4. félévi beszámoló

Kunsági-Máté Sándor
E-mail: kunsagisandor@gmail.com
PhD program: Részecskefizika és csillagászat
Témavezető: Dobos László
A dolgozat címe: Gépi tanulás a csillagászatban

2020. május 29.

1. Bevezetés

Az extragalaktikus csillagászatban a vöröseltolódást becslő eljárásoknál már régóta használnak Deep Learning és gépi tanulási módszereket. A photo-z eljárás során a galaxisképekből kiredukált színindexek alapján adunk becslést a fotometrikus vöröseltolódásra.

A gyakorlatban sokszor előfordul azonban, hogy a galaxisok látóirányába esnek tejútrendszerbeli csillagok (előtércsillagok) is. Emiatt a meghatározott színindexek nem csak az adott galaxishoz fognak tartozni, továbbá maga a galaxis sem egyetlen színkomponensből tevődik össze (pl. HII régiók, csillagközi port tartalmazó régiók), ezért az integrált magnitúdók használata nagyobb szórást eredményezhet a fotometrikus vöröseltolódás számításában.

Fontos megemlítenünk továbbá, hogy az újabb égboltfelmérési programok (pl.: Large Synoptic Survey Telescope, LSST) a jelenleginél sokkal érzékenyebb fotometriát fognak végezni, vagyis az univerzum nagyobb térfogatát tudják majd vizsgálni. Emiatt a felvételeken sokkal nagyobb számsűrűséggel lesznek jelen az észlelt galaxisok, és így megnő a látszólag átfedő galaxisok előfordulási valószínűsége is ([1]). A színindexek meghatározásánál ebben az esetben is különösen fontos a két objektum megfelelő szétválasztása.

Kutatómunkám központi témája a galaxisok képeinek színtérbeli/csillagpopulációk szerinti szegmentálása a Dark Energy Survey (DES) G, R, I, Z színszűrőiben készült felvételek alapján.

2. Kutatás az első három félévben

A galaxisképek szegmentációjához először a Shi & Malik képszegmentációs algoritmus [2] használhatóságát vizsgáltam meg. A módszer során elkészítjük a kép pixeleinek összekötöttségi

gráfját, ahol a gráf csúcsai a pixelek, a köztük lévő élek súlyai pedig a pixelek közötti hasonlóságot jellemzik. A legfontosabb feladat a W súlymátrix elkészítése, ahol a mátrix egyes elemeit az ún. súlyfüggvénnyel határozzuk meg:

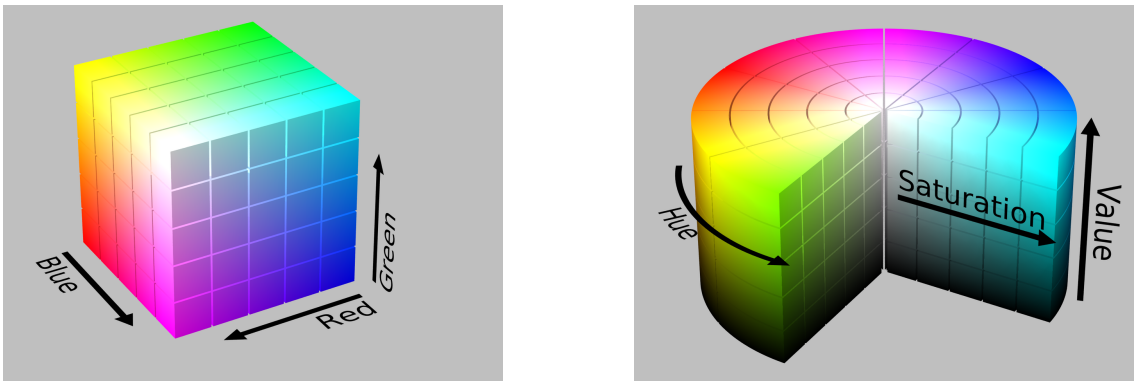
$$W_{ij} = e^{-\frac{||F(i)-F(j)||_2^2}{\sigma}} \tag{1}$$

Itt az F az úgynevezett "feature" vektort jelöli, mely elemei például lehetnek az egyes színindexek. Mivel a színeket az egyes színszűrők különbségeiből határozzuk meg, ezért az SDSS-nél jobb minőségű DES képeken jelenlévő zaj is jelentős relatív zajt okoz a színtérben az egyes pixelekre vonatkozóan. Ennek kezelésére az egyik megoldás az lehet, ha az egyes súlyokat a relatív zaj értékekkel súlyozzuk (vagyis a σ jelentse a relatív zajt). Ez a képszegmentációs eljárás iteratív módon hajtható végre, ahol minden egyes iteráció során a pixeleket két halmazba soroljuk (részletes leírás ld. 1. félévi beszámoló).

