Machine Learning com dados da colaboração Planck para reconhecimento de padrões

Diogo Ayres Rocha

PCI-CBPF

Engenheiro Eletrônico (CEFET-RJ)

Sumário

Introdução

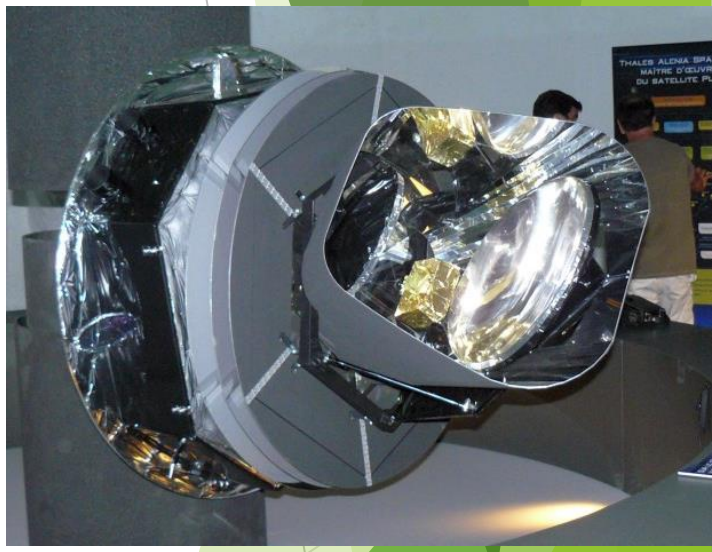
- Cosmic Microwave Background (CMB)
  - Arno Penzias e Robert Wilson
  - Tem origem no início do universo (Big-Bang)
  - O calor gerado remanescente gera esse ruído de fundo devido à expansão e resfriamento do Universo
  - ~2.7 K
  - A radiação é praticamente uniforme preenchendo todo universo
  - O estudo de pequenas variação dessa radiação podem conter informações sobre a origem do universo, assim como a evolução e o conteúdo

Objetivo

- Através de métodos de clusterização achar padrões nos mapas de radiação cósmica de fundo provenientes da colaboração PLANCK
- Entender o uso do formato HEALPIX utilizado pela colaboração para formato dos mapas
- Comparar métodos de clusterização

Sonda Espacial Planck

- A sonda Planck foi lançada em 2009
- Obter dados mais precisos que a sonda Wilkinson Microwave Anisotropy Probe [WMAP]
- Dados adquiridos pela colaboração tiveram uma precisão e resolução maiores que todos os experimentos anteriores
- Para medir, a sonda possuía 2 instrumentos:
  - Low Frequency Instrument (LFI)
  - High Frequency Instrument (HFI)
- Para medir temperaturas tão baixas (~2.7K) com precisão, um dos instrumentos fazia uso do ruído Johnson em uma resistência resfriada, onde devido a agitação térmica causada pela radiação incidente (CMB + foreground emissions), uma tensão aparece nos terminais da resistência atuando como um termômetro



Formato de dados da colaboração

Table Browser

rs Help

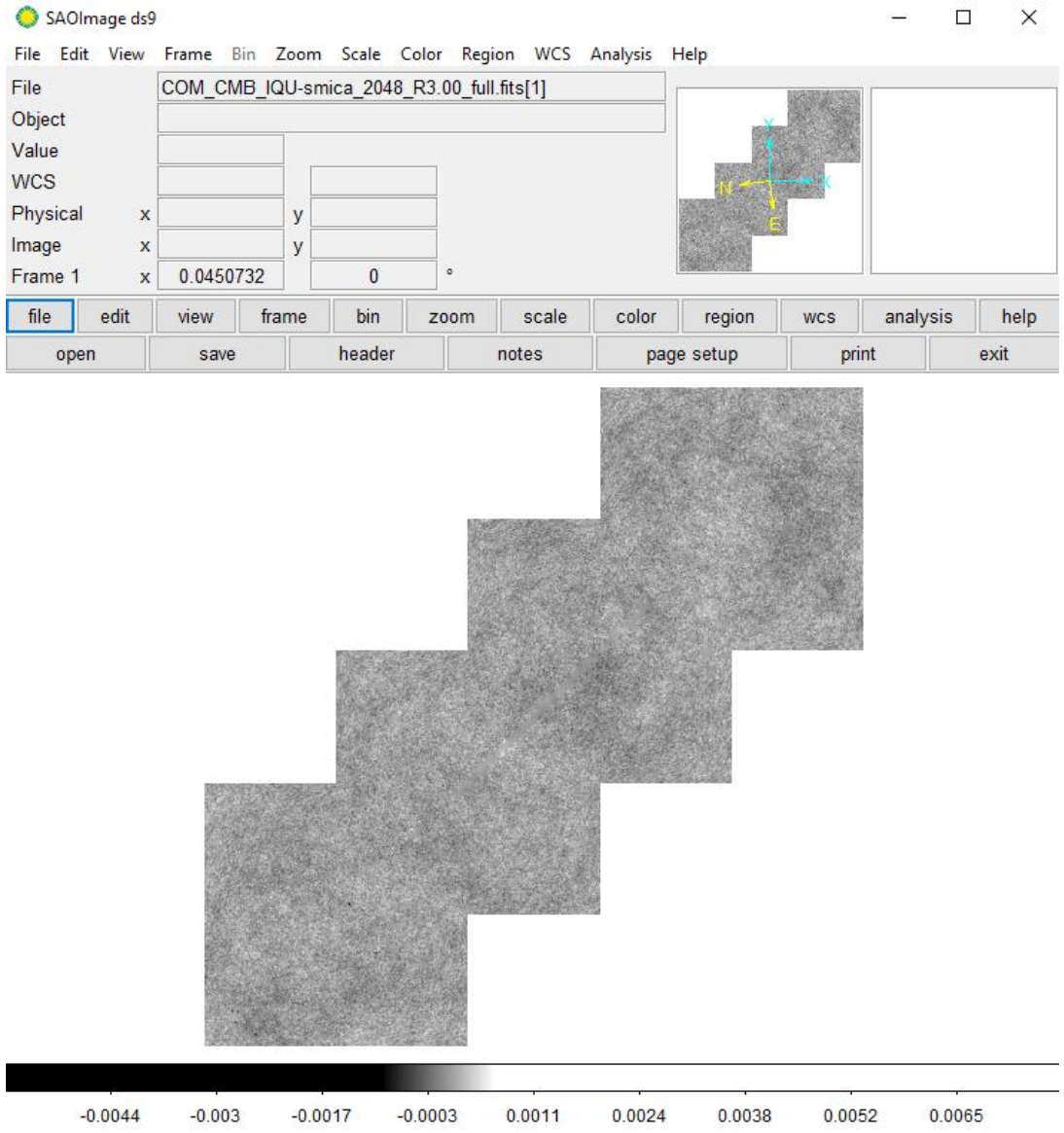
r for 1: COM_CMB_IQU-smica_2048_R3.00_full.fits

