João Valério de Souza Neto

Análise topológica em imagens 3D de otólitos de peixes: explorando padrões de densidade e de morfologia

Recife - PE

Fevereiro de 2024



Ministério da Educação
Universidade Federal Rural de Pernambuco
Departamento de Estatística e Informática
PPG em Biometria e Estatística Aplicada

Análise topológica em imagens 3D de otólitos de peixes: explorando padrões de densidade e de morfologia

Tese considerada adequada para obtenção do grau de Doutor em Biometria e Estatística Aplicada, defendida e aprovada unanimamente em 23 de fevereiro de 2024.

Área de concentração:
Biometria e Estatística Aplicada

Linha de pesquisa:
Modelagem e Métodos Computacionais

Orientador: Paulo José Duarte Neto
Coorientador: Wilson Rosa de Oliveira Jr

Recife - PE

Fevereiro de 2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

N469a Neto, João Valério de Souza
Análise topológica em imagens 3D de otólitos de peixes: explorando padrões de densidade e de morfologia / João Valério de Souza Neto. - 2024.
114 f. : il.

Orientador: Paulo Jose Duarte Neto.
Coorientador: Wilson Rosa de Oliveira Jr.
Inclui referências e apêndice(s).

Tese (Doutorado) - Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Biometria e Estatística Aplicada, Recife, 2024.

1. Tomografia 3D. 2. Big data. 3. Custo computacional. 4. Análise topológica de dados. 5. Estoques pesqueiros. I. Neto, Paulo Jose Duarte, orient. II. Jr, Wilson Rosa de Oliveira, coorient. III. Título

CDD 519.5

Ministério da Educação
Universidade Federal Rural de Pernambuco
Departamento de Estatística e Informática
PPG em Biometria e Estatística Aplicada

Tese de doutorado

Análise topológica em imagens 3D de otólitos de peixes: explorando padrões de densidade e de morfologia

João Valério de Souza Neto

Tese considerada adequada para obtenção do grau de Doutor em Biometria e Estatística Aplicada, defendida e aprovada unanimamente em 23 de fevereiro de 2024.

Orientador:

Paulo José Duarte Neto
UFRPE/DEInfo/PPGBEA

Banca examinadora:

Antonio Celso Dantas Antonino
Universidade Federal de Pernambuco
Externo à UFRPE

Francisco Marcante Santana da Silva
Universidade Federal Rural de Pernambuco/UAST
Externo ao PPG

Antonio Samuel Alves da Silva
Universidade Federal Rural de Pernambuco
Interno

Viviane Moraes de Oliveira
Universidade Federal Rural de Pernambuco
Interno

Aos alunos que tive
e
aos que terei.

Agradecimentos

Toda honra, toda glória e todo agradecimento apenas a Deus, por

- me oportunizar a escolha deste doutorado, estar chegando ao final, cada lauda deste documento, toda benção em cada etapa e por ter sido em seu tempo e não no meu;
- iluminar meu orientador com a ideia desse trabalho, por me passar a mesma, os dados e o desafio. Cada reunião, discussão, disciplinas e ensinamentos;
- trazer à luz a ideia de reduzir os dados diante da pouca capacidade computacional, algo que se tornou o pilar desta tese;
- cada discussão matemática com meu coorientador, pelo seu tempo e disciplinas;
- meu filho Augusto, nascido no pico da pandemia, impulsionando esperança;
- meu filho Aurélio, pitada a mais de desafio dando mais sabor a conquista;
- colocar meu amigo Helma no caminho, sua motivação no pré-projeto e apoio em Recife;
- colocar meu amigo Bruno Roberto no caminho, me esperançando ao não desespero na parte computacional madrugadas a dentro no socorro às dúvidas de programação;
- meus colegas de curso. Pela ajuda em Recife e pelos dias e noites de estudos *online*;
- o amigo Haroldo, técnico adm da Rural, por seu ap, compreensão e companheirismo;
- minha esposa Antonia, por sempre aliviar meu estresse;
- estar alcançando esse grau com minha mãe e meu pai em vida;
- meus amigos Ceiça e Evaldo, por me abrirem as portas do Núcleo de Tecnologia do município;
- meus familiares, amigos, ações, momentos, circunstâncias e quaisquer outras coisas e/ou pessoas que não me recordei de agradecer;
- principalmente seu amor não merecido em minha vida.