A kapott eredményeim alapján a Shi&Malik algoritmus nem bizonyult alkalmasnak a tipikusan nagy fényességgradienssel és kis színkontraszttal (a fényességértékek relatív szórásához képest sokkal kisebb színindexbeli relatív szórással) rendelkező galaxisképek színek szerinti szegmentációjára.

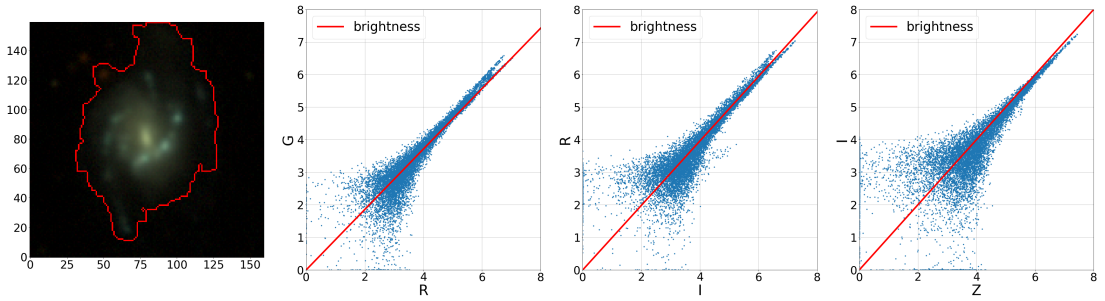
Ezek után a K-means klaszterezési eljáráshoz fordultam. Itt a legfontosabb feladat az egyes pixelek színét legjobban jellemző "tulajdonságvektorok" előállítása. Fontos szempont, hogy olyan koordinátarendszert használjunk, mely nem tartalmazza a pixelek fényességét. Erre azért van szükségünk, mivel az azonos színű, de eltérő méretű csillagpopulációk fényességben különbözhetnek egymástól.

A digitális fotófeldolgozásban használt RGB és HSV színmodellek mintájára a galaxisok pixeleinek jellemzésére is bevezethetünk egy magasabb dimenziós általánosított színmodellt. A HSV színtérmodell legfőbb erőssége, hogy a fényességet és a színkomponenseket egymástól lineárisan független komponensekként reprezentálja, ellentétben az RGB modellel (ld. 1. ábra).



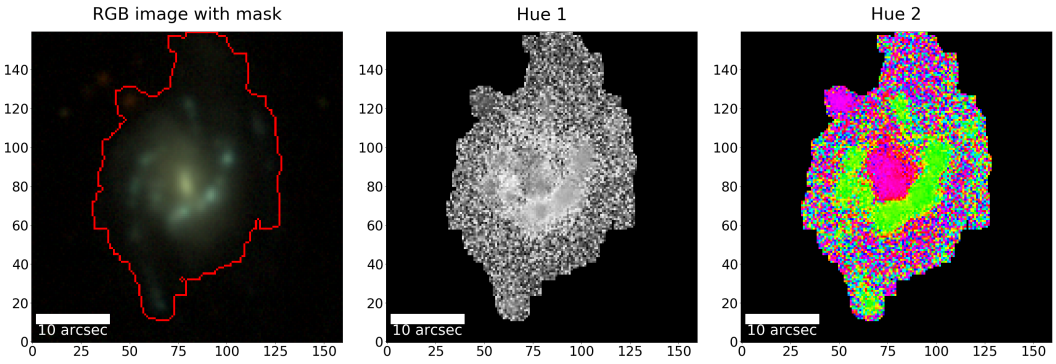
1. ábra. RGB és HSV színtérmodellek (forrás: https://en.wikipedia.org/)

A szegmentáláshoz használt négy (G, R, I, Z) színszűrő egy négydimenziós színteret fog kifeszíteni. Mivel a galaxisképek jelentős fényességgradienssel rendelkeznek – a pixelek intenzitásértékei közelítőleg egy egyenesen helyezkednek el –, ezért a fényesség irányát főkomponens analízissel (PCA) közelíthetjük (ld. 2. ábra).



2. ábra. Kék spirálkarokkal és vörös maggal rendelkező galaxis G, R, I színszűrőkből generált hamisszínes képe, valamint a piros határolóvonalon belül lévő pixelek magnitúdó értékeinek eloszlása a `numpy.arcsinh` függvénnyel eltranszformált 4-dimenziós tér *G-R*, *R-I*, *I-Z* metszeteiben.

