Astrometria - Projeto 1

Determinando a Distância de Aglomerados Abertos com Inferência Bayesiana

André Almeida Trovello[1]

25 de setembro de 2025

Resumo

Modelos de síntese de população estelar nos possibilitam estudar e obter as propriedades de galáxias ou aglomerados a partir de seus espectros integrados. Como todo modelo, estes apresentam imprecisões, podendo elas estarem relacionadas à evolução estelar, função de massa inicial ou ao tipo de espectro estelar utilizado. Os modelos podem ser construídos a partir de bibliotecas estelares empíricas ou teóricas, dos quais as limitações estão relacionadas à cobertura do diagrama HR, no primeiro caso, e ao limite de conhecimento físico que possuímos sobre as atmosferas estelares, no segundo. A literatura se refere ao "efeito sintético" como sendo as alterações nos modelos de população estelar que ocorrem quando uma biblioteca estelar empírica é trocada por uma teórica. Este projeto possui como objetivo verificar as diferenças entre modelos de população estelar, comparando os modelos baseados na biblioteca estelar empírica MILES com os modelos baseados na biblioteca teórica SynCoMiL. Esperamos estudar os efeitos sintéticos, identificando assim as regiões do espectro que mais divergiram e necessitam ser modeladas com maior atenção. Neste primeiro semestre de iniciação científica foram desenvolvidas as seguintes atividades: estudo bibliográfico, manuseamento e limpeza de dados, e reprodução de resultados prévios da literatura. A primeira englobou todo o estudo astronômico sobre modelos de população estelar e técnico (Python e R). A segunda envolveu o aprimoramento dos meus conhecimentos em tratamento de dados. Já a última ateve-se a reproduzir resultados disponíveis na literatura, como forma de verificar se os algoritmos utilizados em meus códigos estavam corretos e poderiam ser utilizados para avaliar novos modelos. Os resultados obtidos se mostraram parcialmente compatíveis com os apresentados no artigo tomado como base para este projeto. No cálculo de diferenças de índices normalizadas, porém, há uma discrepância cuja origem não foi identificada, havendo a suspeita de que esta ocorreu devido ao diferente funcionamento das bibliotecas do Python e do R.

[1] andretrovello@usp.br

1 Introdução

A determinação de distâncias é uma questão relevante na astronomia contemporânea. Em especial, a determinação de distâncias de aglomerados estelares se mostra primordial para refinar modelos de formação e evolução estelar, visto que o cálculo dessa grandeza possibilita a conversão de valores aparentes de magnitudes em absolutos, o que possibilita a obtenção acurada de parâmetros como idade e massa das estrelas que compõem o aglomerado.

Tradicionalmente, a distância destes aglomerados é calculada pela inversão das paralaxes observadas em seus membros, porém, este método se mostra altamente sujeito a erros aleatórios quando são estudados aglomerados mais distantes, já que nestas ocasiões o erro da paralaxe pode apresentar mesma ordem de grandeza que a própria paralaxe em si.

No contexto deste problema, a missão espacial Gaia contribuiu significativamente, fornecendo paralaxes de alta precisão de aglomerados e estrelas. Como toda medida experimental, os dados da Gaia possuem um erro associado, sendo este uma combinação de três fatores principais: erros aleatórios (a incerteza estatística individual de cada fonte), erros sistemáticos (oriundos de fatores como posição, cor e magnitude da estrela, que afetam estrelas próximas no céu de maneira semelhante) e o zero point (um offset de medição sistemático associado ao próprio instrumento Gaia).

Uma população estelar simples (SSP) é um grupo de estrelas que surgiu a partir da mesma nuvem de gás, possuindo assim as mesmas idades e metalicidades. Apesar de serem apenas modelos teóricos, visto que já foi constatado o fato de aglomerados e galáxias não serem compostos por apenas um grupo estelar de propriedades iguais, as SSPs ainda são muito úteis para entendê-los, já que estes corpos podem ser descritos como um conjunto de várias delas.