"Motivos são questionáveis,
matemática é impecável."
Kowalski

Resumo

Otólitos são estruturas calcificas encontradas no ouvido interno dos peixes teleósteos, fundamentais na biologia marinha em estudos de metabolismo, idade, crescimento e na identificação de estoques pesqueiros que podem resultar em práticas sustentáveis de manejo. Uma propriedade relevante dessa estrutura é a sua densidade, porque se refere às modificações na forma cristalina do carbonato de cálcio durante a vida do peixe refletindo em variações de sua forma final. Utilizando micro-tomografia computadorizada, a radiodensidade interna e externa de otólitos de espécies diferentes foram obtidas em uma perspectiva 3D. Porém, ainda não há uma metodologia apropriada que permita descrever e realizar estudos comparativos. Desta forma, buscou-se revelar variações de densidade de otólitos, a partir de imagens 3D de tomografia computadorizada, aplicando a técnica Ball Mapper (BM) da Análise Topológica de Dados (TDA). Inicialmente, preocupou-se em reduzir o custo computacional desta análise aplicando amostragem probabilística e verificando seus efeitos sobre as variações de densidade fornecida pelo grafo BM. Para decidir sobre o tamanho de amostra a própria topologia foi usada para estabelecer o que chamou-se de Validação Topológica da Amostra, que forneceu a resolução mínima com as mesmas informações de densidade que dos dados brutos. A representatividade das amostras foi verificada com testes estatísticos não paramétricos sobre a variável densidade. Com base nas características estruturais da rede, invariantes topológicos permitiram avaliar a similaridade entre grafos. A técnica BM além de revelar padrões de variações de densidade na estrutura do otólito também se mostrou válida como um algoritmo de pré-processamento de imagens tomográficas, permitindo a segmentação de características indesejáveis no objeto de interesse. Em adicional, outra técnica da TDA, Homologia Persistente (HP), foi aplicada aos dados das imagens 3D na expectativa de expor um novo classificador para a forma do otólito. A HP mostrou proeminência mesmo em uma amostra pequena ao separar as classes dos otólitos adequadamente e revelar resultados quantitativos acurados de separação, demonstrando um potencial uso para classificação de otólitos com base em sua estrutura 3D. Por fim, uma análise de regressão demonstrou a possibilidade de estimar idade, comprimento e a radiodensidade dos otólitos pelas características topológicas resultantes da classificação. Com base na análise topológica de imagens 3D de otólitos, foi possível revelar padrões de densidade, segmentar características indesejáveis, classificar otólitos pela estrutura 3D e estimar idade, comprimento e radiodensidade. Tais resultados podem contribuir para estudos de metabolismo, idade, crescimento e manejo sustentável de estoques pesqueiros.

Palavras-chave: Tomografia 3D; *Big data*; Custo computacional; Análise topológica de dados; Estoques pesqueiros.

Abstract

In this thesis, we present a comparative study of otolith density variations using Topological Data Analysis (TDA). Otoliths are calcium carbonate structures found in the inner ears of fish and are commonly used to study age and growth patterns in fish populations. Traditionally, the analysis of otolith density variations has been a computationally intensive task due to the high-dimensional nature of the data. However, TDA offers a promising approach to reduce the data dimensionality and extract meaningful topological information from otolith images. We applied the Ball Mapper algorithm to a dataset of 3D otolith images from different fish species and ages. The algorithm allowed us to construct topological graphs representing the density variations in otoliths. We also explored the use of probabilistic sampling techniques to reduce the data and found that a sample size of 5% provided accurate representations of otolith density variations compared to the full dataset, after a Sample Topological Validation procedure developed here to ensure the efficiency and reliability of the sampling process. Topological invariants of the graphs, such as average clustering, node connectivity, assortativity, shortest path length, efficiency, and others, were used to comparizon between graphs. The comparizon of the topological properties of the full dataset with those of the 5% sample found a high degree of similarity, indicating that TDA with a reduced dataset can capture essential density information. Ball Mapper further allowed us to identify and eliminate dirt or anomalies present in otolith images, further enhancing the accuracy of our analysis. Overall, our study demonstrates the efficacy of TDA in studying otolith density variations with significant computational gains over traditional methods. The reduced data size using probabilistic sampling and the robustness of topological invariants provide valuable insights into the density patterns of otoliths. Another TDA technique, Persistent Homology (PH), was applied to the 3D image data with the expectation of unveiling a new classifier for otolith shape. PH demonstrated prominence even in a small sample by effectively separating otolith classes and revealing accurate quantitative separation results, showcasing potential use for otolith classification based on their 3D structure. Finally, a regression analysis demonstrated the possibility of estimating age, length, and radiodensity of otoliths based on the topological features resulting from the classification.

Keywords: 3D tomography; Big data; Computational cost; Topological Data Analysis; Fish stocks.