I_STOKES	Q_STOKES	U_STOKES	TMASK	PMASK	I_STOKES_INP	Q_STOKES_INP	U_STOKES_INP
1,43179E-5	1,35863E-6	-3,59718E-7	0,	0,	1,57624E-6	2,76046E-6	3,84449E-6
1,47095E-5	6,26007E-7	-1,57705E-6	0,	0,	-1,21124E-5	2,37062E-6	3,91899E-6
1,52535E-5	6,24430E-7	-1,34185E-6	0,	0,	5,41290E-6	2,85037E-6	3,84995E-6
1,56074E-5	-9,84981E-7	8,95875E-7	0,	0,	-7,64925E-6	2,45326E-6	3,94135E-6
1,46945E-5	6,82521E-7	1,18767E-7	0,	0,	-2,82555E-5	1,95950E-6	3,93676E-6
1,68115E-5	1,59002E-7	8,01987E-7	0,	0,	-4,26090E-5	1,53583E-6	3,89961E-6
1,49935E-5	-9,43970E-7	1,31076E-6	0,	0,	-2,39844E-5	2,03515E-6	3,97480E-6
1,73869E-5	-8,49168E-7	3,04878E-7	0,	0,	-3,86226E-5	1,60511E-6	3,95193E-6
1,33982E-5	-1,65636E-6	-8,56168E-7	0,	0,	1,37440E-5	2,88364E-6	3,80290E-6
1,47085E-5	6,48343E-9	2,87981E-6	0,	0,	1,15347E-6	2,48223E-6	3,91188E-6
1,31031E-5	-8,36151E-7	-5,18576E-7	0,	0,	2,28493E-5	2,85992E-6	3,70790E-6
1,43761E-5	1,90890E-6	2,21697E-6	0,	0,	1,14441E-5	2,45706E-6	3,83455E-6
1,50825E-5	1,73112E-6	1,23019E-6	0,	0,	-1,60888E-5	2,06042E-6	3,96249E-6
1,73514E-5	1,15191E-6	-1,42684E-6	0,	0,	-3,21465E-5	1,62751E-6	3,95612E-6
1,54517E-5	2,71743E-6	-3,73142E-7	0,	0,	-6,20495E-6	2,03464E-6	3,90313E-6
1,77065E-5	9,74968E-7	-3,08231E-6	0,	0,	-2,38049E-5	1,60219E-6	3,91483E-6
1,89299E-5	-1,77033E-7	-6,23045E-7	0,	0,	-5,22923E-5	1,10849E-6	3,81026E-6
1,98981E-5	-4,25861E-8	-6,62246E-7	0,	0,	-5,53935E-5	6,86056E-7	3,67225E-6
1,95379E-5	9,70214E-8	8,10826E-8	0,	0,	-4,79741E-5	1,17235E-6	3,87533E-6
1,92591E-5	1,70914E-6	-1,70824E-7	0,	0,	-4,99195E-5	7,45840E-7	3,74853E-6
2,12849E-5	7,15545E-8	9,98004E-7	0,	0,	-5,08359E-5	2,76503E-7	3,48992E-6

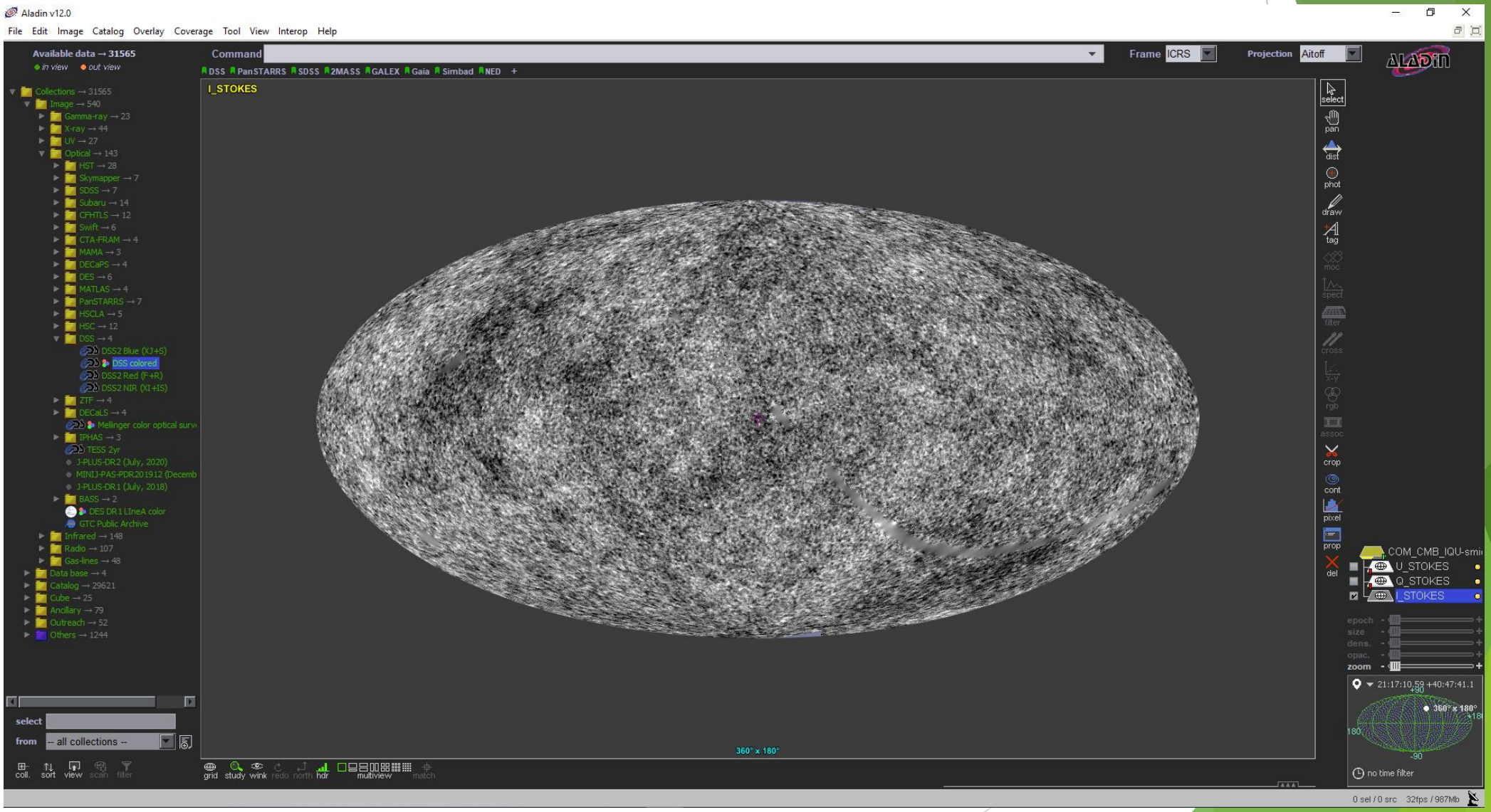
648 Visible: 50.331.648 Selected: 0

- Os dados seguem a padronização HEALPIX
- Basicamente o formato é composto de um vetor 1D em que cada ponto possui um ra e dec associado
- A conversão depende da quantidade de pixels que o dado possui
- Big Endian!

Formato de dados da colaboração



Formato de dados da colaboração

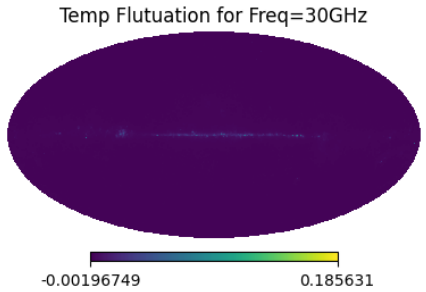


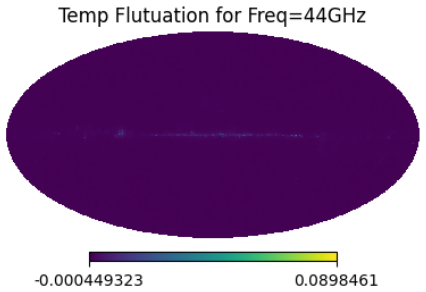
Python

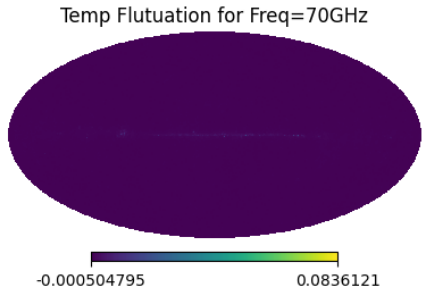
- Bibliotecas usadas
  - Healpy
    - Implementação do healpix no python
  - Numpy
  - Pandas
  - Matplotlib
  - Scikit-learn