Dessa forma, modelos de síntese de população estelar (SPS) possuem um valor incalculável para a astronomia. Sua existência nos permite estudar as propriedades de galáxias e aglomerados em que não é possível resolver individualmente as estrelas, nos possibilitando inferir diversas características sobre a evolução deles. (**??**)

Para a confecção destes modelos, são considerados conceitos de evolução estelar (isócronas), função de massa inicial (IMF), e bibliotecas de espectros estelares. Neste último, a SPS pode ser feita a partir de bibliotecas empíricas ou teóricas. (**??**)

Nas bibliotecas empíricas, as estrelas são observadas, enquanto nas teóricas, os espectros são feitos a partir do conhecimento físico que se possui sobre as atmosferas estelares. Em ambos os casos, existem limitações que afetam a confiabilidade do resultado final. No primeiro, questões como a cobertura do diagrama Hertzprung-Russel (HRD), precisão de parâmetros atmosféricos da estrela, e relação sinal-ruído podem acabar comprometendo. Já no segundo, a limitação do nosso conhecimento sobre atmosfera estelar e opacidade

atômica são fatores agravantes.

As diferenças entre os resultados obtidos têm sido alvo de estudo dos astrônomos por muito tempo. (**??**) discute essas divergências, verificando o impacto que a escolha da biblioteca tem sobre as previsões de cores e espectros das SSPs, e por consequência, sobre a idade e metalicidade inferidas a partir desses modelos. Em linhas gerais, é concluido que, em grande parte dos casos, uma maior cobertura do HRD pode provocar mais diferenças significativas do que as imprecisões dos espectros sintéticos. Entretanto, no caso de certos índices espectrais (CaHK, Ca4227, Fe4668, Fe5270), o efeito sintético é considerável, indicando a necessidade de uma maior atenção na modelagem das grades teóricas.

Portanto, o objetivo deste projeto foi estudar o efeito sintético causado pela biblioteca teórica num modelo de população estelar e compreender em quais destes índices espectrais este efeito foi mais considerável, possibilitando assim que os modellers pudessem entender as regiões do espectro eletromagnético que precisam ser modeladas com maior atenção, melhorando assim a qualidade do modelo.

2 Desenvolvimentos

Neste primeiro semestre de projeto, as atividades se dividiram em três etapas: estudo bibliográfico, manuseamento e limpeza dos dados e reprodução de resultados prévios da literatura.

2.1 Estudo Bibliográfico

Esta primeira etapa teve como finalidade adquirir os conhecimentos necessários para o entendimento, tanto do aspecto astronômico do projeto, quanto do técnico. Dessa forma, inicialmente foi feita a leitura e compreensão de artigos sobre modelos de população estelar (**??**), bibliotecas teóricas e empíricas, bem como as diferenças entre o efeito sintético e de cobertura (**??**) e (**??**). Além disso, sabendo que o último passo do projeto visava utilizar o código Starlight (**??**), foi feito também um estudo desta ferramenta, capaz de, a partir de modelos de síntese de população estelar, destrinchar o percentual de SSPs que compõem uma galáxia observada, além de estimar sua idade, metalicidade e extinção visual.

Compreendida a astronomia, foi então estudada a questão técnica do projeto, centrada basicamente em programação. Dentre todas as possíveis linguagens para a execução das tarefas necessárias, foram escolhidas Python e R pela sua grande versatilidade em análise de dados, sendo a primeira utilizada de fato, e a segunda servindo como base para comparação visto que a orientadora deste projeto, Prof. Paula R. T. Coelho, possui

grande experiência com R, fornecendo assim códigos que pudessem ser estudados para compreender a lógica por trás deles.

Assim, no Python, foram estudadas principalmente as bibliotecas *Numpy*, capaz de realizar operações complexas em arrays, *Pandas*, que possibilita o manuseamento de dados, e *Matplotlib*, que permite a visualização deles. Já no R, foi dada muita ênfase à biblioteca *Tidyverse*, robusto pacote de análise de dados capaz de realizar todas as tarefas já citadas anteriormente.