Lista de figuras

Figura 1 – Ilustração do ouvido interno nos peixes teleósteos 1
Figura 2 – Termos básicos da estrutura morfológica do otólito 2
Figura 3 – Sistema labiríntico em um *Diapterus brevirostris* 7
Figura 4 – Exemplo da variedade de otólitos de diferentes espécies 7
Figura 5 – Exemplo de análise microestrutural e microquímica em otólitos 8
Figura 6 – Exemplo de contagem de anéis em um otólito 9
Figura 7 – Exemplo de formatos de otólito 9
Figura 8 – Ilustração de um ciclo sustentável dos recursos pesqueiros 10
Figura 9 – Observações de variações de densidade em otólitos 11
Figura 10 – Imagem digital de um crânio de peixe identificando otólitos 12
Figura 11 – Exemplo de construção de um complexo simplicial 20
Figura 12 – Construção dos complexos simpliciais de Čech e Rips 21
Figura 13 – Analogia aos números de Betti do toro 23
Figura 14 – Ilustração do cálculo da homologia persistente 25
Figura 15 – Exemplo de homologia persistente sobre o toro 26
Figura 16 – Grafo de Reeb para um toro 27
Figura 17 – Ilustração do Algoritmo Mapper 29
Figura 18 – Grafo da atividade cerebral humana 33
Figura 19 – Imagens 3D originais de otólitos da amostra de estudo 37
Figura 20 – Ilustração da técnica BM 38
Figura 21 – Diagrama com a sistemática da metodologia 46
Figura 22 – Diagrama com a sistemática das análises 48
Figura 23 – Matrizes de dispersão do AS1, comparação entre Pop e amostras a 5% 53
Figura 24 – VTA para o otólito AS1 57
Figura 25 – Distribuição dos voxels por slice para o otólito TO3 e plot do slice Z=325 58
Figura 26 – VTA para o otólito TO3 59
Figura 27 – Plots do slice Z=325 do TO3 com resolução em 1, 5 e 10%. 60
Figura 28 – Grafos BMs do AS1. Comparando densidade entre Pop e Samps a 5% . 62
Figura 29 – Grafos BMs do TO3. Comparando densidade entre Pop e Samps a 5% 63
Figura 30 – μCT do otólito AS3 com um *chunk* detectado pelo grafo BM 65
Figura 31 – Grafos BM do otólito AC8 66
Figura 32 – Grafos BM do otólito AS2 com e sem sujeira 67

Figura 33 – Boxplots das entropias de persistência 73
Figura 34 – Matrizes de característica da classificação qualitativa dos otólitos . . . 74
Figura 35 – Matriz de correlação: entropias vs variáveis do peixe - *Thunnus obesus* 82
Figura 36 – Análise gráfica da regressão em dados dos otólitos da espécie *Thunnus obesus* – caso linear Ey_n vs HU 86
Figura 37 – Análise gráfica da regressão em dados dos otólitos da espécie *Thunnus obesus* – caso exp Ey_n vs HU 87
Figura 38 – Matriz de correlação: entropias vs variáveis do peixe - *Acanthurus coeruleus* 88
Figura 39 – Análise gráfica da regressão em dados dos otólitos da espécie *Acanthurus coeruleus* – caso linear Ex_n vs Age 91
Figura 40 – Análise gráfica da regressão em dados dos otólitos da espécie *Acanthurus coeruleus* – caso exp Ex_n vs Age 92
Figura 41 – Apêndice: Matrizes de dispersão do otólito TO3: comparação entre Pop e amostras a 1% 109
Figura 42 – Apêndice: VTA para o otólito OO1. Strat a 5% 111
Figura 43 – Apêndice: VTA para o otólito TA. Strat a 5% 111
Figura 44 – Apêndice: VTA para o otólito AC42. Strat a 5% 112
Figura 45 – Apêndice: VTA para o otólito HP1. Strat a 5% 112
Figura 46 – Apêndice: Slices de referência (Strat a 5%) do VTA das demais espécies 112
Figura 47 – Apêndice: Topologia dos demais otólitos 113
Figura 48 – Apêndice: Homologia dos otólitos 114

Lista de tabelas

Tabela 1 – Amostra do estudo 36
Tabela 2 – Dados dos peixes e das imagens da amostra do estudo 51
Tabela 3 – Teste de *Kolmogorov-smirnov* sobre distribuições de HU 54
Tabela 4 – Principais estatísticas descritivas de HU 55
Tabela 5 – Tabela de frequência dos voxels por slice Z para o otólito AS1 55
Tabela 6 – ITs para diferentes espécies de otólitos 68
Tabela 7 – Tempo de processamento do cálculo da topologia para alguns otólitos . 69
Tabela 8 – Entropias de persistência dos otólitos 72
Tabela 9 – Resultado da classificação quantitativa dos otólitos 78
Tabela 10 – Regressão: Entropias vs variáveis do peixe - espécie *Thunnus obesus* . . 83
Tabela 11 – Regressão: Entropias vs variáveis do peixe - espécie *Acanthurus coeruleus* 89
Tabela 12 – Apêndice: Tabela de frequência dos voxels por slice Z – otólito TO3 . . 110

