie je homogénny vzhľadom na výšku nad povrchom. Prvým je fakt, že počet fotochemických reakcií sa zvyšuje so znižujúcou sa výškou, pretože sa zvyšuje hustota atmosféry. Druhým dôvodom je, že deexcitácia atómov a molekúl zrážkami je častejšia v nižších nadmorských výškach, čo je taktiež zapríčinené väčšou hustotou. V konečnom dôsledku je emitovanie airglow-u najsilnejšie v určitej strednej výške, kde nastáva kompromis

medzi týmito dvoma javmi (Savigny, 2017).

Najznámejšie vrstvy airglow-u sú (Ghodpage, 2016):

- červená O vrstva (vlnová dĺžka 630 nm) – emisné maximum je vo výške okolo 250 km
- zelená O vrstva (vlnová dĺžka 557,7 nm) – emisné maximum okolo 95 km
- žltá Na vrstva (vlnová dĺžka 589,3 nm) – emisné maximum v 92 km nad zemským povrchom
- OH vrstva (infračervená spektrálna oblasť) – emisné maximum dosahuje vo výške okolo 85 km

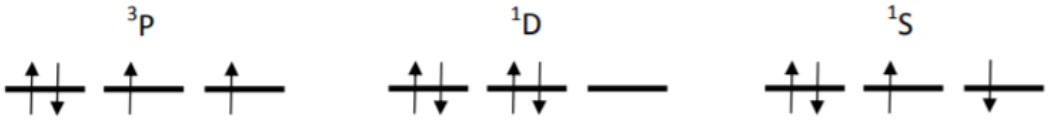
Atomárny kyslík, ktorého životnosť v hornej mezosfére dosahuje aj 1 mesiac, plní hlavnú úlohu pri vzniku airglow-u s vlnovými dĺžkami 630 nm a 557,7 nm. V našej práci práve tieto dve opíšeme podrobnejšie.

2.3 Žiarenie atomárneho kyslíka

Typy airglow-u, na ktorých vzniku sa podieľa atomárny kyslík, sú najštudovanejšie spomedzi všetkých v zemskej atmosfére a sú taktiež prvé, ktoré boli objavené. Pred vysvetlením ich vzniku je vhodné spomenúť možné elektrónové konfigurácie atómu kyslíka.

Atomárny kyslík má celkovo 8 elektrónov a jeho základná elektrónová konfigurácia je $1s^2 2s^2 2p^4$. Nakoľko $2p$ orbitály sú iba čiastočne zaplnené, tak existujú 3 rôzne spôsoby, ako usporiadať elektróny v $2p_x$, $2p_y$ a $2p_z$ orbitáloch. Tieto možnosti sú znázornené na obrázku 2 – 3 (Ho et al., 1995).

Podľa Hundovho pravidla maximálnej multiplicity, základný stav elektrónovej konfigurácie zodpovedá stavu orbitálov s najvyššou multiplicitou. Multiplicita sa



Obrázok 2 – 3 Elektrónové konfigurácie $2p$ orbitálov atómu kyslíka.

vypočíta jednoduchým vzorcom 2S+1, kde S je súčet spinov elektrónov, pričom spin elektrónu môže nadobúdať hodnoty +/- 1/2. Preto základným stavom je ^3P konfigurácia, ktorá má multiplicitu s hodnotou 3. Zvyšné dve konfigurácie majú multiplicitu 1.

Druhé Hundovo pravidlo hovorí, že pri rovnakej multiplicite má najmenšiu energiu konfigurácia s najväčšou hodnotou vedľajšieho kvantového čísla L. Keďže druhá konfigurácia (^1D) má vyššiu hodnotu L, tak jej energia je nižšia. Preto konfigurácia ^1D je prvým excitovaným stavom kyslíka a ^1S je druhým excitovaným stavom (Ho et al., 1995).

Atomárny kyslík, ktorý sa nachádza v základnom 3P stave, je excitovaný na vyššie energetické stavy pomocou slnečného UV žiarenia. Takto vzbudený kyslík sa vracia do základného stavu v dvoch krokoch. Prvý prechod so životnosťou 0,84 s produkuje zelené svetlo (557,7 nm) a druhý so životnosťou 114 s vyprodukuje červené svetlo (630,0 nm).

Všeobecne akceptovaným excitačným mechanizmom O(^1S) stavu je chemiluminiscenčný proces, známy ako Barthova prenosová schéma, ktorá začína rekombináciou dvoch atómov kyslíka.

$$O(^3P) + O(^3P) + M \longrightarrow O_2^* + M$$

Po ňom nasleduje prenos energie.

$$O(^3P) + O_2^* \longrightarrow O(^1S) + O_2$$

Následne je mechanizmus zakončený prechdom z O(^1S) stavu do stavu O(^1D), pri ktorom je vyžiarený fotón s vlnovou dĺžkou 557,7 nm. Týmto spôsobom sa tvorí zelená kyslíková vrstva airglow-u.

$$O(^1S) \longrightarrow O(^1D) + hv(\lambda = 557.7nm)$$

Pre červenú kyslíkovú vrstvu sú dôležité nasledujúce reakcie.

$$O^+ + O_2 \longrightarrow O_2^+ + O(^3P)$$

$$O_2^+ + e^- \longrightarrow O(^1D) + O(^3P)$$

Vyžiarenie fotónu s vlnovou dĺžkou 630 nm sa udeje nasledovne.

$$O(^1D) \longrightarrow O(^3P) + hv(\lambda = 630nm)$$

Zhrnutie všetkých podstatných prechodov medzi jednotlivými elektrónovými konfiguráciami atomárneho kyslíka, ktoré majú vplyv na vznik airglow-u, sú znázornené na obrázku 2 – 4.

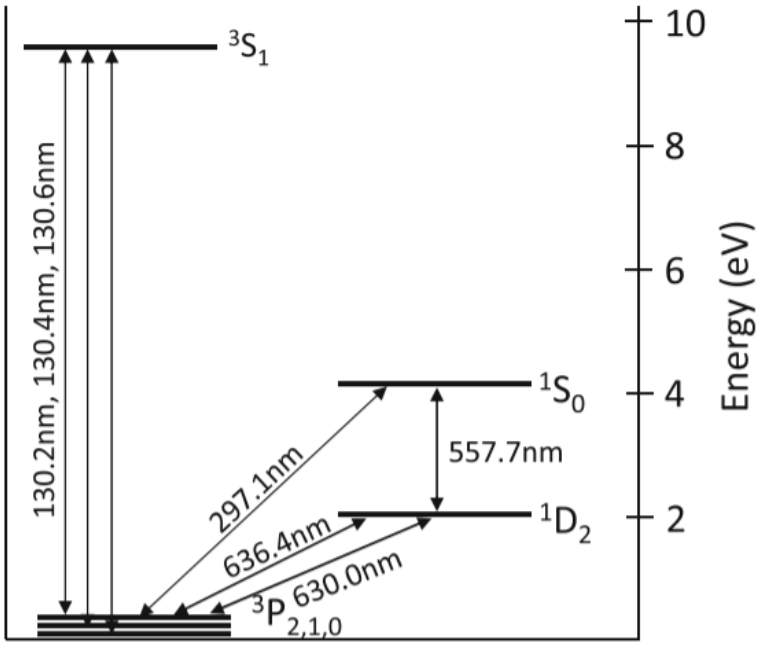
Základna fyzikálna veličina pri výskume airglow-u je tzv. VER (skratka z volume emission rate), ktorá vyjadruje počet emitovaných fotónov z jednotky objemu za jednotku času (Savigny, 2017).

$$[VER] = fotón * cm^{-3} * s^{-1}$$

Avšak VER obyčajne nemôže byť meraný priamo. Pozemné meriacie prístroje, tak ako aj satelity, merajú IER (skratka z integrated emission rate), čo je integrál VER podľa zornej línie daného meracieho prístroja. IER sa vypočíta podľa vzorca:

$$IER = \int_{z=0}^{z_{TOA}} VER(z')\ dz',$$

kde z je výška, z_{TOA} je horná výška atmosféry, v ktorej je už intenzita meraného airglow-u zanedbateľná (Savigny, 2017).



Obrázok 2 – 4 Elektrónové energetické úrovne atomárneho kyslíka podstatné pre emisie airglow-u. Zdroj: (Savigny, 2017)

Jednotkou IER je 1 Rayleigh, čo je jednotka definovaná nasledovne:

$$1R = 10^6 fotónov * cm^{-2} * s^{-1}$$

3 Prehľad technológií

3.1 História umelej inteligencie

Idea mysliacich strojov, ktoré by boli svojimi rozhodnutiami schopné napodobniť činnosť a uvažovanie človeka, je oveľa staršia ako počítače samotné. Za jeden z prvých pokusov takejto imitácie sa považuje robot-rytier Leonarda Da Vinci, ktorý dokázal stáť, sedieť, hýbať rukami, a to už okolo roku 1495 len za pomoci kladiek a lán (Moran, 2006).

Prvé reálnejšie a skutočne automatizované prístupy sa objavili s príchodom výpočtovej techniky po 2. svetovej vojne, kedy boli formulované aj prvé kritéria, ktoré by umelá inteligencia mala spĺňať. Jedným z najznámejších kritérií je tzv. Turingov test (Oppy and Dowe, 2019), ktorý mal rozhodnúť, či je stroj inteligentný, a to len na základe odpovedí na kladené otázky. Test prebiehal tak, že jeden človek v úlohe pýtajúceho kládol otázky dvom respondentom, ktorých nevidel. Na základe odpovedí mal potom rozhodnúť, ktorý odpovedjúci je človek a ktorý je stroj. Ak to nedokázal určiť, tak bol stroj považovaný za inteligentný.

Aj keď v tej dobe bolo vytvorenie umelej inteligencie z pohľadu technológií nemožné, rozprúdilo to aspoň teoretický výskum vo vedeckej komunite, ktorý neskôr upadol z dôvodu nemožnosti otestovania takýchto myšlienok. Zároveň to aj predurčilo vývoj umelej inteligencie, v ktorom vidíme periodické opakovanie sa období na výslní (obdobie „jari“), kedy sa do tejto oblasti vkladali nemalé investície, vznikali enormné očakávania zo strany verejnosti aj vedcov, a ktoré neskôr, kvôli svojej neopodstatnenosti, viedli k obdobiam úpadku (obdobie „zimy“), k všeobecnému sklamaniu a k zavrhnutiu tejto oblasti výskumu ako celku. Spomedzi viacerých spomenieme výročnú konferenciu AAAI (skratka z American Association of Artificial Intelligence) v roku 1984, na ktorej R. Schank a M. Minsky (dvaja poprední výskumníci v oblasti umelej inteligencie) varovali obchodnú komunitu, že nadšenie z umelej inteligencie sa vymklo spod kontroly a pravdepodobne bude nasledované sklamaním.

Tri roky na to sa ich slová naplnili a odvetvie AI sa začalo rúcať.

V dnešnej dobe (približne od roku 2010) sa podľa viacerých expertov (Buhgin, 2017; Olhede and Wolfe, 2018) nachádzame v novej jari umelej inteligencie, ktorá môže, ale nemusí, byť nasledovaná ďalším úpadkom. V prospech tvrdenia, že umelá inteligencia bude naďalej rásť (alebo minimálne nebude upadať), hovorí aj fakt, že takéto systémy a algoritmy už prenikli do nášho každodenného života. Či hľadáme najkratšiu a najrýchlejšiu cestu do nášho cieľa, prezeráme si ponuku internetového obchodu s oblečením, snažíme sa preložiť text z cudzieho jazyka, alebo len počúvame hudbu cez YouTube, prichádzame často bez nášho vedomia do kontaktu s umelou inteligenciou alebo presnejšie s jej časťou nazývanou strojové učenie.

Pojem „strojové učenie“ ako prvý použil Arthur Samuel, americký priekopník v oblasti umelej inteligencie a počítačových hier, počas práce pre IBM v roku 1959 (Samuel, 1959). V tomto smere išlo hlavne o problém klasifikácie pomocou automaticky získaných znalostí z dát. Strojové učenie sa ako oddelená oblasť začalo rozvíjať v 90. rokoch 20. storočia. Hlavným cieľom tejto oblasti tak prestalo byť dosiahnutie všeobecnej umelej inteligencie, ale zamerala sa na riešenie čiastkových problémov s ohľadom na praktický charakter.

Zatiaľ čo dosiahnutie univerzálneho mysliaceho stroja je hudbou budúcnosti (ak je vôbec možné) (Joshi, 2019), na riešenie čiastkových úloh sa umelá inteligencia ukázala byť veľmi silný nástroj.

Ako je vyššie spomenuté, umelá inteligencia v dnešnej dobe zažíva svoj zlatý vek a je to spôsobené hlavne kombináciou troch faktorov, ktoré v minulosti nikdy nenastali súčasne.

Hlavným dôvodom je existencia Big Data (veľkých dát), ktoré sú nevyhnuté pri snahe vytvoriť samostatne sa učiaci systém. Ani ten najdokonalejší žiak sa nič nenaučí, ak nebude mať z čoho. Veľkým prínosom je internet, vďaka ktorému máme k

dispozícií množstvo voľne dostupných dát použiteľných na výskumné účely. Jediným problémom je nutnosť čistenia a predspracovania takýchto údajov, ale aj na tento problém existuje množstvo užitočného softvéru, ktorý bude predstavený aj v tejto práci.

Ďalším významným faktorom je rozvoj grafických procesorov (GPU), ktoré sa hlavne začiatkom tisícročia snažili ich výrobcovia rozšíriť aj medzi vedeckou komunitou s tým, že výrazne zväčšia výpočtovú silu a výkon. Tieto spoločnosti sa dnes zvyknú prezentovať ako tí, ktorí umožnili rozvoj strojového učenia. Určite je pravda, že takýto pokrok by bez nich nebol možný.