2.2 Manuseamento e Limpeza de Dados

Encerrada a primeira etapa teórica, foi então o momento de iniciar de fato a prática, ou seja, aplicar todos os conhecimentos adquiridos para manusear e limpar os dados dos modelos de população estelar.

Foram acessados três modelos, apresentados em (**??**). O primeiro, nomeado cbc, comporta a biblioteca teórica Coelho 14 (C14), que possui uma extensa cobertura do diagrama Hertzprung-Russel (HRD). Neste caso, os arquivos dividem-se em dois formatos: FITS e gz.

Flexible Image Transport System (FITS) é um formato de arquivo muito utilizado em astronomia, capaz de carregar consigo diferentes tipos de informação em formatos variados, sendo eles unidades de dados (HDUs) formados por vetores 1D, 2D ou 3D multi dimensionais, além de ASCII headers capazes de transportar informações adicionais ou introdutórias, como coordenadas. Com essa alta flexibilidade e capacidade de armazenamento, diferentemente de imagens convencionais, FITS permite guardar quantidades considerávis de dados que facilitam visualização e análise. Para a leitura deste formato de arquivo, foi utilizada a biblioteca astropy, construída com foco em resolver questões astronômicas em Python.

Já gz é um arquivo compactado utilizando o software GNU zip (gzip), muito utilizado no Linux e Windows.

O segundo modelo (SPS-M) comporta a MILES, uma biblioteca empírica que possui o fluxo calibrado de aproximadamente 1000 estrelas, cobrindo a faixa de 3540-7410Å do espectro eletromagnético. Já o último modelo (SPS-S) utiliza da biblioteca teórica SynCoMiL, que apresenta todas as suas estrelas geradas por modelagem física, porém, foi limitada a apresentar exatamente a mesma cobertura em termos de comprimento de onda e HRD que a MILES. Tanto na SPS-M quanto na SPS-S, os arquivos foram apresentados nos formatos FITS e txt.

Abertos os arquivos no Python, iniciou-se o procedimento de limpeza dos dados. Foram então removidos os índices que apresentavam dados idênticos em todas as linhas e filtrada a base de dados para que fossem apenas utilizados os pontos onde $7 \leq logage \leq 10$, sendo log age o logaritmo da idade das populações estelares do modelo. Feito isso,

foi concluído o tratamento dos dados, permitindo assim a etapa de visualização e análise destes.

2.3 Reprodução de Resultados Prévios da Literatura

A última tarefa deste primeiro semestre foi a reprodução de resultados prévios apresentados na literatura. Dado que o processo de análise e comparação das bibliotecas é o mesmo em todos os casos, para poder avaliar a qualidade de novos modelos de população estelar foi necessário primeiramente verificar se os algoritmos utilizados no Python reproduziam os mesmos resultados obtidos na literatura. Se isso fosse confirmado, era possível inferir então que os códigos utilizados estavam corretos, podendo assim serem aplicados em modelos mais recentes.

Dessa forma, iniciou-se então um procedimento de reproduzir as figuras do trabalho publicado por minha orientadora em (**??**), dando ênfase naquelas que evidenciavam o efeito sintético, objeto de estudo deste projeto. Desta maneira, a primeira tentativa de reprodução foi a da figura 8 do artigo, apresentada na figura deste projeto. Estes plots demonstram uma comparação direta entre os índices espectrais dos modelos SSP calculados com as bibliotecas MILES e SynCoMiL, sendo possível observar por meio de uma linha 1-1, ou seja, uma reta $f(x) = x$, se estes resultados são equivalentes ou não.