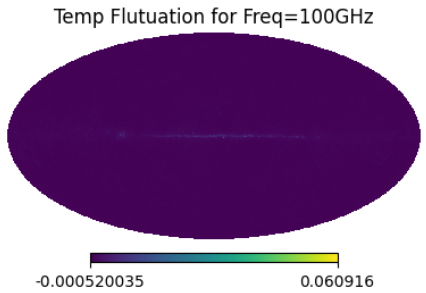
Primeiros plots com healpy

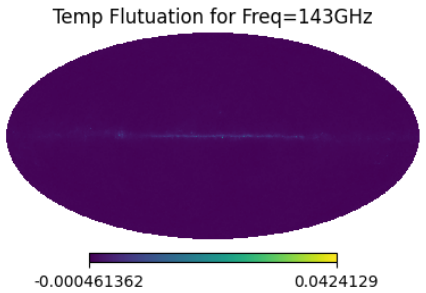
- Mapas de frequência

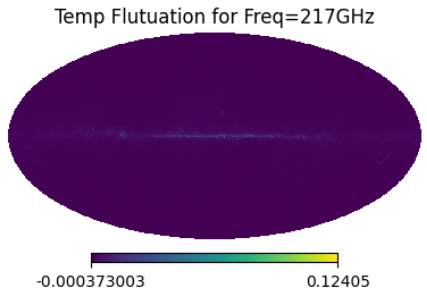


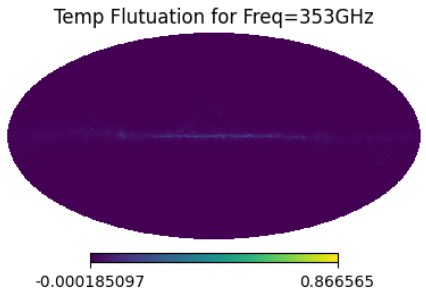


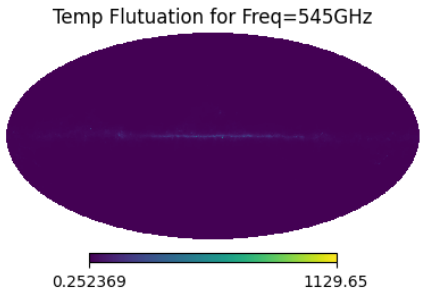


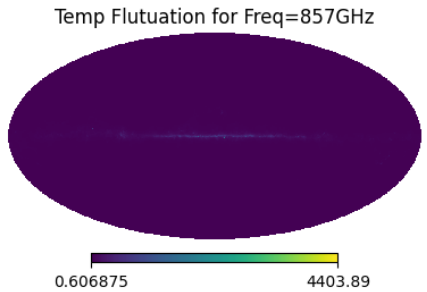






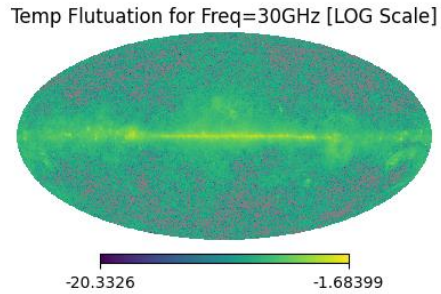


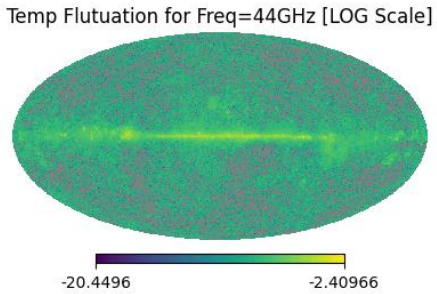


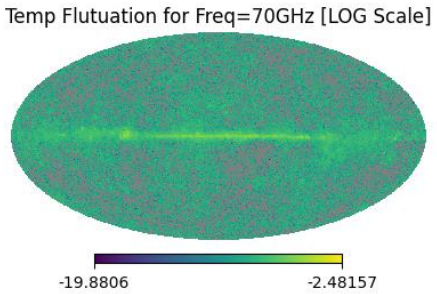


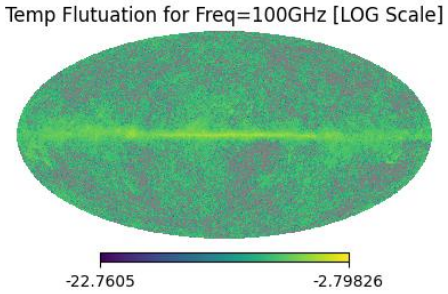
Primeiros plots com healpy

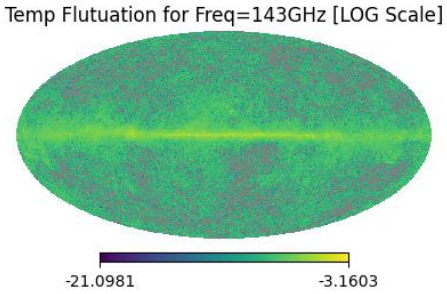
- Mapas de frequência – Escala Logarítmica