A következőkben levetítjük a 4-dimenziós tér pontjait az origón átmenő, a fényesség irányára merőleges hipersíkra, így egy háromdimenziós alteret kapunk, amely pontjait írjuk fel gömbi koordinátarendszerben. Ekkor – a megszokott jelölésrendszert követve – r a szaturációnak, $\theta \in [0, \pi]$ és $\phi \in [0, 2\pi]$ pedig két "hue" szögnek feleltethető meg. A szaturáció a fényesség tengelyétől mért távolságot jelenti, mely azonban jelentős relatív zajt tartalmaz, hiszen a fényességben lévő szóráshoz képest a szaturáció irányában sokkal kevésbé szórnak a pontok. Ennek következtében a szegmentációs eljáráshoz csak a két hue szöget alkalmaztam (ld. 3. ábra):



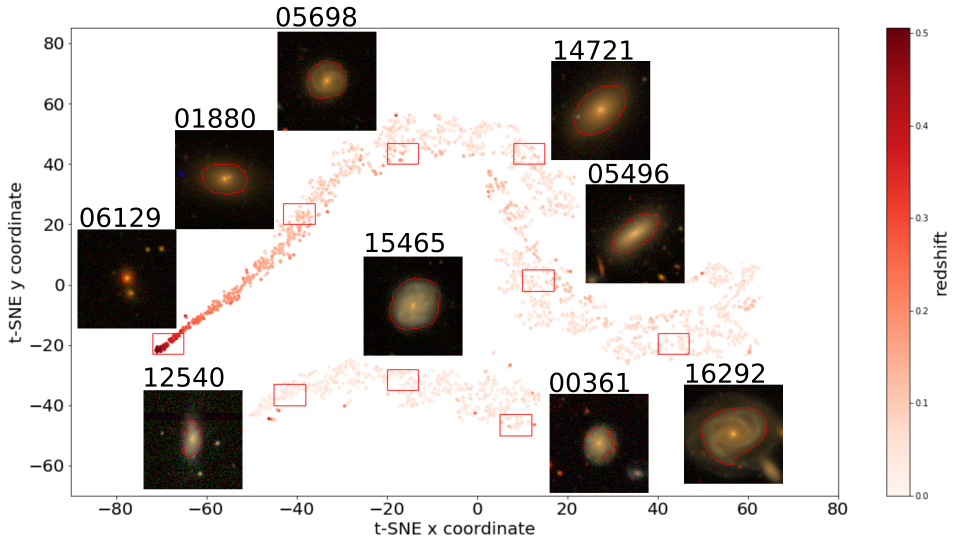
3. ábra. A galaxis RGB képe, valamint a *Hue 1* és *Hue 2* koordináták. Megj.: a "Hue 2" ($\phi \in [0, 2\pi]$) periodikus koordináta, ezért az ábrázoláshoz színkódolást alkalmaztam.

A várakozásoknak megfelelően a mag és a spirálkarokhoz tartozó pixelek jelentősen eltérnek egymástól a Hue 1-Hue 2 által meghatározott térben. Ennek köszönhetően az eltérő színű területeket eredményesen tudtam szuperpixelekre bontani, és azok a színindexek terében is szignifikánsan elkülönültek egymástól (ld. 2. félévi beszámoló). Kutatómunkám során megvizsgáltam a szuperpixelekre kapott színindexek photo-z eljárásoknál történő használhatóságát is. Eddigi eredményeim azt mutatták, hogy nem érhető el nagyobb pontosság a vöröseltolódás

becslésében a szuperpixelekre számolt színindexekből, melynek egyik fő oka a színindexek nagyobb relatív zajának tulajdonítható (ld. 3. félévi beszámoló).