Os resultados reproduzidos por mim estão apresentados na figura . Analisando as figuras obtidas, nota-se que a reprodução dos resultados do artigo foi atingida com sucesso, visto que, em todos os casos, os plots foram equivalentes. Além disso, é perceptível que na maioria dos índices os pontos se mantiveram sobre a linha 1-1, indicando igualdade entre as duas bibliotecas.

Entretanto, alguns índices apresentaram divergências consideráveis, sendo eles Ca4227, Fe4668, Mg1, Mg2, Fe5406, Fe5709 e B4000Vn. Nestes casos, houve uma disparidade entre os valores da MILES (SPS-M) e SynCoMiL (SPS-S), onde os pontos acabaram distoando da linha 1-1. Além disso, houveram casos como H83889, H93885 e H103798 onde, apesar de haver equilíbrio sobre a linha 1-1, houve uma dispersão maior dos pontos em torno dela, sendo interessante estudá-los também.

Após a confecção da figura , houveram algumas dificuldades na tentativa de criar as próximas figuras. Primeiramente, tentou-se reproduzir a parte correspondente ao efeito sintético da figura 11 do artigo, visto que esta utilizava do mesmo grupo de dados que as figuras anteriormente mostradas. Neste caso, foi feito um boxplot das diferenças entre os valores dos índices espectrais das bibliotecas MILES e SynCoMiL (Δidx):

$$\Delta idx = I_{SPS-M} - I_{SPS-S}$$

em que I é qualquer índice espectral. Feita esta subtração, foi então aplicada uma normalização da forma

$$z = \frac{x - \mu}{\sigma}$$

onde z é definido pela subtração do valor inicial de Δidx (x) pela média do conjunto (μ), sendo este resultado dividido pelo desvio padrão (σ). Os resultados obtidos no artigo estão contidos na figura .

O boxplot gerado pelo Python para reproduzir o artigo está contido na figura . É perceptível, quando comparados o plot do artigo com aquele feito por mim que houve grande diferença entre eles. Nota-se que, enquanto a figura apresentou uma crescente nos resultados, indo de valores médios de -2,5 a aproximadamente 0, a média da figura se manteve centrada neste segundo valor, não havendo muita variação em torno desta região. Além disso, os tamanhos das caixas em cada índice não correspondem aos obtidos por minha orientadora e seus colaboradores, demonstrando que não houve compatibilidade entre eles.

Antes de discutir o por quê esta divergência pode ter ocorrido, vamos fornecer mais um exemplo.

Outra plot que tentou-se refazer, foi o gráfico de densidade presente na figura 7 do artigo (reproduzido nesse texto na figura). Neste caso, foram utlizados os valores de Δidx calculados anteriormente e plotadas suas densidades.

Novamente com o objetivo de estudar e enfatizar o efeito sintético, foram plotados os gráficos apenas referentes a ele, apresentados na figura . Avaliando as figuras e , percebe-se mais uma vez que não houve correspondência entre os resultados. Quando analisado o eixo x, é notável que, em alguns casos, a largura da distribuição não apresentou o tamanho esperado, como para os índices TiO1 e TiO2, onde, apesar das distribuições apresentarem a mesma forma, no artigo foi obtida uma ordem de grandeza de 10^{-2}, enquanto nos meus resultados, esta foi de 10^{1}. Outro fator interessante é o fato de, em alguns casos, o formato da distribuição divergir, como no caso do índice Fe5015, em que os picos do artigo e deste projeto aparentam estar invertidos.

Muitas foram as tentativas para avaliar e identifcar o motivo dos resultados estarem se diferenciando tanto. A princípio, foi levantada a hipótese de eu estar manuseando de maneira errada os arquivos FITS, sendo então estes substituídos pelo formato txt. Entretanto, após a troca, a diferença persistiu.