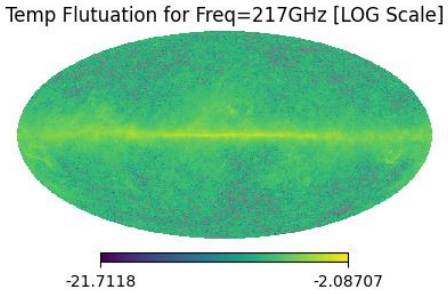


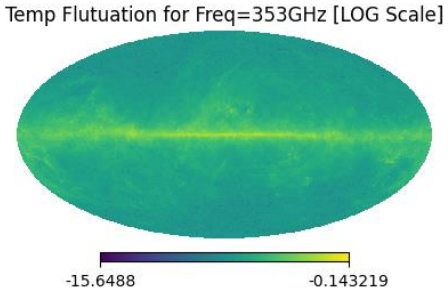


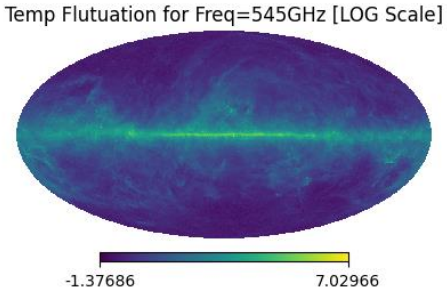


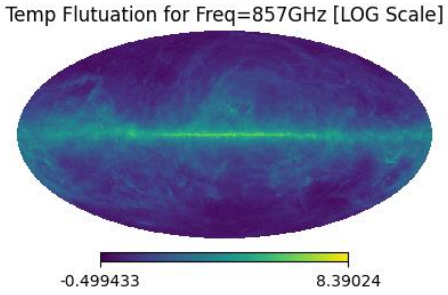






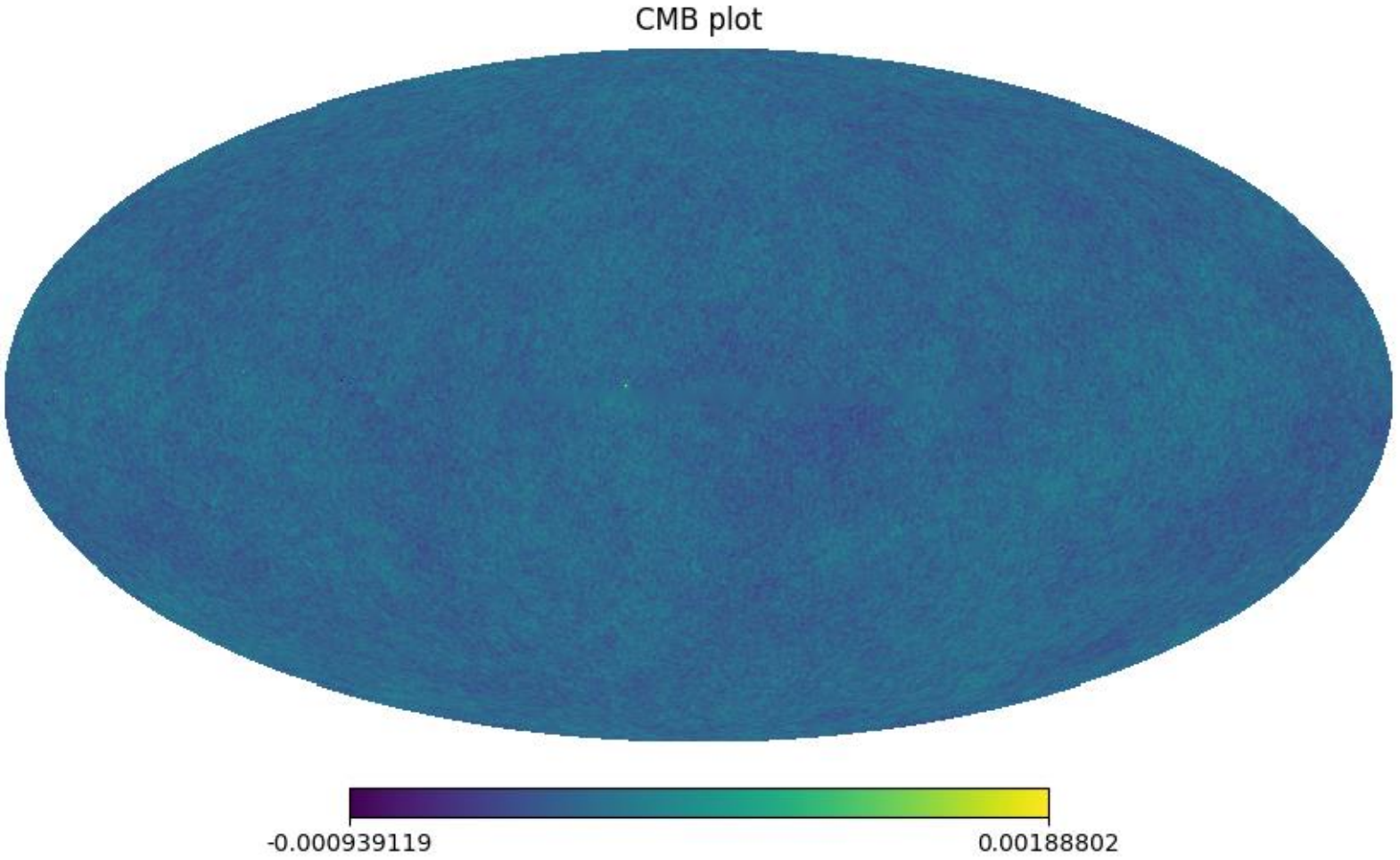






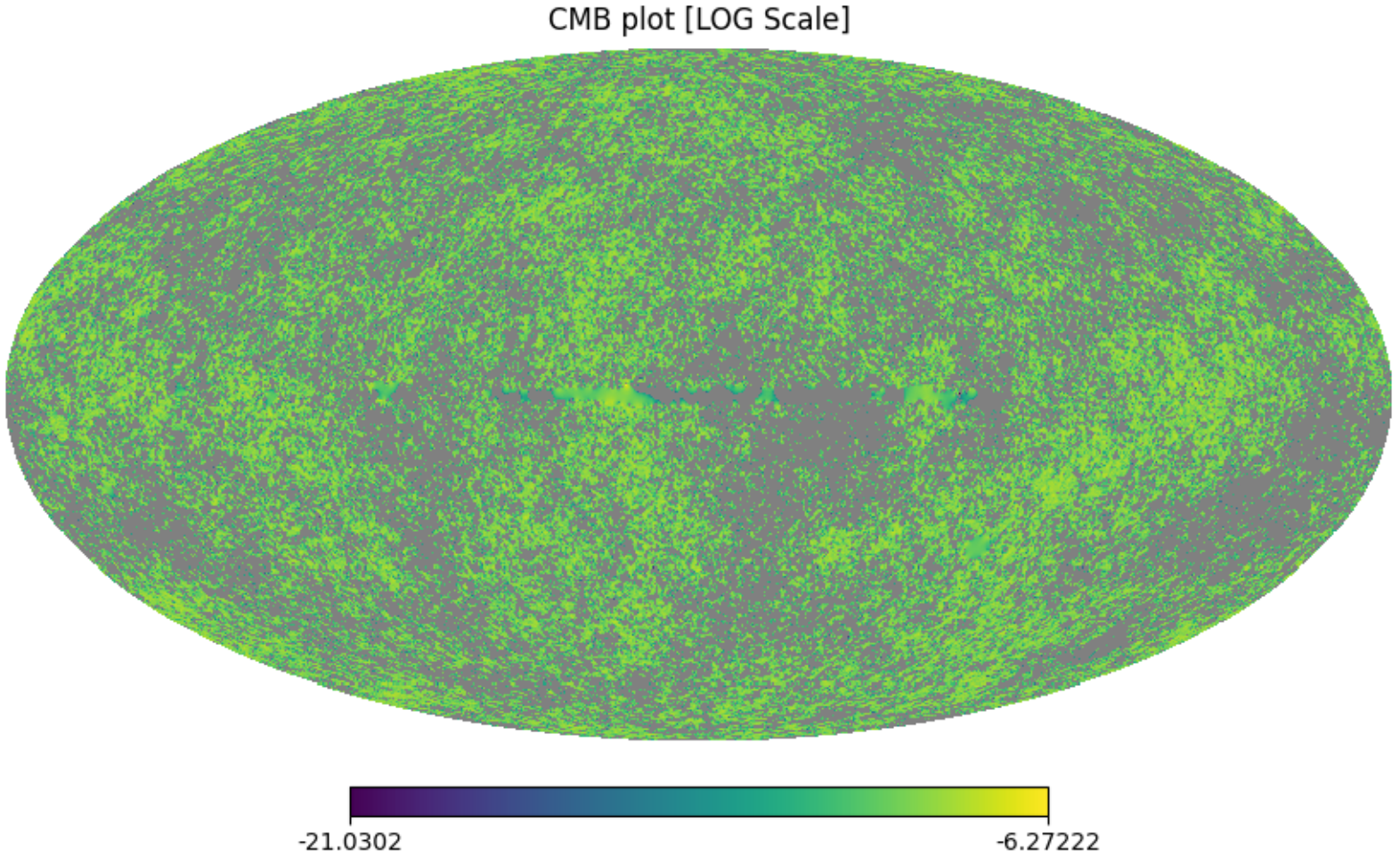
Primeiros plots com healpy

- CMB Plot



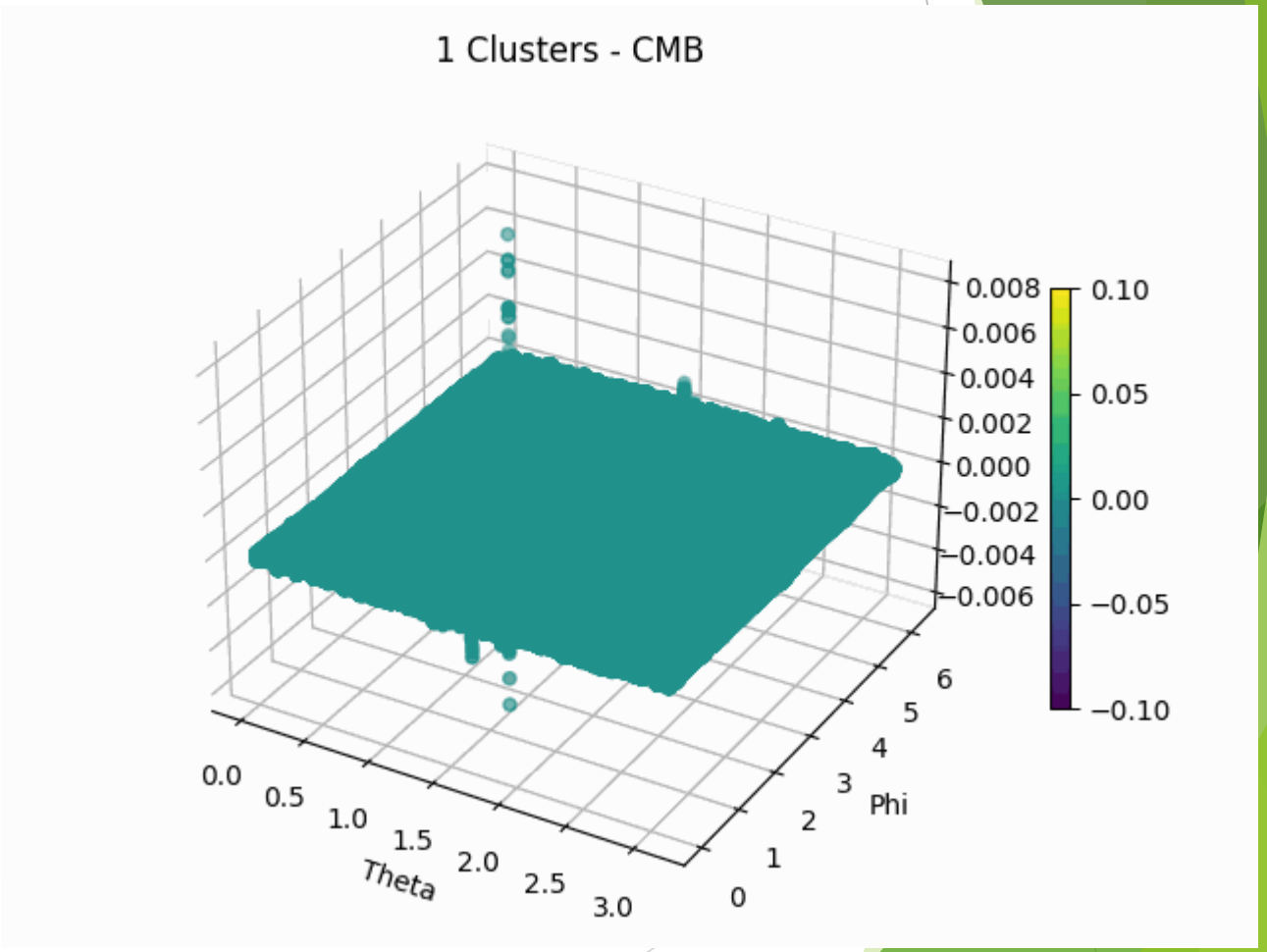
Primeiros plots com healpy

- CMB Plot



K-Means


ClusterSelection N=1 CMB
-1
1

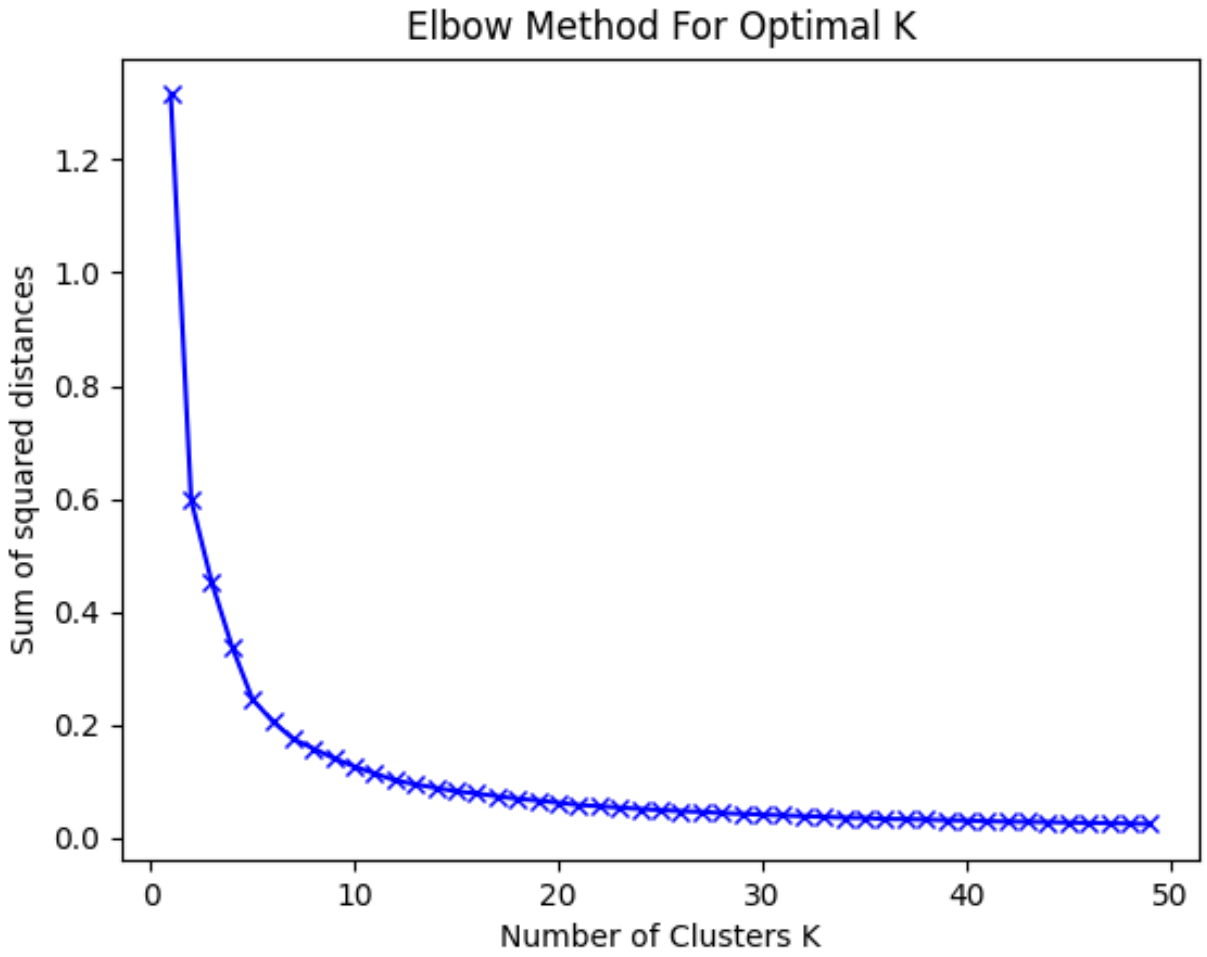

1 Clusters - CMB
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Theta
0
1
2
3
4
5
6
Phi
0.008
0.006
0.004
0.002
0.000
−0.002
−0.004
−0.006
0.10
0.05
0.00
−0.05
−0.10

K-Means

- Como escolher o número de clusters?
- O que seria o ideal?
- Método Elbow (“Joelho”)
  - É realizado uma análise no gráfico de variância dos dados em relação ao número de clusters
  - Na região do “cotovelo” (curva) não existe mais ganho significativo no aumento do número de clusters
  - Esse região é escolhida como número de clusters ideal para ser usada
- Dividir o mapa em regiões menores e procurar por padrões
  - Healpy trabalha como padrão com sistema de orientação esférica
  - `phi = ra`
  - `π/2 - dec = theta`