A posledným, tretím faktorom je množstvo investícií (finančných, ale aj personálnych a časových), ktoré sa vkladajú do výskumu strojového učenia. Hlavne veľké firmy, ako napríklad Google, Microsoft a Facebook, disponujú takými obrovskými množstvami dát, že bez týchto algoritmov by ich nedokázali spracovávať. Z tohto pohľadu je krok urobiť ich vlastné systémy a algoritmy verejne dostupné logický. Okrem nich sa na tvorbe nových postupov v strojovom učení podieľa aj nevedecká komunita nadšených programátorov, ktorá taktiež nemalým dielom prispela k úspešnosti rozvoja tohto odvetvia. Množstvo dnes používaných knižníc umelej inteligencie pre jazyk Python vzniklo práve z takejto iniciatívy. Vďaka týmto nástrojom je možné vytvárať samostatne sa učiace aplikácie bez znalosti štatistiky a matematiky, stačia základné programátorské zručnosti. To umožňuje aj výskumníkom z iných oblastí, ako je napríklad medicína, ekonomika, energetika alebo meteorológia, aplikovať tieto nové postupy aj na svoje odvetvie a urýchliť tak pokrok, alebo dokonca objaviť niečo celkom nové.

3.2 Strojové učenie pre vesmírny výskum

Výnimkou nie je ani vesmírny výskum, ktorý môžeme považovať za jeden z najstarších (prvé astronomické pozorovania a texty sa datujú do obdobia starovekej Číny,

2000 rokov p.n.l). Hlavne kvôli veľkému množstvu dát zo satelitov a observatórií po celej Zemi sa javí ako prirodzený kandidát na použitie strojového učenia. V praxi už fungujú systémy ako automatické vyhľadávanie exoplanét (Zucker and Giryes, 2018; Pearson et al., 2017) alebo detekcia gravitačných vĺn (George and Huerta, 2018), ktoré dokazujú, že ak sa použije na správny problém, vie strojové učenie ušetriť množstvo času.

Takéto aplikácie so sebou prinášajú aj isté problémy. Jedným z nich je príliš náročná, väčšinou až nemožná extrakcia znalostí z natrénovaných modelov. Pri zložitých úlohách s množstvom premenných a parametrov sa síce dokážu modely strojového učenia správne natrénovať a následne poskytovať vynikajúce výsledky, lenže nevieme, ako k tým výsledkom prišli a nedokážeme z nich extrahovať presné fyzikálne zákonitosti. Je to tzv. prístup black-box (čierna skrinka), pri ktorom model získa znalosti len z poskytnutých dát.

Opačným prístupom je white-box (biela skrinka), kde model pracuje priamo s fyzikálnymi vzorcami a simuláciami. Takýto model je pre ľudí zrozumiteľnejší, ale pravdepodobne nikdy nedosiahne pri niektorých problémoch požadovanú presnosť, pretože vo vesmíre na seba pôsobí veľa dejov súčasne v rôznych mierkach a časoch, ktorých simulácia by si vyžadovala enormné výpočtové nároky. Taktiež príčina a následok niektorých javov nasledujú v čase veľmi blízko za sebou.

Vhodným prístupom by mohol byť tzv. grey-box (sivá skrinka), ktorý je medzi predchádzajúcimi dvoma metódami návrhu modelov. Opiera sa o platné, časom overené fyzikálne zákonitosti a zároveň pomocou strojového učenia vylepšuje výsledok o znalosti, ktoré nám chýbajú, alebo majú príliš náhodný charakter.

3.3 Použité knižnice

V tejto podkapitole sa venujeme opisu softvérových nástrojov použitých v našej práci. Na analýzu dát a modelovanie bol použitý programovací jazyk Python (Van Ros-

sum and Drake, 2009). Je to vysokoúrovňový dynamický programovací jazyk, ktorý je vyvíjaný ako open-source projekt. Pri implementácií algoritmov strojového učenia je častou voľbou hlavne vďaka dobrému pomeru rýchlosť – úroveň abstrakcie, rozsiahlej komunite vývojárov a používateľov a v neposlednom rade kvôli množstvu užitočných knižníc s pravidelnými aktualizáciami. V našej práci sme použili knižnice Pandas, Scikit learn a Keras.

Knižnica Pandas (pandas development team, 2020) je určená na uľahčenie manipulácie s dátami a ich analýzu. Názov je odvodený od slovného spojenia „panel dát“. V mnohých veciach sa inšpirovala programovacím jazykom R, s cieľom stať sa najvýkonnejším dostupným analytickým nástrojom. Medzi hlavné podobnosti patrí ukladanie dát do dataframe-ov, jednoduché zaobchádzanie s chýbajúcimi údajmi, rýchle zlučovanie a rozdeľovanie aj veľkých tabuliek a efektívna správa pamäte. Pre našu prácu boli užitočné aj možnosti reorganizácie dát na iné časové jednotky a zoskupenie dát podľa určitých hodnôt.

Knižnica scikit learn (Pedregosa et al., 2011) alebo tiež sklearn obsahuje implementácie mnohých známych algoritmov strojového učenia a vďaka jednoduchosti používania je obľúbená ako vo vedeckej, tak v komerčnej oblasti. Ponúka širokú škálu algoritmov bez učiteľa, aj s učiteľom. Pri väčšine z nich je možnosť výberu medzi regresiou a klasifikáciou. Mnohé sú napísané v jazyku Cython (jazyk kombinujúci rýchlosť jazyka C so syntaxou Python-u), vďaka čomu sklearn dosahuje vyššie výpočtové rýchlosti v porovnaní s inými knižnicami zameranými na strojové učenie.

Keras (Chollet et al., 2015) je open-source knižnica zameraná na neurónové siete a hlboké učenie. Obsahuje mnohé, často používané stavebné bloky neurónových sietí, ako sú aktivačné funkcie, metriky, optimizátory a rôzne typy neurónových vrstiev. Umožňuje pomocou zopár riadkov kódu vytvoriť komplexný model, ktorý sa po natrénovaní dá ľahko exportovať a nasadiť aj na smartfónoch a edge zariadeniach.

Hlavnou výhodou je, že je používateľsky príjemný a vďaka modulárnemu prístupu umožňuje rýchlo vykonať akékoľvek zmeny modelu.

3.4 Regresná analýza

Regresná analýza je označenie štatistických metód používaných pre odhad vzťahov medzi závislou premennou a jednou alebo viacerými nezávislými premennými. Regresia je jedným z najviac aplikovaných štatistických postupov v astronómií (Isobe et al., 1990).

3.4.1 Lineárna regresia

Lineárna regresia je najjednoduchším algoritmom regresnej analýzy. Je používaná na kvantitatívne vyjadrenie lineárnej korelácie medzi dvoma alebo viacerými javmi. Pre vytvorenie modelu je potrebné určiť závislé a nezávislé premenné. Premenná, ktorú chceme modelovať, resp. predikovať, sa nazýva závislá. Ostatné premenné sa nazývajú nezávislé a podľa ich počtu bude výstupom lineárnej regresie priamka (1 nezávislá premenná) alebo rovina (2 a viac nezávislých premenných).

Princíp metódy spočíva v nájdení takej priamky (resp. plochy), ktorá je najbližšie k čo najväčšiemu počtu bodov (trénovacích dát). Vzťah pre lineárnu regresiu zapíšeme nasledovne:

$$Y = a + bX + \epsilon,$$

kde Y je závislá premenná, X je nezávislá premenná, a a b sú modelované koeficienty, ktoré sa v jednotlivých iteráciách učia. b vyjadruje sklon regresnej priamky, a vyjadruje odskok regresnej priamky od nulového bodu. Hodnota ϵ je náhodná chyba, ktorá zachytáva ostatné faktory, ktoré majú vplyv na Y a sú nezávislé od X (Pant, 2019).

Cieľom trénovania je, aby bola chybová funkcia čo najmenšia, najčastejšie sa používa stredná kvadratická chyba (mean squared error), ktorá pomocou predikovanej

a skutočnej hodnoty vypočíta ich štvorcový rozdiel. Vzťah je daný nasledovne:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i')^2,$$

kde n je počet predikcií, Y je vektor skutočných hodnôt a Y‘ je vektor predikovaných hodnôt.

Hodnoty koeficientov a a b sa veľmi často vypočítavajú metódou znižovania gradientu (Gradient descent). Táto metóda začína s náhodnými hodnotami koeficientov, pre ktoré vypočíta pomocou parciálnych derivácií sklon chybovej (hodnotiacej) funkcie a určí smer, v ktorom ich treba meniť, aby sa chyba znížila. Na základe toho sa upravia koeficienty a znova sa vypočíta sklon. Tento proces sa iteračne opakuje, kým sa nedosiahne požadovaná presnosť (Pandey, 2019). Dôležité je taktiež zvoliť správnu hodnotu učiaceho parametra, pretože, ak je príliš malý, tak učenie trvá dlho, no ak je príliš veľký, tak metóda nemusí dosiahnuť požadovanú presnosť.

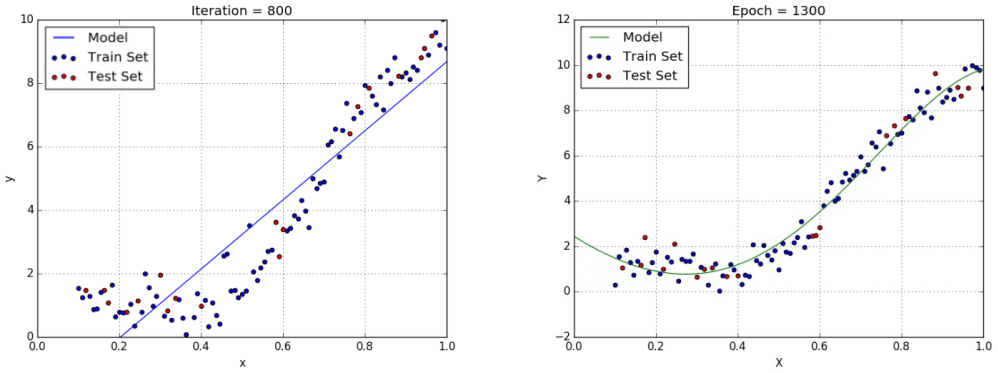
3.4.2 Polynomiálna regresia

Polynomiálna regresia, na rozdiel od lineárnej, dokáže aproximovať aj nelineárne korelácie (viď. obrázok 3 – 1). Je to forma regresnej analýzy, v ktorej je vzťah medzi závislou premennou a nezávislou premennou (premennými) modelovaný ako polynóm n-tého stupňa, teda výstupom je krivka (1 nezávislá premenná) prípadne plocha (2 a viac nezávislých premenných). Všeobecný vzťah vieme zapísať nasledovne (Pant, 2019):

$$Y = a + bX + cX^2 + dX^3 + ...X^n + \epsilon,$$

kde a, b, c, d... sú koeficienty polynómu, Y je závislá premenná, X je nezávislá premenná a ϵ je náhodná chyba. V prípade, že máme viac nezávislých premenných, do polynómu sa pridajú aj ich kombinácie (ak takáto kombinácia nepresahuje najvyšší stupeň polynómu). Spôsob učenia je podobný ako pri lineárnej regresii, teda využíva sa metóda znižovania gradientu s tým rozdielom, že v prípade viacerých koeficientov polynómu sa hľadá minimum vo viacrozmernom priestore, pričom počet rozmerov priestoru je rovnaký ako počet koeficientov polynómu.

Najčastejšie využívaná je polynomiálna regresia druhého stupňa (kvadratická regresia) a polynomiálna regresia tretieho stupňa. Pri vyšších stupňoch je tento algoritmus náchylný na preučenie, pretože sa ľahko naučí aproximovať aj šum, ktorý sa často nachádza v reálnych dátach. Polynomiálna regresia je taktiež citlivá na osamotené dáta, ktorých odchýlka od zvyšných dát je príliš veľká a sú potencionálne chybné (tzv. outliers). Takéto anomálie často vznikajú chybou počas merania alebo spracovania dát.



Obrázok 3 – 1 Porovnanie učenia Lineárnej Regresie (vľavo) a Polynomiálnej regresie (vpravo) na dáta s nelineárnou koreláciou. Zdroj: (Pant, 2019)

3.5 Neurónová sieť

Neurónová sieť je masívne paralelný procesor, schopný uchovávať znalosti a využívať ich pri riešení problémov. Jej vznik bol inšpirovaný biologickými štruktúrami v ľudskom mozgu, ktorý napodobňuje v dvoch aspektoch (Sinčák and Andrejková, 1996):

- poznatky sú zbierané v neurónovej sieti počas učenia
- medzineurónové spojenia (synaptické váhy) sú využívané na ukladanie znalostí.

Princíp jej fungovania spočíva v prepojení veľkého počtu umelých neurónov usporiadaných do vrstiev, medzi ktorými sa šíri signál podobne, ako sa šíri vzruch medzi

biologickými neurónmi. Štruktúra umelého neurónu je na obrázku 3 – 2 a skladá sa z nasledujúcich častí (Skalski, 2018):

- x_1 až x_n sú vstupy do neurónu. Ich počet závisí od počtu príznakov, ktorými je popísaný jeden prvok trénovacej (resp. testovacej) množiny, ak ide o neurón vstupnej vrstvy. Počet vstupov pre neurón skrytej vrstvy je rovný počtu neurónov na vrstve predchádzajúcej.

- w_1 až w_n sú synaptické váhy, ktoré prepájajú jednotlivé neuróny a sú nositeľom znalostí

- b je bias, čo je konštanta, ktorá plní podobnú funkciu ako priesečník s osou y v rovnici priamky - umožňuje posunúť aktivačnú funkciu a zlepšiť učenie

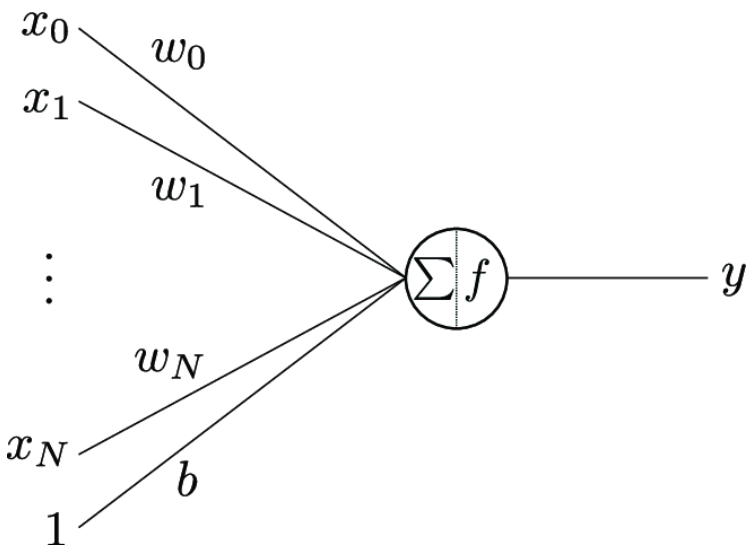
- $\sum$ je suma súčinov vektora vstupov a vektora váh plus bias. Je to hodnota, ktorá sa použije ako vstup do aktivačnej funkcie.