Após isso, o foco foi mudado não para o formato do arquivo e como utilizá-lo, mas sim para as operações matemáticas que estavam envolvidas, já que, em todos os casos problemáticos, houve algo em comum: a subtração dos valores dos índices (Δidx). Assim, foi inicialmente debatido que pode ter ocorrido uma falta de correspondência nos valores subtraídos, ou seja, Δidx estava sendo calculado realizando a diferença de valores espectrais com idades diferentes. Dessa forma, foi então realizada uma correlação cruzada entre as tabelas MILES e SynCoMiL considerando os fatores log age e metalicidade (Z)

como determinantes. Apesar de resultar em um formato mais parecido com o esperado nas densidades (figura), ainda assim algumas diferenças persistiram.

Foi cogitado, por fim, que a utilização de linguagens de programação distintas pode ter causado erro nos resultados finais. Originalmente, os plots e a análise do artigo foram todos feitos em R, enquanto neste projeto foi usado o Python. Obviamente, isto não deveria afetar os plots, já que ambas são linguagens muito robustas e muito competentes no tratamento de dados. Entretanto, visto que cada uma delas possui bibliotecas diferentes, estas possuem nuances em suas funções embutidas que podem ter feito com que as operações matemáticas fossem feitas de formas distintas, resultando assim nestes resultados incompatíveis.

Infelizmente não foi possível encontrar o motivo principal destes problemas, visto que, como me graduei, não darei seguimento nas atividades da IC, partindo agora para a pós graduação. Entretanto, como continuarei trabalhando com astronomia extragalática, Python e R, utilizarei os conhecimentos e hipóteses levantados ao longo deste projeto em minhas futuras pesquisas.

3 Desempenho Acadêmico

No segundo semestre de 2024, referente à primeira metade da minha iniciação científica, estive matriculado nas disciplinas de 'Introdução à Cosmologia Física' e 'Introdução ao Caos' nas quais fui aprovado com as médias finais 7,5 e 9,1, respectivamente. Além disso, concluí todos os créditos requeridos pelo meu curso, me graduando assim no Bacharelado em Física pelo Instituto de Física da Universidade de São Paulo (IFUSP). Fui aceito no programa de Mestrado em Astronomia do IAG, onde continuarei minhas atividades acadêmicas.

4 Conclusões e Perspectivas

Ao longo deste projeto de IC, foi possível conhecer mais sobre a definição e os usos de modelos de síntese de população estelar, bem como sobre programação, tratamento de dados e análise estatística.

Ao longo do semestre, foram desenvolvidas as atividades de estudo bibliográfico, onde pude compreender mais sobre o funcionamento dos modelos e sobre o modo de operação das linguagens Python e R; manuseamento e limpeza de dados, em que me foi permitido trabalhar diretamente com os modelos de população estelar, compreendendo mais sobre como filtrar e administrar conjunto de dados; e reprodução de resultados prévios da literatura, na qual foram refeitos os gráficos publicados no artigo base deste projeto.

Estes resultados obtidos foram parcialmente compatíveis com o esperado. Para comparações diretas entre os índices espectrais de modelos construídos com as bibliotecas MI-

LES e SynCoMiL, nossos resultados reproduzem os da literatura, revelando semelhança grande entre as bibliotecas. Além disso, foram evidenciados também alguns índices espectrais divergentes que merecem ser reelaborados, como Ca4227, Fe4668, Mg1, Mg2, Fe5406, Fe5709 e B4000Vn.

Já quando foi feito o cálculo de diferenças de índices normalizados, não houve reciprocidade entre os resultados obtidos aqui e os da literatura, de forma que foram cogitados diversos fatores que pudessem ter acarretado essa questão, sendo o mais provável algo relacionado às diferenças de funcionamento entre as bibliotecas do R e do Python.

Devido à conclusão da minha graduação, infelizmente não será possível permanecer investigando o que causou estes erros, porém, seguirei minha vida acadêmica no mestrado em Astronomia no IAG-USP, onde estudarei a morfologia de galáxias disco, por meio do código Capivara. Como utilizarei de análise de dados e da linguagem R novamente, todo o conhecimento astronômico e técnico que aprendi ao longo dessa IC me serão muito úteis nesta nova etapa.

References

References