Lista de abreviaturas e siglas

μCT	Microtomografia computadorizada de raios X
Big data	Grande volume de dados
HU	Densidade óssea dos otólitos em Unidades Hounsfield
TDA	Análise Topológica de Dados, herdada do inglês *Topological Data Analisys*
BM	Algoritmo Ball Mapper ou Grafo Ball Mapper
HP	Homologia Persistente
ONU	Organização das Nações Unidas
NOAA	NOAA Fisheries é uma agência dos Estados Unidos que faz parte da *National Oceanic and Atmospheric Administration* (NOAA)
Pop	População
Samp(s)	Amostra(s)
Simp	Amostra ou Amostragem Aleatória Simples
Syst	Amostra ou Amostragem Aleatória Sistemática
Strat	Amostra ou Amostragem Aleatória Estratificada Proporcional
VTA	Validação Topológica da Amostra
Rips	Complexo simplicial Vietoris-Rips
Čech	Complexo simplicial de Čech
score	*OOB score*
AIC	Critério de Informação de Akaike

Lista de símbolos

$\mathbb{R}^n$	Conjunto dos números reais de dimensão n
$\mathbb{E}^n$	Espaço euclidiano n-dimensional
E	Espaço topológico
X_i ou $\mathbb{X}_i$	Nuvem de pontos ou matriz de dados tratada como espaço topológico
$B(x,r); B_r(x)$	Bola de raio r centrada em x
ε	Parâmetro do Algoritmo Ball Mapper
G	Grafo abstrato
$H_k(X_i)$	k-ésimo grupo de homologia do espaço topológico X_i
β_j	Número de Betti para o j-ésimo grupo de homologia
Ex	Entropia calculada sobre o grupo de homologia H_0 (Componentes conectados)
Ey	Entropia calculada sobre o grupo de homologia H_1 (Vazios unidimensionais)
Ez	Entropia calculada sobre o grupo de homologia H_2 (Vazios bidimensionais)
Ex_n	Entropia Ex normalizada
Ey_n	Entropia Ey normalizada
Ez_n	Entropia Ez normalizada
ρ	Coeficiente de correlação de Spearman

Sumário

1 Introdução 1

2 Fundamentos teóricos 6

2.1 Otólitos 6

2.2 Imagens Digitais para estudos de Otólitos 12

2.3 Análise Topológica de Dados 13

2.3.1 Homologia Persistente 23

2.3.2 Mapper 27

2.3.3 Aplicações a Imagens Digitais 32

3 Objetivos 35

3.1 Objetivo Geral 35

3.2 Objetivos Específicos 35

4 Material e Métodos 36

4.1 Amostra e aquisição das imagens 36

4.2 Ferramentas da Análise Topológica de Dados 37

4.2.1 O Algoritmo *Ball Mapper* 37

4.2.2 Invariantes Topológicos sobre grafos 39

4.2.3 Homologia Persistente 40

4.3 Sistemática da Análise 46

5 Resultados e Discussões 50

5.1 Imagens de otólitos reduzidas por amostragem probabilística 52

5.2 Validação topológica das imagens reduzidas 56

5.3 Topologia de otólitos como perspectiva para explorar densidade 61

5.4 *Ball Mapper* como ferramenta de segmentação para otólitos 64

5.5 Invariantes Topológicos como ferramenta para comparação entre grafos 67

5.6 Homologia Persistente como ferramenta para classificação de otólitos 70

5.6.1 Classificação qualitativa 74

5.6.2 Classificação quantitativa 77

5.7 Relacionamento entre características topológicas e variáveis do peixe . 81

6 Conclusão 93

Referências 95

APÊNDICE A Algoritmo usado para a extração dos voxels e valores de HU . 107
APÊNDICE B Matrizes de Dispersão do otólito TO3 109
APÊNDICE C Tabela de frequência dos voxels do otólito TO3 110
APÊNDICE D VTA em otólitos das demais espécies 111
APÊNDICE E Topologia dos demais otólitos 113
APÊNDICE F Homologia persistente dos otólitos 114