- f je aktivačná funkcia, ktorej úlohou je transformovať vstupný signál na výstupný. Od nej závisí, či sa neurón „aktivuje“.

- y je výstupný signál neurónu, ktorý sa použije ako vstupný signál pre nasledujúcu vrstvu

Existujú viaceré typy aktivačných funkcií, pričom medzi najpoužívanejšie patria sigmoidálna funcia, ReLu funkcia alebo hyperbolický tangens.

- sigmoidálna funkcia -
$$f(x) = \frac{1}{1 + e^{-x}},$$
kde x je vstup a e je Eulerovo číslo. Táto funkcia má plynulý sklon, čo zamedzuje „skokom” na výstupe, výstup je v rozmedzí 0 - 1, vďaka čomu je dobrá pri predikcií pravdepodobnosti. Nevýhodou je, že je málo citlivá na príliš vysoké alebo príliš nízke vstupy.



Obrázok 3 – 2 Model neurónu - základnej stavebnej jednotky neurónových sietí. Zdroj: (Ioannou, 2017)

- ReLu funkcia -

$$f(x) = max(0, x),$$

kde x je vstup. Výhodou je rýchla konvergencia, nevýhodou je necitlivosť v blízkom okolí nuly a pre záporné čísla.

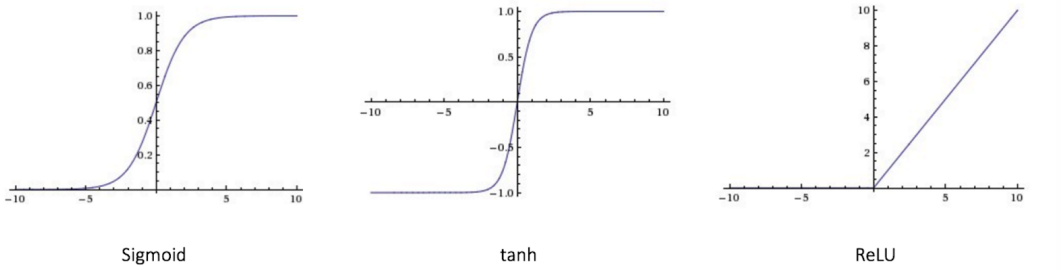
- TanH / hyperbolický tangens -

$$f(x) = tanh(x) = \frac{2}{1 + e^{-2x}} - 1,$$

kde x je vstup a e je Eulerovo číslo. Táto funkcia je veľmi podobná sigmoidálnej, avšak výstup je z rozsahu -1 až 1. Má stršmí sklon ako sigmoidálna, vďaka čomu konverguje rýchlejšie, ale taktiež je necitlivá na príliš vysoké alebo príliš nízke vstupy.

Neuróny sa následne usporiadajú do vrstiev, pričom poznáme 3 typy:

- vstupná vrstva – do nej vstupujú dáta z reálneho sveta
- skrytá vrstva – jej neuróny dostávajú dáta zo vstupnej vrstvy, prípadne z predchádzajúcich skrytých vrstiev. Neurónová sieť ich môže mať viacero.



Obrázok 3 – 3 Grafické zobrazenie priebehu aktivačných funkcií sigmoid, ReLu a TanH.

- výstupná vrstva – jej výsledok sa vracia do reálneho sveta

Činnosť neurónovej siete vieme vo všeobecnosti rozdeliť na fázu učenia, v ktorej dochádza k zmene synaptických váh (získavaniu znalostí) a fázu života, v ktorej sa získané znalosti aplikujú na riešenie nejakého problému (napr. klasifikácia, optimalizácia, predikcia) (Sinčák and Andrejková, 1996).

To, ako sa synaptické váhy menia, závisí od typu učenia neurónovej siete. V prípade kontrolovaného učenia je najčastejšou metóda spätného šírenia chyby (backpropagation). Prebieha tak, že sa jeden prvok trénovacej množiny privedie na vstupnú vrstvu a dopredným šírením prejde cez celú sieť až získame odpovedajúci výstup pre momentálne synaptické váhy. Rozdiel tejto a očakávanej hodnoty je chyba siete. Spätným šírením tejto chyby upravíme váhy, pričom príslušná zmena sa vypočíta nasledovne (Sinčák and Andrejková, 1996):

$$\Delta w_{ij}(t) = -\gamma \frac{\partial J(t)}{\partial w_{ij}(t)},$$

kde w_{ij} je synaptická váha medzi i-tým a j-tým neurónom, γ je učiaci koeficient a $J(t)$ je chybová funkcia. Následne tento postup zopakujeme pre všetky prvky trénovacej množiny, a tým vykonáme jednu epochu učenia neurónovej siete. Ak je celková chyba menšia ako požadovaná presnosť, tak učenie ukončíme. Ak nie, začneme ďalšiu epochu.

Tento spôsob úpravy váh sme použili aj v podkapitole Lineárna regresia pre úpravu koeficientov a a b, keďže ide o metódu postupného znižovania gradientu.

3.6 Random Forest

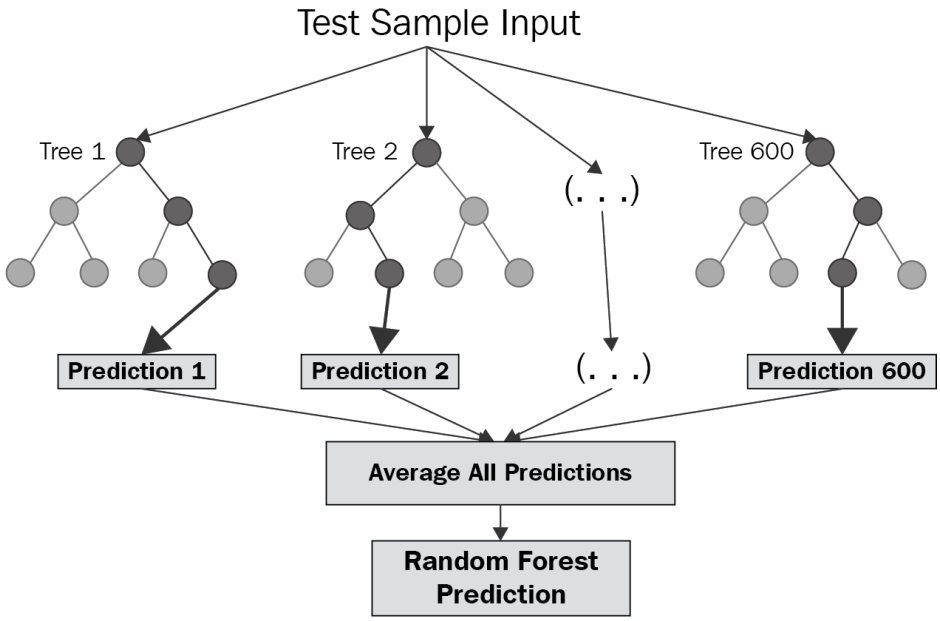
Random Forest (Náhodný les), ako už názov napovedá, je metóda strojového učenia, ktorá pracuje s náhodnými rozhodovacími stromami. Je vhodná ako pre klasifikáciu, tak aj predikciu (Breiman, 2001).

Základnou jednotkou tohto algoritmu je rozhodovací strom, ktorý sa skladá z koreňového uzlu, medziľahlých uzlov a listových uzlov. Cieľom je rozdeliť dáta do tried, na základe jednotlivých atribútov. Na začiatku sa nachádzame v koreňovom uzle, ktorý obsahuje celý dataset. Keďže chceme, aby bol výsledný strom čo najjednoduchší, tak za deliaci atribút zvolíme ten, ktorý má najväčší informačný zisk (najviac od seba odlíši jednotlivé objekty). Týmto sa vytvoria medziľahlé uzly, v ktorých proces opakujeme. Rozhodovanie sa ukončí, ak už nemám ďalšie atribúty, podľa ktorých by sme mohli objekty deliť, alebo uzol obsahuje už iba objekty jednej triedy. V takom prípade tento uzol nazývame listový.

Problémom rozhodovacích stromov je, že sú značne citlivé na trénovacie dáta, čo môže často viesť k preučeniu. Random Forest tento problém rieši tak, že vytvorí mnoho paralelných rozhodovacách stromov, ktoré sa v procese učenia nijako neovplyvňujú (viď. obrázok 3 – 4) (Chakure, 2019). Následne sa pre daný údaj zistí výsledná hodnota jednotlivých stromov a finálna hodnota (resp. trieda) je daná ich priemerom (resp. modusom).

Aby sa zabezpečilo, že jednotlivé stromy nebudú rovnaké, tak do ich vzniku zasahujú dve náhodné hodnoty:

- každý strom má pri učení k dispozícií iba určitú časť náhodne vybratých atribútov, čo zabezpečí, že celkový model nebude príliš závisieť iba od jedného konkrétneho atribútu so silnou informačnou hodnotou
- každý strom má pri učení k dispozícií iba určitú časť náhodne vybratých prvkov trénovacej množiny, čo bráni celkovému preučeniu modelu



Obrázok 3 – 4 Diagram zobrazujúci schému algoritmu Random Forest. Zdroj: (Chakure, 2019)

3.7 XGBoost

XGBoost je najmladším algoritmom spomedzi všetkých vyššie predstavených a z určitého pohľadu spája ich pozitívne vlastnosti. Vyvinutý bol na Washingtonskej Univerzite ako výskumný projekt. Od svojho uvedenia tento algoritmus vyhral množstvo súťaží týkajúcich sa strojového učenia a dolovania dát, pričom za zmienku stoja napríklad výzvy stránky Kaggle z roku 2015, kde z 29-tich víťazných riešení 17 využívalo XGBoost (Chen and Guestrin, 2016).

XGBoost (Extrémne zosilnenie gradientu) je založený na metóde Gradient Boosting. Ten vytvára predikčný model vo forme viacerých slabších predikčných modelov, ktorými sú zväčša rozhodovacie stromy. V tomto aspekte je podobný s vyššie popísaným algoritmom Random Forest, avšak existujú medzi nimi rozdiely hlavne v tom, ako sú jednotlivé rozhodovacie stromy vytvárané a kombinované.

Kým Random Forest vytvára kompletné stromy z náhodne vybratých vzoriek paralelne, Gradient Boosting vytvára stromy sekvenčne. Prvý strom má slabú pre-

dikčnú silu, ale každý nasledujúci sa učí predikovať to, čo predchádzajúci nedokázal, až kým sa nevytvorí dostatočne silný model. Inak povedané, každý nasledujúci strom sa pokúsi predikovať chybu, teda rozdiel medzi predikovanou a skutočnou hodnotou z predchádzajúceho stromu. Je potrebné spomenúť, že aj keď Gradient Boosting opisujeme v kontexte s rozhodovacími stromami, tak tento algoritmus je možné aplikovať aj na hocijaké iné modely, avšak najlepšie výsledky dosahuje práve so stromami (Mandot, 2019).

XGBoost sa na rozdiel od Gradient Boosting-u nesnaží predikovať chybu, ale gradient chybovej funkcie, teda najvhodnejší smer znižovania chybovej funkcie. Po tom, čo sú všetky objekty rozhodovacieho stromu rozdelené do listových uzlov, je možné vypočítať priemerný gradient, ktorý je následne vynásobený učiacim parametrom. Následne sa pre každý listový uzol urobí krok v smere vypočítaného gradientu (Mandot, 2019). Takto sa zníži chybová funkcia pre všetky prvky listového uzlu a vytvorí sa nový strom. V konečnom dôsledku XGBoost oproti Gradient Boosting-u obsahuje nasledujúce vylepšenia (Morde and Setty, 2019):

- kým Gradient Boosting vytvára stromy po jednom, XGBoost je schopný paralelného trénovania, čím sa značne zvýši rýchlosť učenia
- penalizuje príliš komplexné stromy - obsahuje reguláciu maximálnej hĺbky stromov, čo predchádza preučeniu, a taktiež zvyšuje rýchlosť výpočtov
- obsahuje vstavanú krížovú validáciu po každej iterácií, vďaka ktorej nie je potrebné zadávať počet epoch trénovania
- dokáže pracovať aj s chýbajúcimi hodnotami