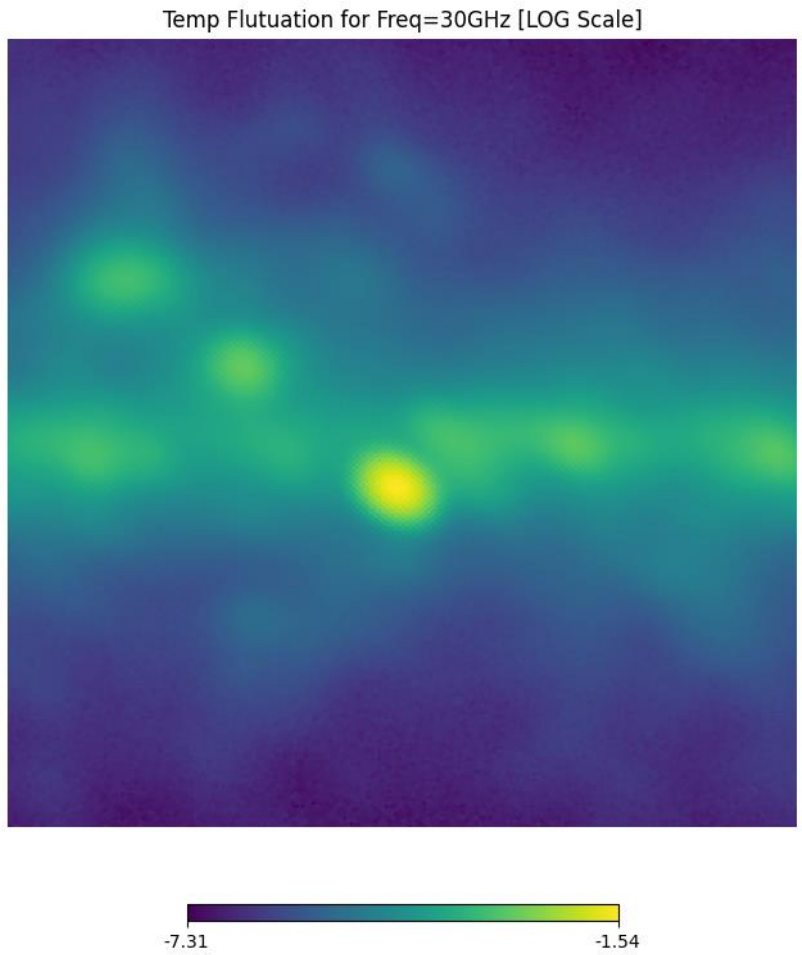
K-Means

K-Means

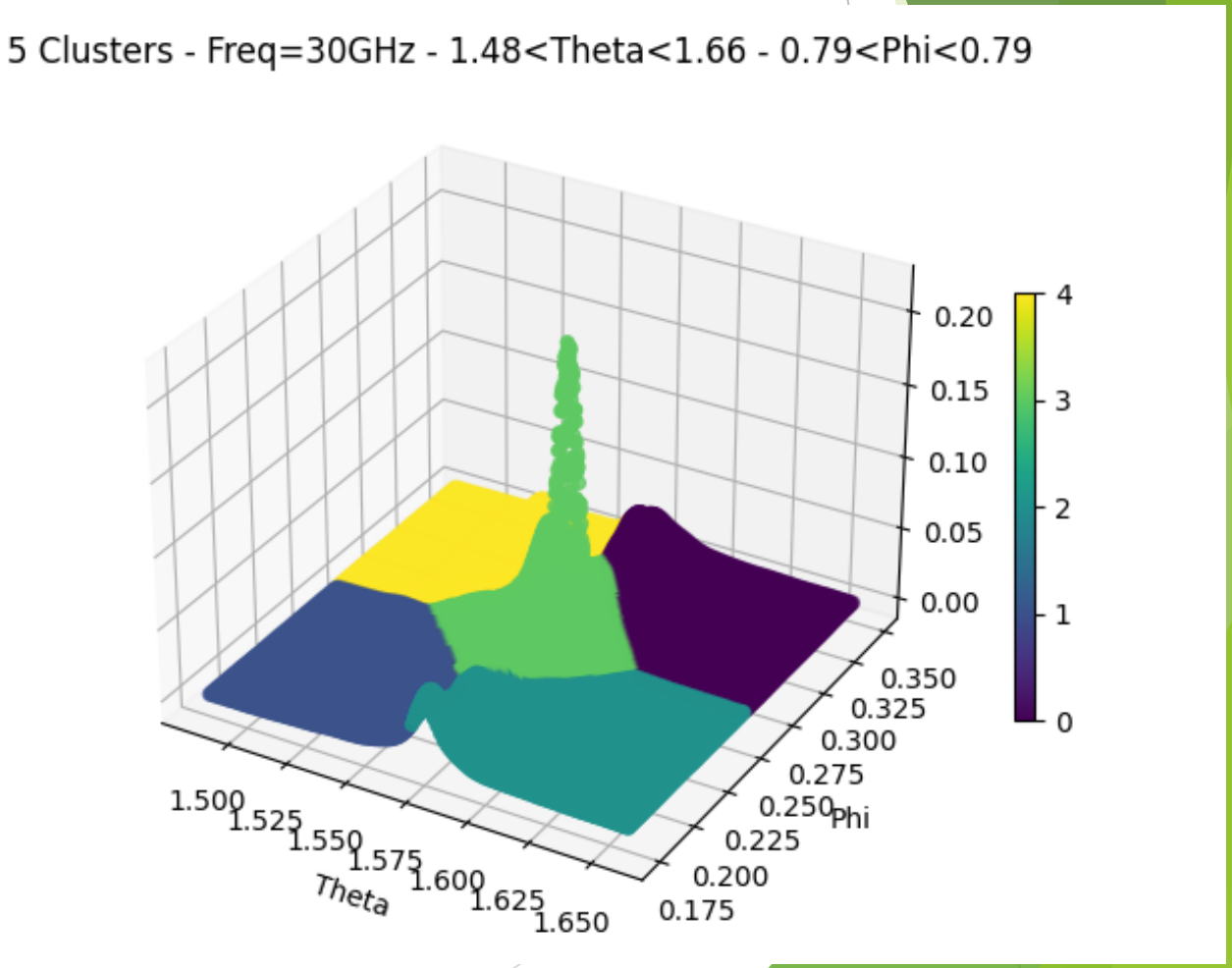


- Para a região ao lado a região do “cotovelo” da curva se situa entre 5 e 8 clusters
- A partir dessa análise é escolhido o valor ideal de clusters para realizar a classificação