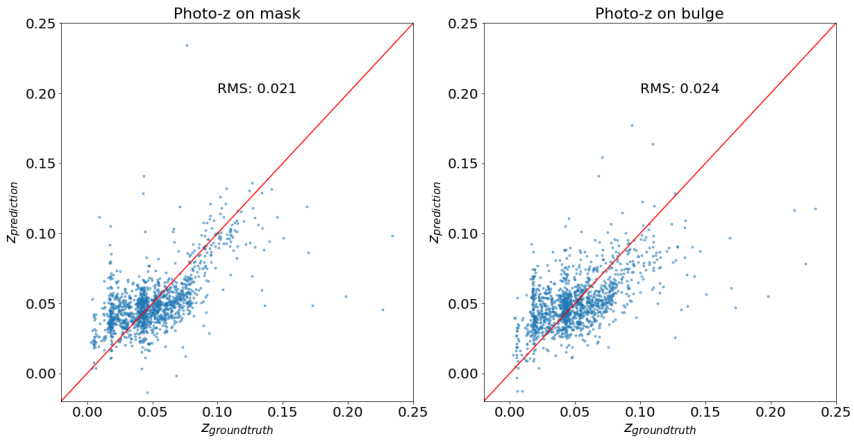
3. Kutatás a negyedik félévben

A negyedik félév elején a szegmentáló algoritmust ismét a photo-z becslés szempontjából vizsgáltam meg. Elsősorban olyan galaxisoknál várunk észrevehető javulást a vöröseltolódás-becslés pontosságában, ahol egymással összemérhető kiterjedésben vannak jelen az eltérő csillagpopulációk a galaxisképeken. Ennek a feltételnek leginkább a spirálgalaxisok felelnek meg (spirálkarok, mag), ezért szükséges volt a szegmentáláshoz használt mintából a spirálgalaxisok leválogatása. Ezt úgy végeztem el, hogy a galaxisokra külön-külön meghatároztam a fényesség irányába mutató 4-dimenziós egységvektorokat, melyek más és más irányba mutatnak aszerint, hogy kékebb vagy vörösebb galaxisról van szó. Ezután a t-SNE [3] eljárás segítségével egy 2 dimenziós síkra képeztem le a vektorok koordinátáit (ld. 4 ábra).



4. ábra. Galaxisok fényességének irányába mutató egységvektorok leképezése 2 dimenziós altérre a t-SNE eljárás segítségével. A piros téglalapok által meghatározott területről véletlenszerűen választott galaxisok képeit ábrázoltam. A piros színkódolás a galaxisok vöröseltolódását jelöli.

Az adatokat K-means klaszterezéssel két csoportra osztottam, így összesen kb. 1300 spirálgalaxist azonosítottam. Ezekre elvégeztem a szegmentációt, majd a magra és a teljes galaxisra (maszk) meghatároztam a színindexeket, és összehasonlítottam a lokális lineáris regresszióval kapott vöröseltolódásértékeket (ld. 5. ábra).



5. ábra. Lokális lineáris regresszióval kapott vöröseltolódás becslés a maszkon illetve a bulgehoz tartozó szuperpixelen számolt színindexek alapján.

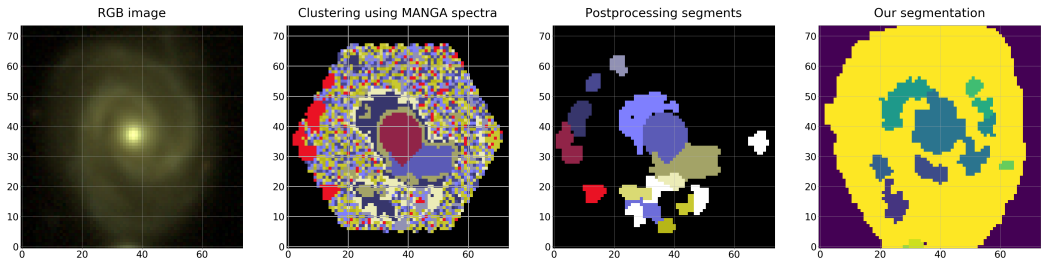
A kapott RMS értékek azt mutatják, hogy továbbra sem sikerült jobb eredményt elérni a szuperpixelek felhasználásával. Az eredményt befolyásoló tényezők között említhetjük: (i) kevés számú spirálgalaxis; (ii) a galaxisok túlnyomó többsége közeli (z<0.1), ahol tipikusan alacsony pontossággal rendelkezik a photo-z becslés (mely egyik oka a galaxisok nagyobb morfológiai változatossága); (iii) a szuperpixelek színindexei nagyobb relatív hibával rendelkeznek, mint a maszk esetében.

A photo-z becslés mellett a szegmentáló algoritmus verifikálásával foglalkoztam. Az SDSS égboltfelmérési program MaNGA elnevezésű küldetése során kb. 10000 galaxison végeztek el integrálismező-spektroszkópiát, ahol a galaxis több száz pontjában mérték meg a spektrumot. A spektrumok egy független módszert biztosítanak a galaxisképek szuperpixelekre bontásához, amellyel az általam készített szegmentáló algoritmus eredményét összehasonlíthatjuk.