4 Postup riešenia

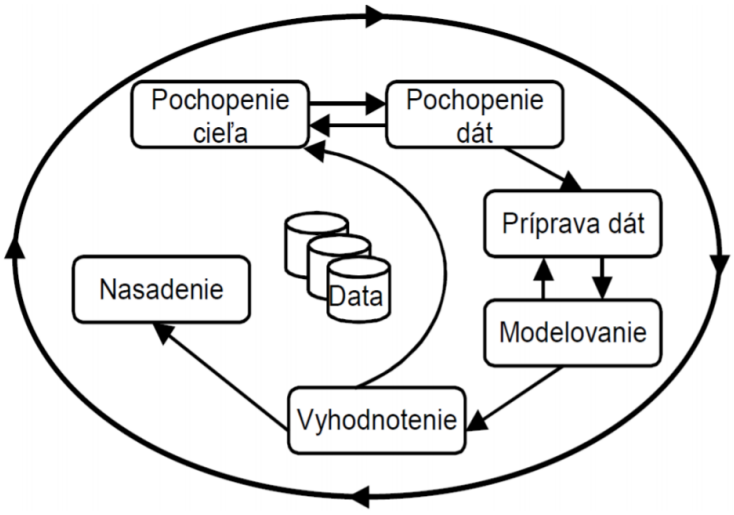
V rámci našej práce sme sa počas výskumu riadili metodikou CRISP-DM (skratka z Cross-Industry Standard Process for Data Mining). Táto metodika bola sformulovaná v polovici 90-tych rokov s cieľom štandardizovať a urýchliť proces dolovania dát. V pomerne krátkom čase sa dostala do povedomia dátových analytikov a začala sa využívať ako vo výskumnej, tak aj v obchodnej oblasti. Od jej vzniku sa dátová veda posunula míľovými krokmi dopredu a začali sa riešiť komplexnejšie problémy, ktoré si vyžadujú pokročilejšie algoritmy. Ale vo všetkých týchto projektoch aj naďalej zostalo nevyhnutné pochopenie problematiky, získavanie dát a ich následné spracovanie, tvorba riešenia problému a nasadenie výsledkov do praxe. Aj vďaka tomu, že metodika CRISP-DM pokrýva všetky tieto kľúčové kroky, je aj v dnešnej dobe spoľahlivým a užitočným nástrojom pri riešení najnovších výziev v oblasti dátovej analytiky. Metodika CRISP-DM je tvorená nasledovnými krokmi (viď. obrázok 4 – 1) (Chapman et al., 2000):

- pochopenie cieľa – v tomto kroku sa stanoví, čo je cieľom projektu, aké kroky je potrebné vykonať na jeho dosiahnutie, vytvorí sa zoznam potrebných prostriedkov, zváži sa možné obmedzenia a vytvorí sa projektový plán
- pochopenie dát – zahŕňa činnosti ako hľadanie zdroja dát a ich získanie, vytvorenie si základného prehľadu o ich obsahu, vykreslenie vizualizácií, identifikovanie možných problémov s dátami a vypočítanie základných štatistických charakteristík, akými sú napríklad priemer, rozptyl alebo početnosti
- príprava dát – táto fáza obsahuje činnosti ako vybratie relevantných podmnožín pre náš projekt, transformáciu dát do potrebného formátu, spájanie dát z rôznych zdrojov, generovanie odvodených atribútov a v neposlednom rade čistenie dát, ktoré zahŕňa napríklad odstránenie duplicitných hodnôt a spracovanie chýbajúcich údajov
- modelovanie – v tejto časti je potrebné na začiatku zvoliť vhodnú metriku,

ktorá bude hodnotiť úspešnosť modelov. Následne sa vyberú a aplikujú cieľuprimerané modelovacie techniky, pri ktorých by sa mal venovať dostatok času aj hľadaniu vhodných hyperparametrov. Keďže množstvo algoritmov má rozličné nároky na formát dát, tak je často nevyhnutné vrátiť sa o krok späť k príprave dát a uskutočniť zmeny.

- vyhodnotenie výsledkov – v tomto kroku sa zosumarizujú výsledky zo všetkých predchádzajúcich krokov a zhodnotí sa, či bol dosiahnutý cieľ, ktorý bol stanovený na začiatku. V tejto fáze je predstavený najlepší model spolu s nastaveniami jeho parametrov, aby bol výsledok reprodukovateľný. Vhodné je aj navrhnúť ďalšie kroky, ktoré by sa mohli vykonať v nasledujúcich projektoch pre dosiahnutie ešte lepších výsledkov.

- nasadenie – posledným krokom metodiky CRISP-DM je uvedenie získaných poznatkov a modelov do praxe. Pre tento účel je potrebné vytvoriť plán nasadenia, ktorý bude obsahovať všetky nevyhnutné postupy a návod, ako ho vykonať.



Obrázok 4 – 1 Diagram zobrazujúci jednotlivé kroky metodiky CRISP-DM

4.1 Pochopenie cieľa

Žijeme v dobe, kedy umelá inteligencia preniká okrem odbornej sféry aj do našich súkromných životov. Je len málo oblastí, v ktorých si zatiaľ algoritmy strojového učenia nenašli uplatnenie. V obchodnej oblasti dokonca prestávajú byť konkurenčnou výhodou a pomaly sa stávajú nevyhnutnosťou pre udržanie sa na trhu. Čo sa týka výskumu, je ich dopad miernejší, nakoľko v dohľadnej dobe určite nepripravia vedcov o prácu, ale sú pre nich veľmi užitočným pomocníkom. Miera využitia sa však líši v závislosti od vedeckej disciplíny. Prirodzene, bolo vytvorených viac aplikácií v odboroch s veľkým množstvom dát, ako sú napríklad medicína alebo astronómia. Avšak aj tu je stále priestor na zlepšovanie a hľadanie nových možností uplatnenia strojového učenia, čomu sme sa budeme venovať aj v našej práci.

Cieľom našej práce je zistiť, či je s pomocou dnešných algoritmov možné modelovať intenzitu airglow-u a ak áno, akú presnosť sme schopní dosiahnuť. Keďže ide o jav, na ktorý vplýva veľké množstvo rôznych faktorov, tak by sme taktiež chceli určiť, ktoré majú najväčší podiel na predikcii výslednej intenzity v rámci použitých algoritmov a overiť opodstatnenie týchto výsledkov vzhľadom na platné fyzikálne zákony. Tieto informácie potenciálne môžu poukázať aj na nepriame faktory, ktorých vplyv nemusí byť na prvý pohľad zrejmý a budú predmetom ďalšej výskumnej činnosti.

Na dosiahnutie týchto cieľov je potrebné získať vhodné dáta s dostatočnou početnosťou. Následne sa na základe charakteru dát vyberú vhodné predikčné algoritmy, pre ktoré sa upraví formát údajov do potrebnej podoby. Pred modelovaním je potrebné zvoliť hodnotiacu metriku, ktorá jasne posúdi úspešnosť algoritmov medzi sebou pri predikcií jednej emisnej čiary, tak ako úspešnosť jedného algoritmu pri modelovaní airglow-u s rôznymi vlnovými dĺžkami. Následne je nutné venovať dostatok času nastaveniu hyperparametrov a výberu relevantných podmnožín atribútov. Nakoniec sa vykonajú samotné predikcie a vyhodnotia sa ich výsledky.

V rámci potrebných zdrojov pre túto prácu budú postačujúce verejne dostupné (open-source) softvérové prostriedky špecifikované v kapitole 3.

4.2 Pochopenie dát

Dáta, ktoré chceme modelovať, pochádzajú z observatória Abastumani v Gruzínsku, ktoré leží na 41°45 ′ severnej šírky a 42°49 ′ východnej dĺžky v nadmorskej výške 1650 metrov. Merania boli vykonávané medzi rokmi 1957 až 1993, pričom nie sú rovnomerne rozložené v čase, keďže pozorovania boli možné iba za priaznivého počasia. Dáta sme získali vo forme excelovských tabuliek, kde jeden súbor predstavoval merania za jednu noc. V súčte ide o 1433 súborov, čo je v porovnaní so všetkými 13 514 nocami počas 37 rokov $10,6\%$ všetkých nocí.

Tento dataset sa dá podľa obsahu a formy tabuliek rozdeliť na tri časti. Prvá časť obsahuje dáta z rokov 1957 až 1961, do druhej časti sa dajú zaradiť merania medzi rokmi 1962 a 1974. Do poslednej tretej časti patria zvyšné dáta, teda roky 1975 až 1993.

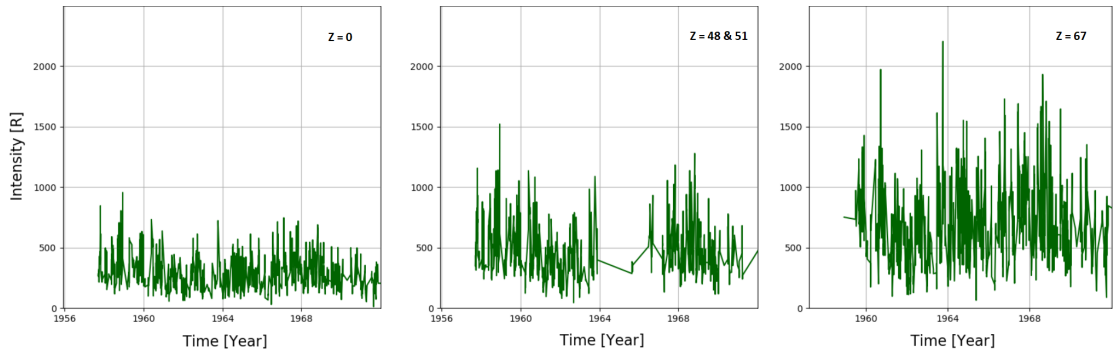
Prvá časť obsahuje údaje o airglow-e s vlnovými dĺžkami 557,7 nm, 589,3 nm a 630 nm. Ku každému riadku prislúcha časový údaj, kedy boli dané intenzity namerané v miestnom časovom pásme, vo formáte HH:MM. Taktiež obsahuje údaj Z, čo je zenitový uhol, ktorý určoval uhol osi fotometra od zenitu. Druhá časť datasetu je rovnaká ako prvá, avšak obsahuje naviac aj intenzitu airglow-u v OH vrstve a k nej samostatný časový údaj.

Tretia časť datasetu obsahuje intenzity airglow-u pre vlnové dĺžky 557,7 nm, 630 nm a intenzitu pre OH vrstvu. Intenzita zodpovedajúca 589,3 nm už viac nebola zaznamenávaná, rovnako ako zenitový uhol, ktorý chýba v tejto časti datasetu a nie je ani jednoznačne určená jeho hodnota pre tieto dáta. To môže do modelovania vniesť určitú chybu. Táto časť taktiež neobsahuje jeden spoločný časový údaj, ale pre každú vrstvu bol čas zapisovaný zvlášť.

Najprv sa bližšie pozrieme na intenzity airglow-u samostatne pre jednotlivé vlnové dĺžky, aby sme vedeli zhodnotiť kvalitu dát a určiť kroky, ktoré bude potrebné vykonať v rámci prípravy dát. Ako prvú zanalyzujeme vlnovú dĺžku 557,7 nm.

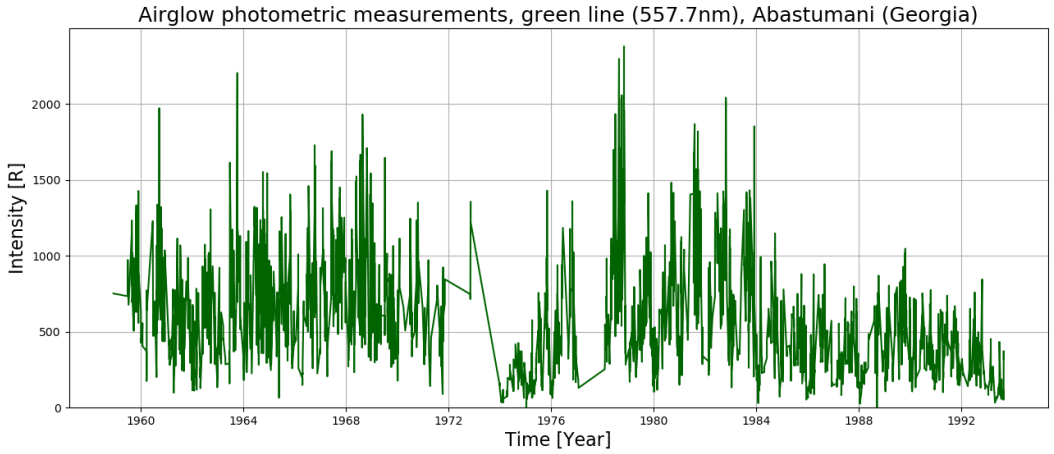
Jednotlivých meraní máme spolu 36 045, čo je dostatočne veľká vzorka. Treba však spomenúť, že rozloženie v čase nie je rovnomerné, keďže pozorovanie bolo možné iba za priaznivého počasia a v observatóriu v Abastumani sa venovali aj iným meraniam a výskumu, čo sa odzrkadlilo aj na rôznej početnosti dát pre jednotlivé roky. Medzi rokmi 1957 až 1974 bolo vykonaných 9623 meraní a od roku 1975 bolo vykonaných zvyšných 26 422 meraní.

Prvým problémom je zenitový uhol v prvej a druhej časti datasetu, respektíve jeho absencia v tretej. Intenzity boli merané pre 4 rôzne zenitové uhly: 0, 48, 51 a 67. Početnosti dát pre jednotlivé uhly sú v rovnakom poradí nasledovné: 3012, 1105, 956 a 3106 údajov. Po ich vykreslení je zrejmé, že nemôžu byť použité spoločne, kvôli značným rozdielom v rozsahu ich intenzít (viď. obrázok 4 – 2). Uhly 48 a 51 sme spojili do jedného grafu, keďže sú dostatočne blízko pri sebe. Na základe analýzy rozsahu pre jednotlivé roky a pre jednotlivé zenitové uhly sme dospeli k záveru, že tretia časť datasetu je najpodobnejšia so zenitovým uhlom 67, a tak tieto údaje spojíme.



Obrázok 4 – 2 Grafy zobrazujúce intenzity airglow-u pre vlnovú dĺžku 557,7 nm pre jednotlivé zenitové uhly.

Vykreslenie všetkých dát pre vlnovú dĺžku 557, 7 nm je na obrázku 4 – 3. Vidieť, že kompletne chýbajú dáta pre roky 1972 až 1974 a pre rok 1977, ale to by nemal byť problém. Štatistické informácií sú zosumarizované v tabuľke 4 – 1.