Selecionando uma região específica



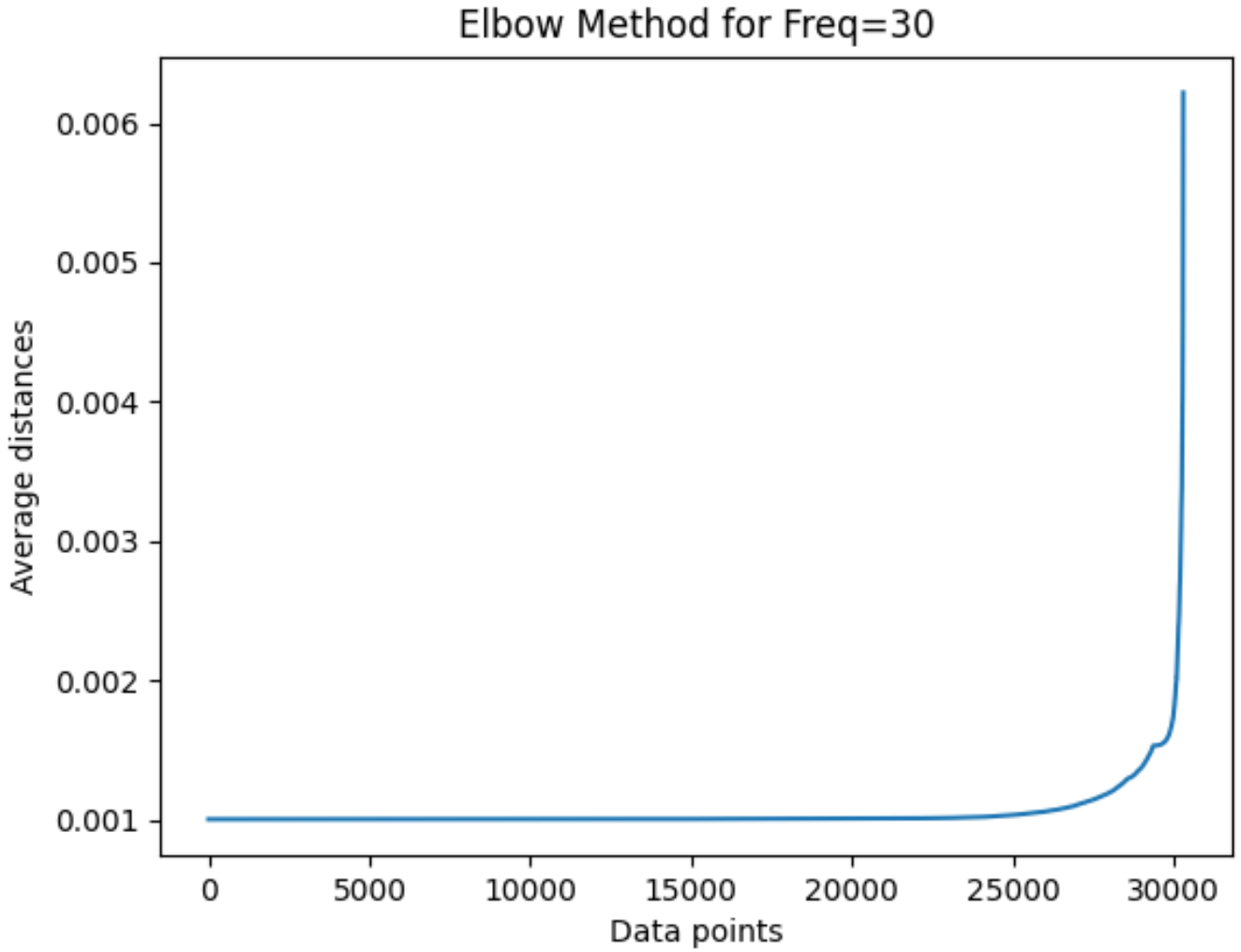
Selecionando uma região específica



DBScan

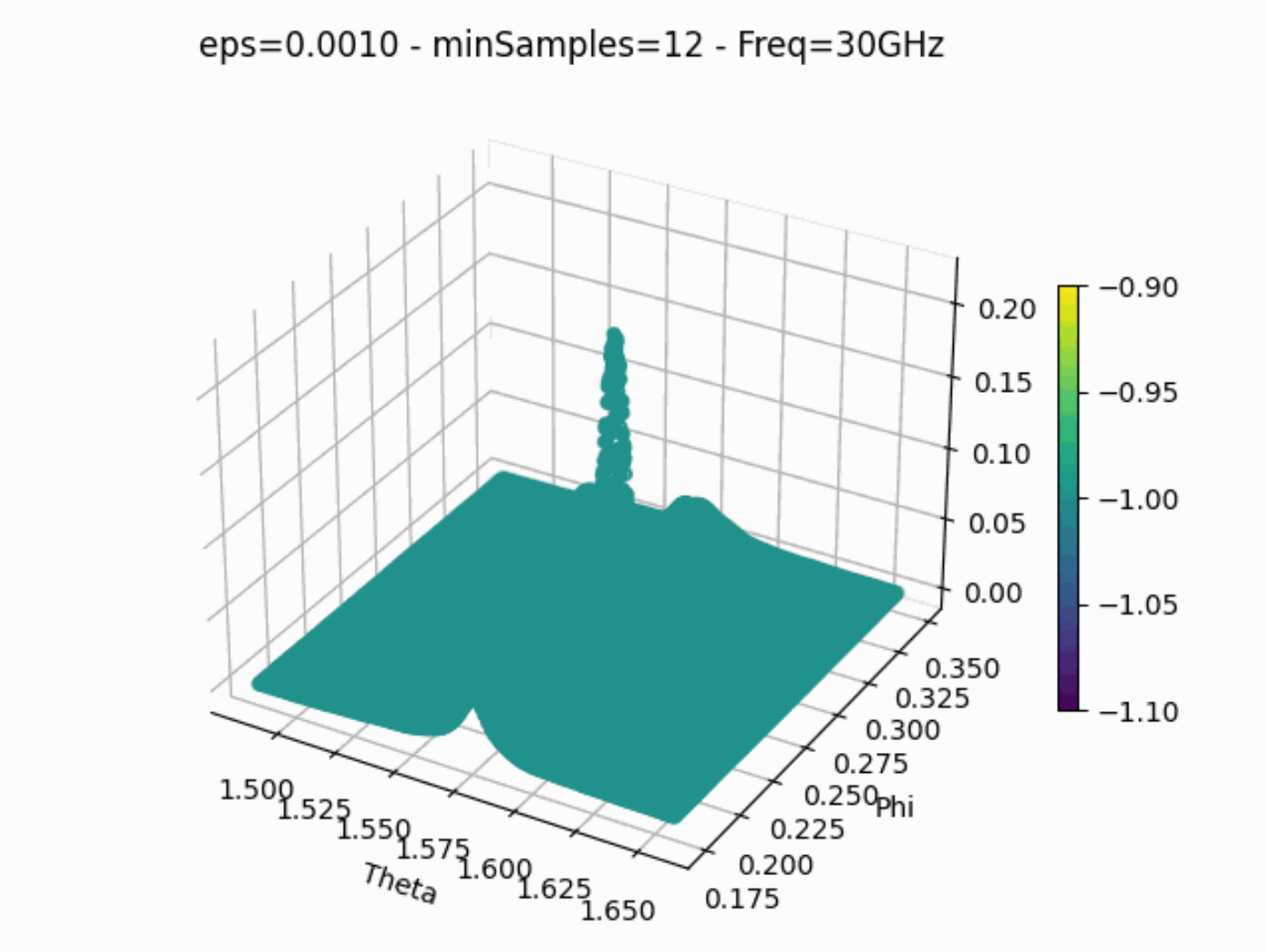
- Diferentemente do método K-Means, o DBScan funciona classificando por densidade
- No método não há controle direto sobre o número de clusters, mas sim a definição de como o cluster será formado
  - Épsilon: É o raio em torno de cada ponto que é definido como espaço para formação do cluster
  - Número mínimo de elementos para um conjunto encontrado ser considerado um cluster
  - Os elementos que ficam fora dessas condições são considerados outliers
- Assim como no K-Means, o DBScan possui um método para guiar a escolha de parâmetros ótimos
  - Para guiar a escolha dos parâmetros ótimos
    - Nmin = (no mínimo) 2*dimensão dos dados
    - Épsilon pode ser obtido através do método elbow