Ahhoz, hogy a fényesség szerint ebben az esetben se klaszterezzünk, az egyes spektrumokat a 90. percentilissel normáltam. Ezután a `UMAP` [4] dimenzióredukciós eljárással egy síkra vetítettem a spektrumokat. A vetítés során a következő metrikát alkalmaztam az n elemű s_1 és s_2 spektrumok távolságának definiálásához:

$$D(s_1, s_2) = \exp\left(-\max\left(\frac{\mathrm{corr}(s_1, s_2)}{N}\right)\lambda\right) \tag{2}$$

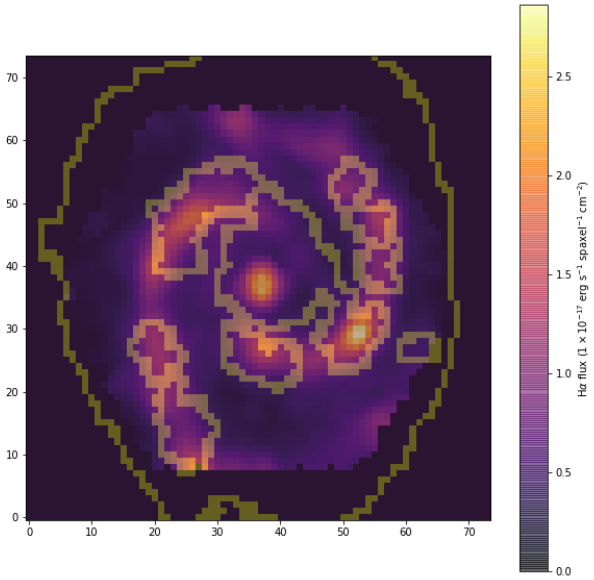
, ahol λ egy skálafaktor, a $\mathrm{corr}(s_1, s_2)$ a spektrumok normálás nélküli keresztkorrelációját jelenti (a `numpy.correlate()` függvény ilyen), N pedig egy $2n-1$ elemű vektor, ami megfelelő sorrendben tartalmazza a keresztkorreláció során aktuálisan átfedő elemek számát, azaz ezzel lenormáljuk a korreláció eredményvektorát. A levetített spektrumokat Kmeans klaszterezéssel csoportosítottam, majd a szegmentáló algoritmusomhoz hasonló utófeldolgozással elkészítettem a szegmenseket. Az előzetes eredményeket a 6. ábrán láthatjuk.



6. ábra. Balról jobbra: A 8154-12704 MaNGA azonosítójú galaxis RGB képe; a UMAP által levített 2 dimenziós térben klaszterezett spektrumok elhelyezkedése a képen; PSF (pont-terülési függvény) félértékszélességénél kisebb területek elhagyása; valamint a színszűrők felhasználásával kapott szegmensek.

Az eredmények azt mutatják, hogy leginkább csak a középső, nagy jel/zaj arányú területeken fednek át a kétféle módszerből kapott szegmensek.

Ezután összevetettem a szegmenseket a Hα régiók eloszlástérképével is (ld. 7 ábra):



7. ábra. A 8154-12704 MaNGA azonosítójú galaxis Hα térképe, valamint a halvány sárga körvonallal jelzett szegmensek.

Láthatjuk, hogy a meghatározott szegmensek nagyrészt egybeesnek a 656 nm-hez tartozó Hα régiókkal. Ennek az lehet az egyik magyarázata, hogy ezek a régiók csak a DES r szűrőjében láthatók, így a színszűrőkből képzett színtérben ezek a területek szignifikáns klasztert fognak alkotni.

4. Publikációk

A félév alatt a szegmentáló algoritmus elméleti és technikai hátterét összefoglaltam, a módszerről szóló publikáció jelenleg kb. 80%-os készültségi szinten van, várhatóan júniusban küldjük be elbírálásra.

5. Tervezett konferenciák

ADASS nemzetközi konferencia 2021-ben (2020-as kiírás: https://adass2020.es/)

6. Oktatási és egyéb tevékenység a 4 félév alatt

- Programozási alapismeretek (progalapf17va) laborgyakorlatok az 1. és 3. félévben
- pontozói részvétel a Keszthelyen megrendezett 13. Nemzetközi Csillagászati és Asztrofizikai Diákolimpián (IOAA, 2019. augusztus 2-10.)

7. Elvégzett kurzusok a félévben

- Csillagrendszerek dinamikája I. (FIZ/2/027E), oktató: Dr. Balázs Lajos

Hivatkozások

[1] Daniel M. Jones, Alan F. Heavens, MNRAS, Volume 483, Issue 2, February 2019,

[2] J. Shi and J. Malik. Normalized cuts and image segmentation. In Proc. IEEE Conf. Computer Vision and Pattern Recognition, pages 731-737, 1997.

[3] van der Maaten, L.J.P.; Hinton, G.E. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9:2579-2605, 2008.

[4] McInnes, L, Healy, J, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018