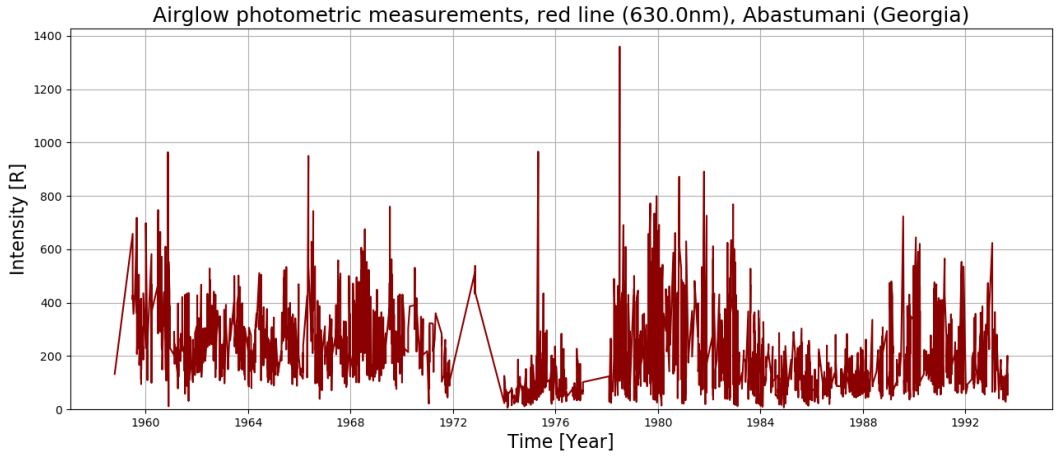
Obrázok 4 – 3 Intenzity airglow-u namerané pre vlnovú dĺžku 557,7 nm medzi rokmi 1957 a 1993.

Dáta pre červenú kyslíkovú čiaru, teda pre airglow s vlnovou dĺžkou 630 nm, sú podobne početné ako dáta vyššie popísanej zelenej kyslíkovej čiary. Celkovo ich je 35 598, pričom 9224 z nich bolo nameraných pred rokom 1974 a 26 374 bolo zaznamenaných od roku 1975 po rok 1993.

Okrem podobného počtu majú aj rovnaké nedokonalosti. Pre jednotlivé zenitové uhly rovné 0, 48, 51 a 67 obsahujú v danom poradí 2995, 1093, 852 a 3088 hodnôt. Opäť sme s dátami od roku 1975 spojili dáta so Z = 67 a výsledný priebeh intenzity pre červenú emisnú vrstvu je na obrázku 4 – 4. Opäť neobsahuje žiadne hodnoty pre roky 1972-1974 a rok 1977.

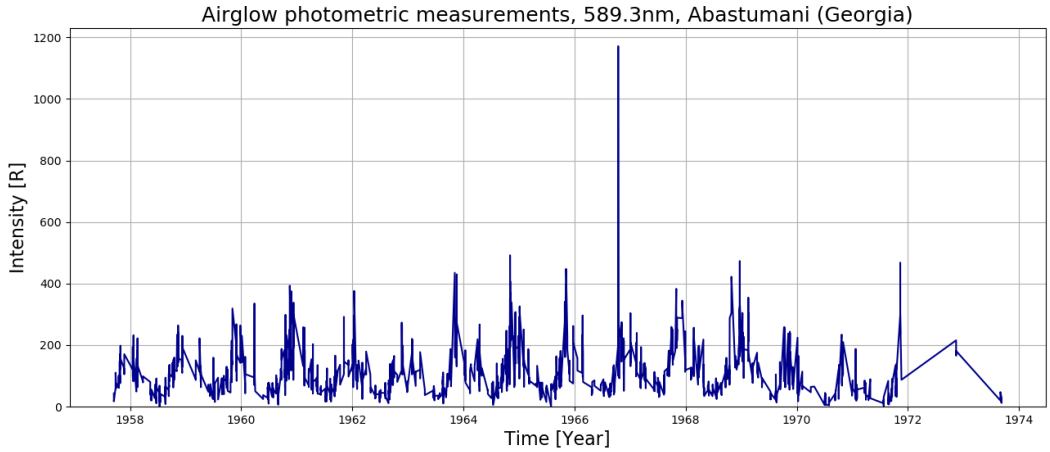
Charakteristické štatistické hodnoty o tejto vrstve (viď. tabuľka 4 – 1) napovedajú, že je značne slabšia ako zelená vrstva.

Pre žiarenie hornej vrstvy atmosféry s vlnovou dĺžkou 589,3 nm máme najmenej údajov zo všetkých. Táto emisia sa nachádza v našom datasete iba medzi rokmi



Obrázok 4 – 4 Intenzity airglow-u namerané pre vlnovú dĺžku 630,0 nm medzi rokmi 1957 a 1993.

1957 až 1972. Konkrétne ide o 652 nocí, čo za 18 rokov predstavuje 9, 9%. Celkový počet nameraných hodnôt je 9091, pričom sa opäť tieto údaje zbierali pre 5 rôznych zenitových uhlov. Najpočetnejší je uhol 67°, pre ktorý dataset obsahuje 3036 meraní. Vizualizácia tejto vrstvy so spojenými všetkými zenitovými uhlami je na obrázku 4 – 5.



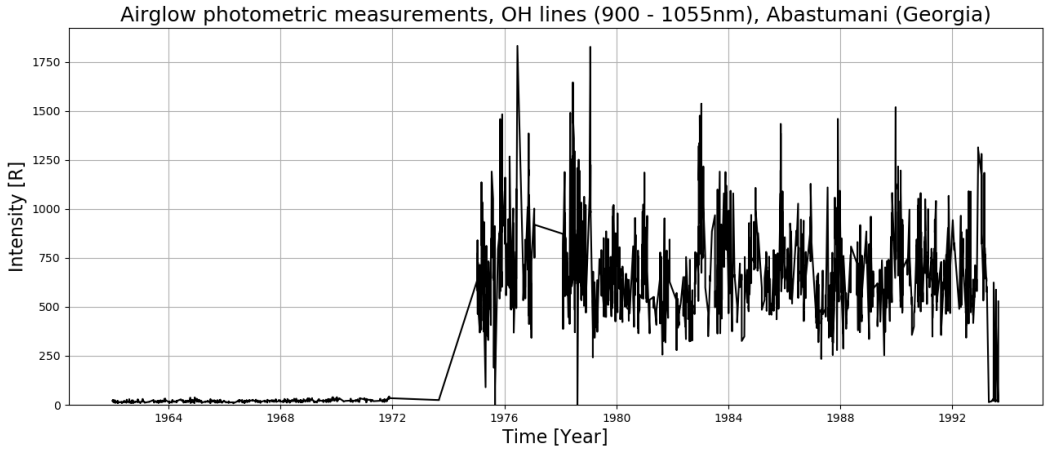
Obrázok 4 – 5 Intenzity airglow-u namerané pre vlnovú dĺžku 589,3 nm medzi rokmi 1957 a 1974.

Poslednou airglow emisiou v našom datasete je OH vrstva. Medzi rokmi 1975 a 1993 bolo vykonaných 25 295 meraní, avšak pred rokom 1975 iba 3660 meraní. Táto

	I 557,7	**I 630,0**	**I 589,3**	**I OH**
Počet dát	36 045	35 598	9 091	25 295
Maximálna hodnota	2 379,5 R	1 360 R	1 534 R	1 832,9 R
Minimálna hodnota	0 R	6 R	0 R	0 R
Priemer	572,9 R	183,5 R	154 R	433,7 R
Smerodajná odchýlka	344,8 R	125 R	106,5 R	372,6 R

Tabuľka 4 – 1 Tabuľka zobrazujúca štatistické charakteristiky dostupných dát pre jednotlivé vrstvy airglow-u.

vrstva, na rozdiel od predchádzajúcich troch, pokrýva širší rozsah vlnových dĺžok, konkrétne infračervenú spektrálnu oblasť v rozsahu 900 - 1055 nm, čo ju robí ťažšie predikovateľnou, keďže závisí ešte od väčšieho počtu procesov prebiehajúcich v horných vrstvách atmosféry. Naviac, hodnoty v našom datasete sú nekonzistentné, čo je jasne viditeľné na obrázku 4 – 6. Pred rokom 1975 bol pravdepodobne na merania použitý iný prístroj, prípadne boli hodnoty inak prepočítavané. Keďže nepoznáme spôsob, ako tieto údaje zjednotiť, tak sú pre nás spolu nepoužiteľné. Jedinou možnosťou je modelovať dáta iba po roku 1975.



Obrázok 4 – 6 Intenzity airglow-u namerané pre OH vrstvu medzi rokmi 1957 a 1993.

Ak chceme predikovať hodnoty airglow-u s dostatočnou presnosťou, je potrebné

získať relevantné atribúty. Tieto dáta sme získali z verejne dostupnej databázy NASA zvanej OMNIWeb. Dostupná je na adrese: (NASA, (accessed March 14, 2020). Obsahuje 49 parametrov kozmického počasia rozdelených do 5 kategórií: magnetiké pole, plazma, odvodené parametre, indikátory a častice. Keďže najstaršie dostupné dáta v tejto databáze sú z novembra 1963, tak aj airglow hodnoty budeme môcť použiť na predikciu len od tohto dátumu.

Prvou kategóriou sú dáta týkajúce sa medziplanetárneho magnetického poľa (skr. IMF z Interplanetary Magnetic Field). Je to časť magnetického poľa Slnka, ktoré je slnečným vetrom unášané do medziplanetárneho priestoru. Kvôli rotácii Slnka sa toto pole, tak ako slnečný vietor, šíri v tvare špirály. Vzniká v regiónoch na Slnku, kde je magnetické pole „otvorené", teda materiál, ktorý sa pozdĺž neho šíri, sa nevracia na povrch Slnka (ako v prípade „uzavretých" siločiar), ale uniká do medziplanetárneho priestoru. Keď sa takéto pole stretne so zemským magnetickým poľom, pričom ich siločiary sú orientované navzájom opačne alebo antiparalelne, môžu sa spojiť a dôjde k výmene energie, hybnosti a hmoty. Pre naše potreby sme zo 14-tich atribútov vybrali 2: Vector B Magnitude, ktorý vyjadruje veľkosť vektora IMF a RMS field vector, ktorý vyjadruje maximálnu chybu merania. Veľkosť IMF sa meria v nT (nano Tesla) a v blízkosti Zeme dosahuje od 1 nT do 37 nT s priemerom okolo 6 nT.

Ďalšou kategóriou je plazma, ktorá obsahuje 12 rôznych atribútov. Ide o parametre slnečného vetra (solar wind, SW), ktorý je tvorený časticami (zväčša protónmi a elektrónmi), ktoré unikli z gravitačného poľa Slnka vďaka svojím vysokým energiám. Pre nás sú dôležité SW Plasma Temperature [K], čo je teplota uniknutej plazmy, SW Proton Density [počet častíc v cm3], ktorá vyjadruje hustotu slnečného vetra a SW Plasma Speed [km/s], čo je rýchlosť častíc. Každému z týchto atribútov prislúcha jeden ďalší atribút vyjadrujúci jeho odchýlku, ktoré sú pomenované sigma-T, sigma-n a sigma-V.

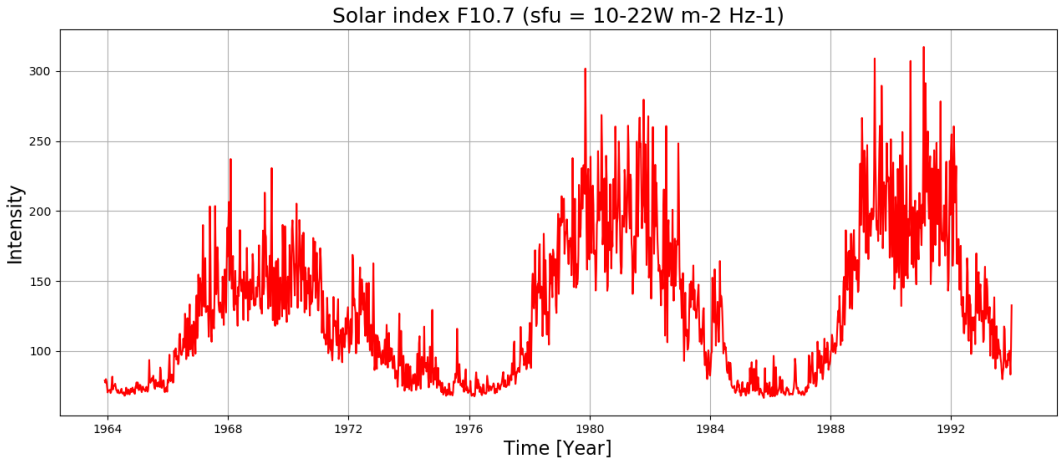
Treťou kategóriou sú odvodené atribúty, ktoré vznikli výpočtom zo základných atribútov. Napríklad tlak slnečného vetra (Flow pressure) sa vypočíta z jeho hustoty a rýchlosti. Z tejto kategórie nepoužijeme žiadny atribút, keďže by boli redundantné.

Nasledujúcou kategóriu sú rôzne typy indexov. Kp index pochádza z nemeckých slov „planetarische kennziffer" a je voľne preložiteľný ako planetárny index. Na svete je 13 magnetometrických staníc v stredných geografických šírkach, ktoré merajú geomagnetické fluktuácie vo svojej oblasti a zaznamenávajú ich ako K-hodnoty, ktoré sú v rozsahu od 0 po 9. Kp index je priemerom týchto 13-tich K-hodnôt.

Atribút Dst index je skratkou z anglického „disturbance storm time" a je to meranie geomagnetickej aktivity, s cieľom odhadnúť závažnosť magnetických búrok. Vyjadrený je v nT a počíta sa ako priemer horizontálnej zložky zemského magnetického poľa zo štyroch blízko-rovníkových observatórií.

Indexy AE, AL a AU sú odvodené zo záznamov magnetometrov v oblasti polárneho kruhu a opisujú úroveň narušenia magnetického poľa. Hodnoty horizontálnych magnetických zložiek sú vykreslené relatívne k ich pokojovým hodnotám tak, aby sa prekrývali. Horná obalová krivka je AU index (amplitude upper) a dolná je AL index (amplitude lower). AE index (auroral electrojet) je definovaný ako rozdiel indexov AU a AL. Pre naše potreby použijeme iba AE index, keďže zvyšné dva by boli redundantné.