DBScan

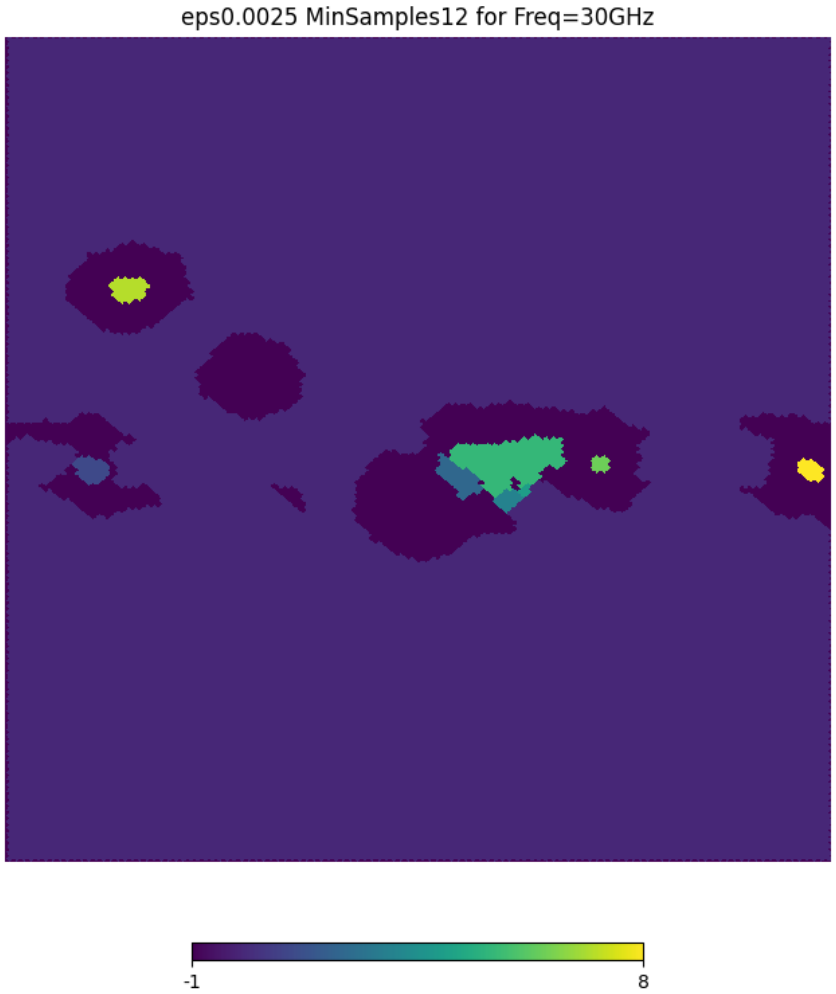


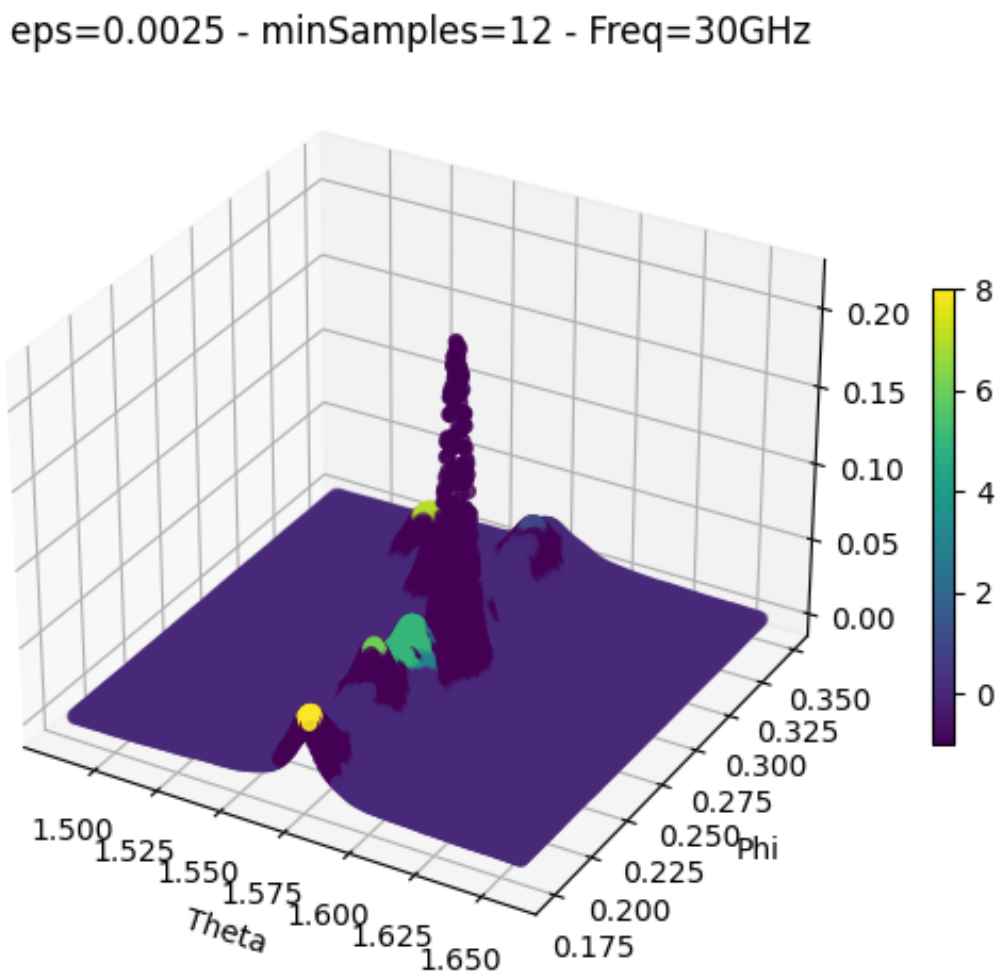
- Assim como usado no K-Means para seleção do valor do número de clusters, no DBScan, a região de inclinação da curva indica um valor ótimo para o raio (élpsilon)
- Nesse caso algo em torno de 0.002

Análise da mesma região selecionada

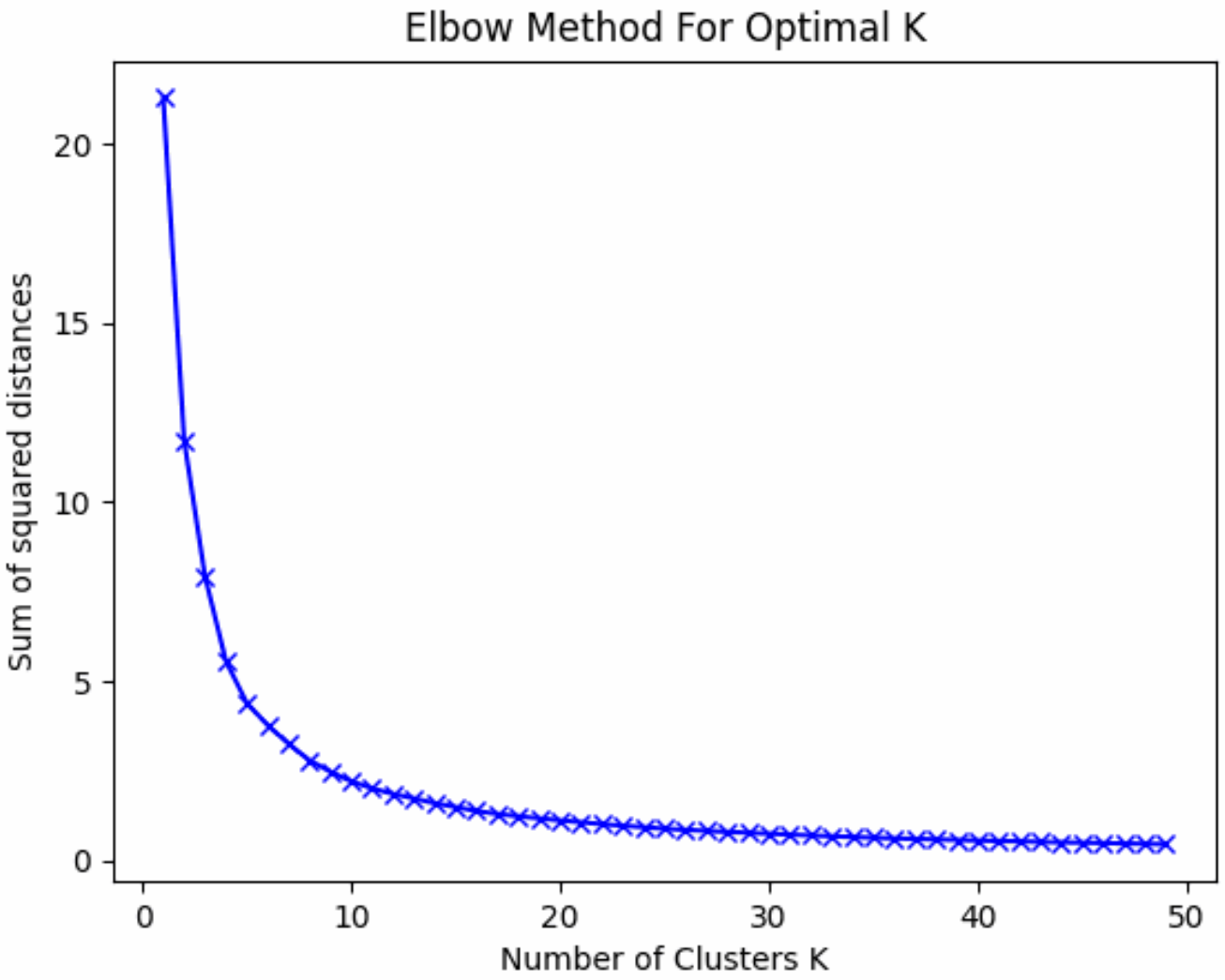


Análise da mesma região selecionada





Varrendo o Mapa



Conclusão

- O trabalho foi uma prova de conceito na análise de padrões utilizando dados de CMB da colaboração Planck
- O método do K-Means se mostrou eficiente em regiões menores para separação de dados com intensidades diferentes
- O método DBScan em função da busca por densidade nos dados consegue colocar dados que estão em um mesmo plano em um único conjunto e separando os picos nos dados, mas perdendo regiões onde a densidade é muito menor
  - Observar que o pico maior onde a densidade é claramente menor, os dados são classificados como outliers

Referências

- [1] https://ui.adsabs.harvard.edu/link_gateway/1965ApJ.[3]..142..419P/doi:10.1086/148307
- [2] https://doi.org/10.48550/arXiv.1506.01907
- [3] http://dx.doi.org/10.1051/0004-6361/201525820
- [4] https://doi.org/10.2307/2346830
- [5] https://doi.org/10.1023/A:1009745219419