Atribút F10.7 index vyjadruje množstvo žiarenia emitované Slnkom s vlnovou dĺžkou 10,7 cm. Je to vynikajúci indikátor slnečnej aktivity, keďže na rozdiel od ostatných slnečných parametrov je ho možné merať aj za nepriaznivého počasia zo zemského povrchu. Meraný je v jednotkách s.f.u. (z angl. solar flux unit) a jeho hodnota sa pohybuje v rozmedzí od 50 do 300 s.f.u. v závislosti od slnečnej aktivity. Tento atribút dobre koreluje s mnohými emisiami ultrafialového žiarenia, ktoré ovplyvňujú vrchnú atmosféru a nami študované javy. Z obrázku 4 – 7 vidno, že jeho priebeh závisí od 11-ročného cyklus slnečnej aktivity.



Obrázok 4 – 7 Intenzita index-u F 10.7 nameraná medzi rokmi 1964 až 1993.

Atribút Lyman alpha opisuje intenzitu spektrálnej čiary vodíka, ktorá vzniká pri prechode elektrónu z druhého orbitálu na prvý a je sprevádzaná emisiou fotónu s vlnovou dĺžkou 121,5 nm. Atribút R (sunspot No.) vyjadruje počet škvŕn na privrátenej strane Slnka. Keďže obidva tieto atribúty značne korelujú s index-om F10.7, tak pre nás nie sú potrebné. Poslednou kategóriou databázy OMIWeb sú častice. Táto obsahuje počet častíc pre rôzne úrovne energií väčšie ako: 1 MeV, 2 MeV, 4 MeV, 10 MeV, 30 MeV a 60 MeV. Keďže každý atribút s menšou energiou obsahuje aj všetky hodnoty predchádzajúcich vyšších energií, tak nám postačí atribút Proton Flux (>1Mev) a ostatné sú redundantné.

Našim posledným zdrojom dát je databáza NRLMSISE, spravovaná NASA a dostupná na stránke: (NASA/CCMC, (accessed May 24, 2020). Obsahuje údaje o zložení atmosféry v konkrétnom čase, v konkrétnej výške a nad určitým miestom. Pre nás sú dôležité hlavne atribúty o počte častíc atomárneho vodíka H, kyslíka O, dusíka N, molekulového kyslíka O2, dusíka N2, celková hustota atmosféry a teplota.

4.3 Príprava dát

Prvou úlohou je načítať údaje z tabuliek a uložiť ich do formátu, s ktorým by sa ľahšie manipulovalo. Na načítanie prvej a druhej časti sme vytvorili skript xls2pd-57-74.py, ktorý prečíta údaje a uloží ich do štyroch stĺpcov podľa typu meraného airglow-u. Tretiu časť dokáže načítať a spracovať skript xls2pandas.py.

Problémom je variabilita tabuliek, vzhľadom na to, že názvy stĺpcov sa menia, teda existuje viac variantov pre jednu čiaru airglow-u. Údaje nezačínajú vždy na rovnakom riadku, ani nenasledujú striktne za sebou (obsahujú prázdne riadky). Taktiež je potrebné pridať dátum ku každému meraniu, keďže čas je zaznamenaný iba v hodinách a minútach a prislúchajúci dátum je len v názve súboru. Pri tomto úkone si treba dať pozor na posledné dni v mesiacoch, aby sa nestalo, že sa dáta pridajú do minulého mesiaca. Následne je potrebné konvertovať čas z lokálneho časového pásma (GMT+4) na UTC, keďže všetky ostatné dáta používajú práve tento formát.

Po úspešnom načítaní všetkých airglow dát je potrebné odstrániť tie údaje, pri ktorých meraní ešte nebola astronomická noc (Slnko bolo vyššie ako $-18°$ pod obzorom) a taktiež tie, keď bol Mesiac nad obzorom, keďže to mohlo vniesť určitú nepresnosť do meraní. Na tento účel sme využili pyhton knižnicu ephem, ktorá poskytuje veľmi presné astronomické výpočty. Kvôli tomuto kroku nám počet dát pre zelenú kyslíkovú vrstvu klesol z 35 799 na 22 938 a pre červenú kyslíkovú vrstvu z 35 598 na 23 240, bol však potrebný, keďže prístroje merajúce airglow sú citlivé aj na takéto nepriame svetlo. Tieto hodnoty sme následne nechali ako atribúty v datasete a sú označené skratkami: SA - Sun altitude, MA - Moon altitude a AU - astronomical unit. Prvé dva menované sú vyjadrené v stupňoch, tretí atribút je hodnota blízka 1, ktorá vyjadruje momentálnu vzdialenosť Zem - Slnko.

Ďalším krokom je prepočítanie dát na celé hodiny, pretože najmenšie možné intervaly dát z OMNIWeb-u sú práve hodinové. Využili sme funkciu „resample“ v knižnici Pandas, ktorá údaje v celom datasete transformovala. V prípade, že pre

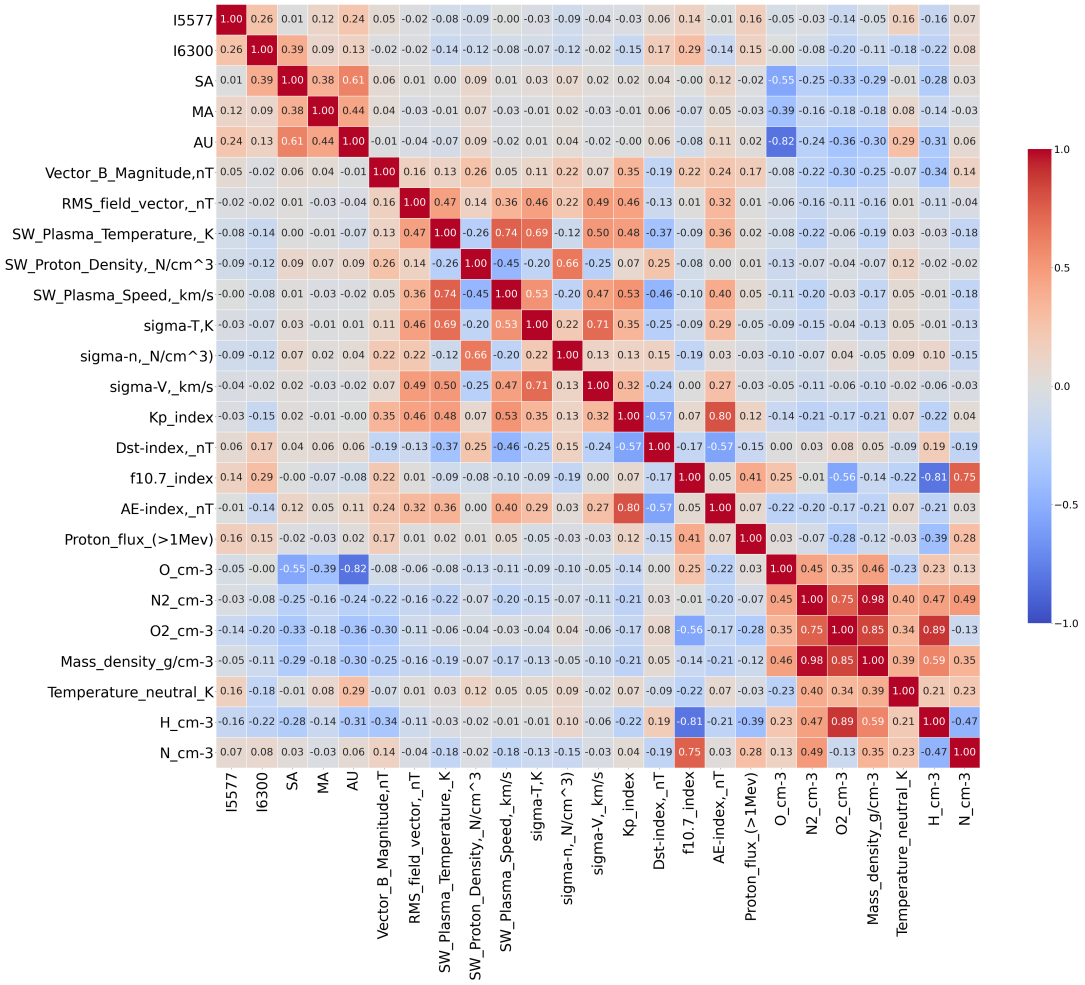
jednu hodinu bolo vykonaných viac meraní airglow-u, tak ako konečnú hodnotu sme použili ich priemer. Týmto prepočtom nám klesol počet dát pre vlnovú dĺžku 557,7 nm na 5174 a pre vlnovú dĺžku 630 nm na 5224.

Dáta z databázy OMNIWeb-u taktiež neboli vo vhodnom formáte. Stiahnuteľné sú ako jeden súbor, v ktorom sú stĺpce oddelené jednou alebo viacerými medzerami. Stiahli sme všetky dostupné stĺpce, a tak vznikol dataset s 50 stĺpcami a s 263 808 riadkami. Tieto dáta obsahovali chýbajúce hodnoty, ktoré boli reprezentované najvyšším možným číslom zapísateľným do daného stĺpca (pre stĺpec s štvorcifernými hodnotami bola chýbajúca hodnota reprezentovaná ako 9 999). Všetky tieto hodnoty sme konvertovali na hodnotu NaN, ktorú využíva pandas pre tieto prípady.

V ďalšom kroku sme na získanie potrebných údajov z databázy NRLMSISE zadali na stránke geografické súradnice observatória v Abastumani a výšku 97 km pre zelenú airglow čiaru a 250 km pre červenú airglow čiaru. Vzhľadom na to, že stránka nepodporuje jednoduché získanie dát v hodinových intervaloch pre viac rokov, stiahli sme si údaje namerané o polnoci každého dňa.

V poslednej fáze prípravy dát sme spojili všetky získané hodnoty a vytvorili finálny dataset. Vymazali sme všetky riadky, pre ktoré sme nemali žiadnu hodnotu airglow-u a výsledný rozmer nášho celkového datasetu je teda 5 588 riadkov a 65 stĺpcov. Rozhodnutie stiahnuť celkovú databázu OMNIWeb-u vyplynulo z toho, že je praktické mať k dispozícií aj zvyšné atribúty, o ktorých sa nepredpokladá, že majú pre nás veľkú hodnotu, ale v rámci dodatočnej analýzy sa môže ukázať opak. Takto už predpripravené dáta môžu urýchliť prípadný ďalší výskum tejto problematiky.

V rámci našich potrieb sme preto vytvorili aj menší dataset, ktorý budeme využívať na účely predikcie. Tento je podmnožinou vyššie uvedeného a obsahuje iba atribúty vybraté v predchádzajúcej kapitole - pochopenie dát. Následne sme si vykreslili korelačnú maticu finálneho dataset-u, ktorá je na obrázku 4 – 8.



Obrázok 4 – 8 Korelačná matica finálneho dataset-u.

4.4 Modelovanie

V tejto kapitole sa budeme venovať implementácií algoritmov, ktoré boli predstavené v kapitole 3. Na modelovanie airglow-u najprv využijeme jednoduchšie algoritmy ako sú lineárna a polynomiálna regresia, následne použijeme neurónovú sieť, ktorá je však náročnejšia na správne nastavenie hyperparametrov a nakoniec využijeme algoritmy založené na princípe rozhodovacích stromov, teda Random Forest a XGBoost.

Pred samotnou tvorbou modelov je potrebné zvoliť si správnu hodnotiacu metriku, ktorá objektívne vyhodnotí úspešnosť predikcie. Prirodzeným kandidátom je

najjednoduchšia metrika, teda mean absolute error (skr. MAE), čo v preklade znamená priemerná absolútna chyba. Vzťah pre jej výpočet je veľmi podobný so vzťahom na výpočet mean squared error, ktorý sme definovali v rámci teoretickej časti lineárnej regresie, avšak namiesto druhej mocniny je použitá absolútna hodnota. MAE je definovaná nasledovne:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid (y_i - x_i) \mid,$$

kde y_i je predikovaná hodnota, x_i je skutočná hodnota a n je počet vzoriek. Túto metriku budeme používať na porovnanie algoritmov pri predikcií jednej čiary. Neodráža však úspešnosť konkrétneho algoritmu pri predikovaní viacerých airglow vrstiev. Keďže zelená čiara je niekoľkonásobne intenzívnejšia ako červená, tak pre porovnanie úspešnosti použijeme metriku mean absolute percentage error (skr. MAPE), čo v preklade znamená priemerná absolútna percentuálna chyba. Túto metriku knižnica sklearn neobsahuje, preto sme si ju museli naprogramovať. Je definovaná vzťahom:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \mid \frac{x_i - y_i}{x_i} \mid,$$

kde y_i je predikovaná hodnota, x_i je skutočná hodnota a n je počet vzoriek. Táto hodnota sa zvykne násobiť krát 100, aby sme dostali vyjadrenie v percentách.

Dôvod, prečo nepoužívame iba MAPE je, že favorizuje prípady, v ktorých je skutočná hodnota väčšia ako predikovaná a penalizuje opak. Napríklad ak $x_i = 200$ a $y_i = 100$, tak MAPE je 50%. Ak však $x_i = 100$ a y_i je 200, tak MAPE má hodnotu 100%. V obidvoch týchto prípadoch je však MAE rovná 100. Ďalším problémom je, že ak by bola skutočná hodnota rovná nule, tak by výpočet stroskotal na nepovolenom delení nulou. Preto sme sa rozhodli používať pri predikcií aj MAE aj MAPE.

Pred modelovaním sme si pripravili funkciu na rozdelenie dát na trénovaciu a testovaciu množinu a implementovali sme v nej normalizáciu dát. Aby nedošlo k tzv. data leakage (prekl. úniku dát), čo je nežiaduci prípad, keď sa nejakým spôsobom

dostane informácia z testovacích dát do trénovacích a umelo sa vylepší výsledok, tak sme normalizáciu vykonali až po rozdelení dát na testovaciu a trénovaciu množinu. Na začiatok sme pomer rozdelenia nastavili na 80 ku 20.

Prvým algoritmom, ktorý sme použili, je lineárna regresia. Hodnoty aproximuje obyčajnou priamkou a je naším najjednoduchším algoritmom. Preto jej výstup budeme považovať za základnú predikciu, ktorú sa zložitejšími algoritmami budeme snažiť prekonať.

Lineárna regresia nedokáže pracovať s chýbajúcimi hodnotami, takže odstránime všetky nekompletné riadky, kvôli čomu nám pri 23 atribútoch zostane z 5 588 riadkov iba 1504. Pre zelenú kyslíkovú vrstvu je MAE v testovacej množine 224 R a v trénovacej 219 R. MAPE sa v tomto prípade zastavilo na veľkosti chyby $68,6\%$ pri testovaní a $66,2\%$ pri trénovaní.

Druhou možnosťou je chýbajúce hodnoty doplniť namiesto ich vymazania. Vyskúšali sme aj tento prístup a NaN hodnoty sme nahradili priemerom daného stĺpca. Rizikom je, že sa kvôli tomu do datasetu dostane určitá chyba v prípade, že je počet chýbajúcich hodnôt veľmi vysoký. Tento predpoklad sa aj potvrdil. Pre zelený airglow stúpla v testovacej množine absolútna chyba na 234 R a absolútna percentuálna chyba dokonca až na 181%. Tieto čísla dokazujú, že zameranie sa iba na hodnoty jednej metriky môže v konečnom dôsledku spôsobiť veľké skreslenie výsledkov. Skok o 120% pri MAPE je pravdepodobne práve dôsledkom vyššie spomínaného favorizovania nižších predikčných hodnôt.

Ak zopakujeme tento pokus aj pre červenú kyslíkovú čiaru, dostaneme zaujímavý výsledok. Pri vymazaní chýbajúcich hodnôt bola pre testovaciu množinu MAE rovná 72 R a MAPE 69%, naopak pri ich doplnení priemerom pre rovnakú množinu MAE stúpla na 77 R a MAPE klesla $63,5\%$. Z uvedených výsledkov teda vyplývajú dve veci, ktoré treba brať do úvahy pri zložitejších metódach predikcie. Prvou je, že zvýšenie počtu použiteľných vzoriek za cenu vnesenia chyby do datasetu nezlepšuje

predikciu, ale ani ju výrazne nezhoršuje (aspoň pri lineárnej regresii) a druhou je fakt, že MAPE je citlivejšia na zmeny pri predikovaní intenzity zelenej vrstvy airglow-u, pretože táto obsahuje dáta z väčšieho rozsahu a s väčším rozptylom.

Čo sa týka polynomiálnej regresie druhého stupňa, tak má o čosi lepšie výsledky ako lineárna, ale diametrálny rozdiel to nie je. Počet atribútov stúpol až na hodnotu 299. MAPE sa pri obidvoch vrstvách pohybovala okolo 60%, MAE bola 214 R pre zelenú a 67 R pre červenú vrstvu. Pozoruhodné je, že pri polynomiálnej regresií boli predikcie o niečo lepšie pri nahradení chýbajúcich hodnôt priemerom ako pri ich vymazaní.

Vyskúšali sme aj polynomiálnu regresiu tretieho stupňa, pri ktorej počet atribútov vyšiel až na 2559. V tomto prípade kvôli nedostatočnému počtu vzoriek došlo k úplnému preučeniu. Tento výsledok síce zmiernilo doplnenie chýbajúcich hodnôt, ale stále bolo preučenie veľmi veľké. Problém sa nám podarilo vyriešiť vymazaním najmenej početných atribútov, čím sa celkový počet atribútov znížil na 559 a dosiahli sme najnižšiu chybu spomedzi všetkých regresií. Pre airglow 557,7 nm bola MAE = 189 R a pre airglow 630 nm MAE dosiahlo hodnotu 64 R. Celkovo však ani jedna regresia nie je schopná dosiahnuť výrazné zlepšenie.

Ďalším použitým algoritmom bola neurónová sieť, ktorú sme implementovali pomocou knižnice Keras. Keďže neurónové siete sú známe ako univerzálny aproximátor funkcie, tak by mala byť schopná lepšej predikcie.

Aj keď sú neurónové siete výborným nástrojom na rozpoznávanie obrazu, reči alebo textu, tak pri predikčných úlohách sú citlivejšie na nastavenie hyperparametrov. Keďže v ich prípade sa dá meniť počet skrytých vrstiev, počet neurónov, aktivačné funkcie, learning rate, ako aj počet učiacich epoch, je nájdenie vhodných nastavení zdĺhavejší proces. Pri modelovaní sme pracovali s neurónovou sieťou s jednou skrytou vrstvou. Ako najúspešnejšie nám vyšli aktivačné funkcie ‘relu‘ a ‘tanh‘. Prvá menovaná dosiahla pri predikcií zelenej čiary MAE = 155 R, pričom na to po-

trebovala 200 epoch. Najlepšie výsledky sme dosiahli pomocou ‘tanh‘ funkcie, ktorej výsledná chyba na testovacej množine bola 115 R. Potrebovala však vyšší počet neurónov na jednotlivých vrstvách. Najlepšie pracovala s počtom neurónov 64-32-1. Pre porovnanie ‘relu‘ postačovalo nastavenie 32-16-1. Learning rate sme v oboch prípadoch nastavili na 0,1, pričom po každých 100 iteráciách sme ho prenásobili konštantou 0,8. Týmto sme dosiahli počiatočnú rýchlejšiu konvergenciu k minimu a neskoršiu vyššiu citlivosť, aby minimum nebolo “preskočené“.

Počas trénovania sme pracovali s množinou 23 atribútov a počtom vzoriek 1 504. V rámci testovania sme vyskúšali použiť aj celý dataset 63 atribútov, kde však kvôli vymazaniu chýbajúcich hodnôt klesol počet vzoriek na 621. V tomto prípade bola MAE rovná 120 R, ale došlo k značnému preučeniu.

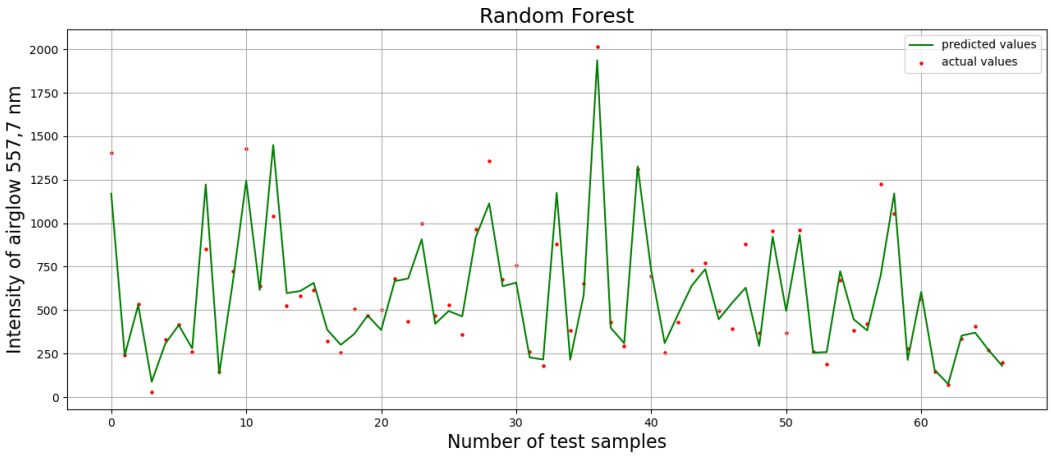
Pri modelovaní červenej čiary sme použili rovnaké nastavenia, teda aktivačnú funkciu ‘tanh‘, počet neurónov 64-32-1, počet epoch 500 a učiaci parameter 0,1. Na testovacej množine sme dosiahli chybu MAE = 48 R. Pri predikovaní tejto čiary oveľa častejšie dochádzalo k preučeniu, čo pripisujeme menšiemu rozptylu intenzít airglow-u v tejto vrstve.

V porovnaní úspešnosti neurónovej siete v predikcii vzhľadom na vrstvy, sa lepšie výsledky dosiahli pri červenej vrstve. MAPE pre vlnovú dĺžku 630 nm sa pohybovala okolo 80%, v prípade zelenej čiary chyba dosiahla približne 85%. Tieto hodnoty sú paradoxne ešte vyššie ako pri použití regresie. Tento jav má vysvetlenie, ktoré podávame v kapitole 4.5.

Nasledujúcim algoritmom, ktorý sme implementovali pri našom modelovaní, je Random Forest. Vďaka generovaniu veľkého počtu rozhodovacích stromov, pričom každý pri trénovaní používa inú podmnožinu dát, je tento algoritmus robustný a nie je citlivý na nastavenie hyperparametrov ako neurónové siete. Už s prednastavenými parametrami dosiahol pri predikcií o trochu lepšie výsledky. Pri trénovaní s množinou 23 atribútov pre zelenú vrstvu dosiahla priemerná absolútna chyba úroveň 108 R.

Jediným hyperparametrom, ktorý sme zmenili, bol počet generovaných rozhodovacích stromov. V základe je nastavený na 100, pre náš problém postačoval počet stromov 30. Chyba sa nijako nezmenila a zvýšila sa rýchlosť výpočtov. Zvyšovanie tohto parametra nemalo nijaký vplyv na výslednú chybu. Dochádzalo však k miernemu preučeniu, keďže na trénovacej množine bola MAE rovná 41 R. Z toho dôvodu sme začali meniť počet vstupných atribútov.

Random Forest poskytuje užitočnú funkciu „feature importance" (významnosť atribútov), ktorá ohodnotí každý atribút na základe toho, ako sa podieľal na predikcii. Túto funkciu sme využili na nájdenie najmenej významného atribútu, ktorý sme následne vymazali z nášho datasetu. Iteračným prístupom sme prešli všetky atribúty a zistili sme, že najlepšiu predikciu dosahujeme pri použití piatich: 'vzdialenosť Zem - Slnko', 'F 10.7 index', 'počet častíc atomárneho kyslíka', 'počet častíc atomárneho dusíka' a 'teplota'. Priemerná absolútna chyba testovacej množiny dosiahla 85 R, čo je naša zatiaľ najlepšia predikcia, a MAPE sa zastavilo na hodnote $21,8\%$. Grafické znázornenie predikcie je na obrázku 4 – 9.



Obrázok 4 – 9 Graf zobrazujúci predikciu intenzity zelenej vrstvy airglow-u pomocou algoritmu Random Forest (pozn. pre prehľadnosť sme z testovacej množiny náhodne vybrali len časť vzoriek).

Pri predikcií červenej vrstvy tento algoritmus nedosiahol až také zlepšenie predik-

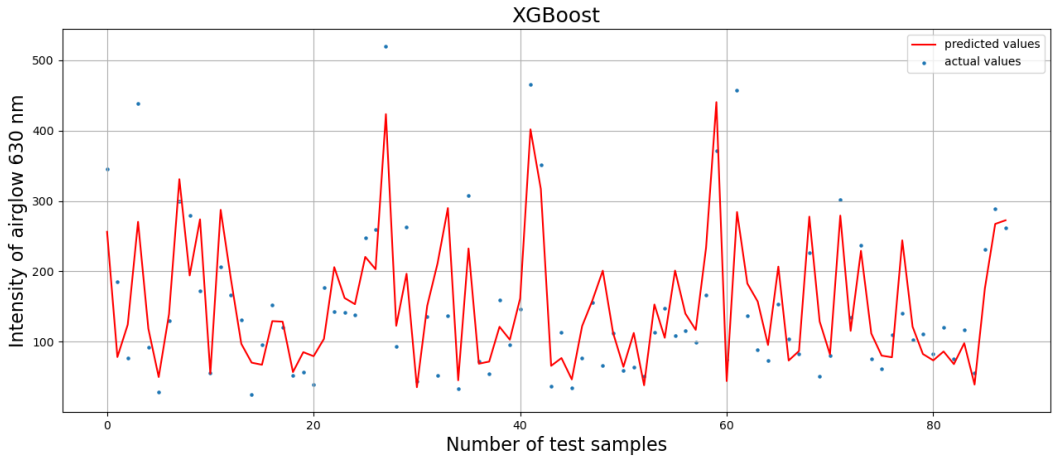
cie. Najnižšia chyba, ktorú sa nám podarilo dosiahnuť, bola 45 R, čo je porovnateľné s presnosťou neurónovej siete. MAPE bola na úrovni 35%.

Posledný algoritmus, ktorý sme použili pri modelovaní, je XGBoost. Ten, podobne ako Random Forest, pracuje s množstvom rôznych rozhodovacích stromov, vďaka čomu nie je potrebné zdĺhavé nastavovanie hyperparametrov. Vyžaduje si to však viac času ako pri Random Forest. V rámci našej implementácie sme testovali rôzne hodnoty učiaceho parametra, maximálnej hĺbky stromov, samotného počtu stromov a hodnotiacu funkciu.

Spomedzi viacerých hodnotiacich funkcií najlepšie výsledky vychádzali s MSE. Čo sa týka maximálnej hĺbky stromov, najideálnejšia hodnota bola v rozmedzí 14 až 17, keďže pri nižších hodnotách chyba mierne stúpala. Parameter počtu stromov, rovnako ako pri Random Forest, zvyšoval zložitosť trénovania a, bohužiaľ, bez lepších predikčných výsledkov. Najzaujímavejšie bolo nastavovať učiaci parameter, keďže tento mal najväčší dopad na výsledky predikcie. Príliš veľká hodnota (>0,3) spôsobovala väčšiu chybu ako Random Forest so základnými nastaveniami, a preto nebola vhodná. Na testovacej množine sme dosiahli najlepšie výsledky s učiacim parametrom medzi 0,3 a 0,2, v závislosti od predikovanej vrstvy. Na trénovacej množine dochádzalo k veľkému preučeniu. Na jeho redukciu bolo potrebné nastaviť učiaci parameter na hodnotu 0,15, ale za cenu väčšej chyby na testovacej množine.

Pre zelenú airglow vrstvu sme opäť dosiahli najlepšiu predikciu pri redukovanom datasete, v tomto prípade s 10-timi atribútmi. Priemerná absolútna chyba klesla až na 77 R, zatiaľ čo MAPE sa pohybovala okolo 21%.

Pre červenú kyslíkovú čiaru sme najnižšiu chybu dosiahli s celkovým datasetom, ktorý bol len očistený od vysoko korelujúcich atribútov. V tomto prípade mala MAE hodnotu 43 R a MAPE 34%. Výsledok predikcie je graficky znázornený na obrázku 4 – 10



Obrázok 4 – 10 Graf zobrazujúci predikciu intenzity červenej vrstvy airglow-u pomocou algoritmu XGBoost (pozn. pre prehľadnosť sme z testovacej množiny náhodne vybrali len časť vzoriek).

4.5 Vyhodnotenie

V tejto kapitole zhrnieme všetky kroky, ktoré sme vykonali pri riešení našej práce, porovnáme použité algoritmy a predstavíme najlepší model. Zároveň zhodnotíme, či boli naplnené ciele definované na začiatku a navrhneme možné vylepšenia a postupy, kam by sa mohol uberať ďalší výskum tejto problematiky.

Základom výskumu bolo pochopenie dát. Množina vybratých atribútov sa ukázala ako dostačujúca a následná analýza dokonca ukázala viaceré zaujímavé korelácie medzi niektorými veličinami, ktorých skúmanie môže priniesť lepšie pochopenie procesov prebiehajúcich v zemskej atmosfére. Výber správnych atribútov však nebol triviálny a bolo potrebné získať astronomické „know-how", ktoré nám poskytli konzultácie so skúsenými výskumníkmi a štúdium odbornej literatúry.

V následnej príprave dát boli údaje očistené od všetkých potenciálnych chýb, ktoré by mohli výsledok pozitívne alebo negatívne skresľovať. Tieto potrebné kroky znížili početnosť dát, ktorá ale zostala aj naďalej postačujúca. Významná informácia bola zachovaná hlavne vďaka veľkým časovým rozostupom medzi meraniami, ktoré

pokryli viaceré slnečné cykly.

Vo fáze modelovania sme postupovali od jednoduchších metód k zložitejším. Naším východiskovým bodom bol výsledok lineárnej regresie, ktorý stanovil základnú chybu. Následne sme na predikciu použili polynomiálnu regresiu druhého a tretieho stupňa. Pri druhom stupni obidve hodnotiace metriky vykazovali zníženie chyby, čo však úplne neplatilo pri regresií tretieho stupňa, kde pri použití rovnakých atribútov chyba narástla niekoľkonásobne. Došlo k výraznému preučeniu, keďže chyba trénovacej množiny bola minimálna. Tento efekt sa podarilo znížiť menším počtom atribútov, celkovo však ako lineárna, tak ani polynomiálna regresia nie sú dostatočne silnými nástrojmi pre takýto typ predikčného problému.

V modelovaní sme pokračovali neurónovými sieťami, pri ktorých sme zaznamenali opäť určité zlepšenie. Časové aj výpočtové nároky potrebné na dosiahnutie týchto výsledkov boli príliš veľké. Nastavovanie vhodných hyperparametrov si vyžadovalo značné množstvo pokusov. Neurónové siete sú nepochybne silným a dynamickým nástrojom v oblasti strojového učenia, ale na riešenie tohto predikčného problému existujú aj lepšie algoritmy.

To nás priviedlo k algoritmom Random Forest a XGBoost. Obidva sú založené na rozhodovacích stromoch, ale každý s nimi pracuje svojím spôsobom. Ako sme už spomínali pri modelovaní, Random Forest prekonal všetky doterajšie metódy už so základnými nastaveniami a s pôvodnou množinou atribútov. Ďalším trénovaním sa nám chybu podarilo ešte znížiť, pričom v tomto procese sme pracovali hlavne s rôznymi podmnožinami atribútov a hyperparametre sme menili iba minimálne.

Algoritmom XGBoost sme pokračovali v nastolenom trende a chybu sme stlačili ešte o trochu nižšie. Tu naša práca spočívala ako v hľadaní najvhodnejšej podmnožiny atribútov, tak aj v nastavovaní správnych hyperparametrov. Hlavnou výhodou XGBoost je schopnosť spracovať aj chýbajúce hodnoty, a vďaka tomu v príprave dát odpadáva nutnosť riešiť dilemu nahradenia alebo vymazania údajov. Naviac sa

	I 577,7		I 630	
	MAE	**MAPE**	**MAE**	**MAPE**
Lin. Regresia	224,60	68,59 %	72,23	63,50 %
Poly. Regresia (2.)	214,91	73,42 %	67,63	57,91 %
Poly. Regresia (3.)	189,83	69,70 %	64,95	54,65 %
Neurónové siete	115,49	85,84 %	48,60	79,75 %
Random Forest	85,73	21,78 %	45,48	35,81 %
XGBoost	**77,48**	**21,14 %**	**43,08**	**33,99 %**

Tabuľka 4 – 2 Tabuľka zobrazujúca úspešnosť predikcií jednotlivých algoritmov na testovacej množine.

pri dopĺňaní vnáša do dát falošná informácia a pri mazaní sa stráca istá informačná hodnota. Keďže sme skúšali trénovanie s úplnými aj s neupravenými dátami, tak vieme povedať, že XGBoost dokáže využiť aj tieto čiastkové informácie na zlepšenie výsledkov.

Z hodnôt uvedených v tabuľke 4 – 2 je vidieť (podľa metriky MAE), že najlepšiu predikciu dosahuje algoritmus XGBoost. Pre predikciu zelenej vrstvy je rovnako použiteľný aj Random Forest. Čo sa týka červenej vrstvy, tak medzi najlepšími troma algoritmami nie sú veľké rozdiely, avšak neurónové siete potrebujú na trénovanie oveľa viac času ako prvé dva algoritmy.

Zaujímavá je však aj metrika MAPE, ktorá pri neurónových sieťach dosahuje podstatne vyššie hodnoty ako pri ostatných algoritmoch, aj keď MAE je nízka. Toto sa dá vysvetliť tým, že neurónové siete svoje predikcie „predimenzujú", teda majú tendenciu predikovať vyššie intenzity airglow-u, ako sú skutočné hodnoty a metrika MAPE takéto predikcie penalizuje. Random Forest a XGBoost zväčša predikujú intenzity nižšie, ako sú reálne hodnoty. Je to vidieť aj na obrázkoch 4 – 9 a 4 – 10, v ktorých predikcie horných extrémov nie sú „dotiahnuté".

Airglow je prevažne lokálny jav, ktorý sa v priebehu noci celkom rýchlo mení. Výsledky dokazujú, že aj napriek tomu je možné ho predikovať pomocou strojového učenia s dostatočne dobrou presnosťou.

Pre dosiahnutie ešte lepších výsledkov by bolo vhodné získať väčšiu množinu dát, s kompletnejšími údajmi. Dnešná technika umožňuje presnejšie merania ako dosahovala technika v časoch, keď vznikal náš dataset, ktorého ozajstná hodnota spočíva v dĺžke obdobia, ktoré pokrýva.

Vhodným predmetom ďalšieho výskumu by bol gray-box model, ktorý by algoritmy strojového učenia využíval spolu s predprogramovanými fyzikálnymi zákonmi ovplyvňujúcimi airglow.

5 Záver

V našej práci sme podali teoretický pohľad na airglow z hľadiska jeho charakteristiky aj z hľadiska príčin vzniku. Opísali sme etapy procesu tvorby predikčného modelu spolu s funkcionalitou jednotlivých metód strojového učenia. Vybrali sme vhodné zdroje dát, ktoré sme zanalyzovali a spracovali do vhodného formátu pre riešenie úlohy predikcie. Realizovali sme samotné modelovanie a vyhodnotili jeho výsledky.

Spomedzi použitých algoritmov strojového učenia sa ako najvhodnejšie na predikciu takýchto javov ukázali XGBoost a Random Forest. Najlepšie zachytili charakter dát aj napriek istému vplyvu náhody na intenzity airglow-u. Naviac, potrebný čas na učenie ako aj výpočtové nároky boli minimálne. Taktiež nevyžadujú žiadne špeciálne postupy prípravy dát, čím urýchľujú predspracovanie a umožňujú plynulý prechod od návrhu k hotovému modelu. V našom prípade sme využili princíp „gray box-u", ktorý spočíval vo využití fyzikálne opodstatnených atribútov, vďaka ktorým sme dosiahli cieľ stanovený na začiatku. Tým sme potvrdili, že metódy strojového učenia sú nielen použiteľné, ale pomaly sa stávajú až nevyhnutnými vo vesmírnom výskume.

Pre dosiahnutie ešte lepších výsledkov je potrebné zozbierať väčší počet dát, prípadne implementovať známe fyzikálne závislosti priamo do modelu a strojové učenie využívať ako doplnok na predikciu nepriamych vplyvov.

Vesmírny výskum je komplexná vedecká činnosť, v ktorej vie umelá inteligencia plniť úlohu plnohodnotného pomocníka. Je veľmi pravdepodobné, že v budúcnosti bude spojenie medzi astronómiou a strojovým učením silnieť a pokrok jedného odvetvia bude znamenať úspech oboch. Preto by sme radi aj v diplomovej práci ďalej rozvíjali toto prepojenie.

Literatúra

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.
URL: *http://dx.doi.org/10.1023/A3A1010933404324*

Buhgin, H. E. (2017). The new spring of artificial intelligence: A few early economies.
URL: *https://voxeu.org/article/new-spring-artificial-intelligence-few-early-economics*

Chakure, A. (2019). Random forest regression.
URL: *https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f*

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. R. H. and Wirth, R. (2000). Crisp-dm 1.0: Step-by-step data mining guide.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system.
URL: *https://arxiv.org/pdf/1603.02754.pdf*

Chollet, F. et al. (2015). Keras.
URL: *https://github.com/fchollet/keras*

George, D. and Huerta, E. (2018). Deep learning for real-time gravitational wave detection and parameter estimation: Results with advanced ligo data, *Physics Letters B* **778**: 64 – 70.
URL: *http://www.sciencedirect.com/science/article/pii/S0370269317310390*

Ghodpage, R. (2016). *Upper atmospheric study using night airglow*, Vol. 1, Lambert Academic Publishing.

Ho, R. Y. N., Liebman, J. F., Valentine, Joan Selverstone", e. C. S., Valentine, J. S., Greenberg, A. and Liebman, J. F. (1995). *Overview of the Energetics and*

Reactivity of Oxygen, Springer Netherlands, Dordrecht, pp. 1–23.
URL: *https://doi.org/10.1007/978-94-007-0874-7_1*

Ioannou, Y. (2017). *Structural Priors in Deep Neural Networks*, PhD thesis.

Isobe, T., Feigelson, E. D., Akritas, M. G. and Babu, G. J. (1990). Linear Regression in Astronomy. I., *apj* **364**: 104.

Joshi, N. (2019). How far are we from achieving artificial general intelligence?
URL: *https://www.forbes.com/sites/cognitiveworld/2019/06/10/how-far-are-we-from-achieving-artificial-general-intelligence/*

Mandot, P. (2019). How exactly xgboost works? - pushkar mandot - medium.
URL: *https://medium.com/@pushkarmandot/how-exactly-xgboost-works-a320d9b8aeef*

Moran, M. E. (2006). The da vinci robot, *Journal of Endourology* **20**(12): 986–990. PMID: 17206888.
URL: *https://doi.org/10.1089/end.2006.20.986*

Morde, V. and Setty, V. A. (2019). Xgboost algorithm: Long may she reign! - towards data science.
URL: *https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d*

NASA ((accessed March 14, 2020)). Interface to produce plots, listings or output files from OMNI 2, `https://omniweb.gsfc.nasa.gov/form/dx1.html`.

NASA/CCMC ((accessed May 24, 2020)). NRLMSISE-00 Atmosphere Model , `https://ccmc.gsfc.nasa.gov/modelweb/models/nrlmsise00.php`.

Olhede, S. and Wolfe, P. (2018). The ai spring of 2018, *Significance* **15**(3): 6–7.
URL: *https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1740-9713.2018.01140.x*

Oppy, G. and Dowe, D. (2019). The turing test, *in* E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, spring 2019 edn, Metaphysics Research Lab, Stanford University.

pandas development team, T. (2020). pandas-dev/pandas: Pandas.
URL: *https://doi.org/10.5281/zenodo.3509134*

Pandey, P. (2019). Understanding the mathematics behind gradient descent.
URL: *https://towardsdatascience.com/understanding-the-mathematics-behind-gradient-descent-dde5dc9be06e*

Pant, A. (2019). Introduction to linear regression and polynomial regression.
URL: *https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb*

Pearson, K. A., Palafox, L. and Griffith, C. A. (2017). Searching for exoplanets using artificial intelligence, *Monthly Notices of the Royal Astronomical Society* **474**(1): 478–491.
URL: *https://doi.org/10.1093/mnras/stx2761*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**: 2825–2830.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers, *IBM Journal of Research and Development* **3**(3): 210–229.

Savigny, C. (2017). Airglow in the earth atmosphere: basic characteristics and excitation mechanisms, *ChemTexts* **3**.

Sinčák, P. and Andrejková, G. (1996). *Neurónové siete Inžiniersky prístup (1. diel).*

Skalski, P. (2018). Deep dive into math behind deep networks - towards data ...
URL: *https://towardsdatascience.com/https-medium-com-piotr-skalski92-deep-dive-into-deep-networks-math-17660bc376ba*

Slanger, T., Cosby, P., Huestis, D. and Bida, T. (2001). Discovery of the atomic oxygen green line in the venus night airglow, *Science (New York, N.Y.)* **291**: 463–5.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*, CreateSpace, Scotts Valley, CA.

Zucker, S. and Giryes, R. (2018). Shallow transits—deep learning. i. feasibility study of deep learning to detect periodic transits of exoplanets, *The Astronomical Journal* **155**(4): 147.
URL: *https://doi.org/10.3847%2F1538-3881%2Faaae05*

Zoznam príloh

Príloha A Systémová príručka

Príloha B Používateľská príručka

Príloha C CD médium obsahujúce všetky zdrojové kódy, bakalársku prácu a prílohy v elektronickej podobe