1 Introdução

Otólitos são complexos biomineralizados presentes no ouvido interno dos peixes teleósteos essenciais para a função de equilíbrio e audição (LEE et al., 2019). O ouvindo interno dos peixes vertebrados (espécies *osteichthyan*) é composto por, três canais semicirculares, sacos óticos, líquido endolinfático, máculas e os otólitos, organizados do seguinte modo, três sacos óticos chamados *sacculus*, *utriculus* e *lagena*, cada um contendo um otólito imerso em endolinfa, chamados, respectivamente, por *sagitta*, *lapillus* e *asteriscus*, todos conectados por células sensoriais ciliares às maculas que enviam sinais elétricos ao cérebro a cada movimento dos otólitos (PANFILI et al., 2002, p. 31). A Figura 1 ilustra e descreve a organização da estrutura do ouvido interno dos peixes teleósteos.

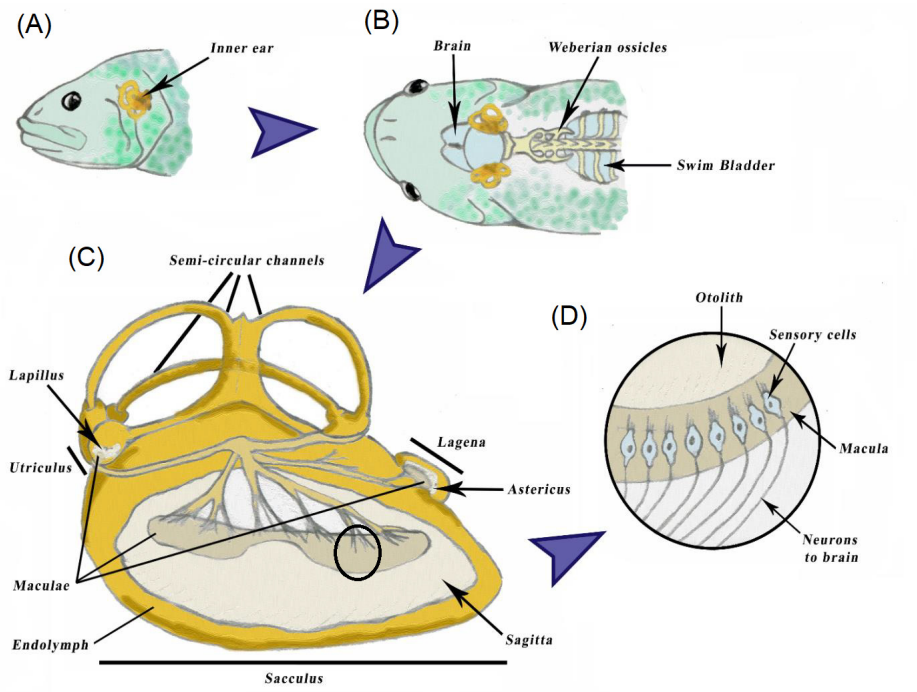


Figura 1 – Ilustração do ouvido interno dos peixes teleósteos. (A) Localização do ouvido interno (inner ear) em relação a cabeça. (B) Destaque dos dois ouvidos internos (direito e esquerdo) em relação ao cérebro (brain), a estrutura óssea (weberian ossicles) e a bexiga natatória (swin bladder) do peixe. (C) Ampliação de um dos ouvidos destacando os canais semicirculares (*semi-circular channels*), os sacos óticos *utriculus*, *lagena* e *sacculus*, as máculas (*maculae*), a endolinfa (*endolymph*) e os otólitos *lapillus*, *asteriscus* e *sagitta*. (D) Destaque da conexão das células sensoriais (sensory cells) ciliares à mácula e ao otólito. Adaptado de Ashworth (2016).

Em termos de funcionamento, quando os peixes teleósteos se movimentam, eles realizam movimentos da cabeça ou corpo, os otólitos de movem na endolinfa estimulando as células ciliadas das máculas. Essas células convertem o movimento dos otólitos em sinais elétricos, que são transmitidos ao cérebro, que interpreta esses sinais para determinar posição vertical e orientação gravitacional contribuindo para o equilíbrio e coordenação motora (POPPER; HOXTER, 1981).

Quanto ao formato, os otólitos encontrados na maioria dos teleósteos possuem formato elíptico e são simétricos em relação aos lados esquerdo e direito com exceção dos peixes chatos e bagres. Têm um sulco característico conhecido por *sulcus acusticus*, cuja função é fazer o contato das células ciliares da mácula ao otólito. Suas faces, classificadas como interna, a que está o sulco, e externa, são convexa e côncava, respectivamente. Sobre tamanho, o *sagitta* é o maior dentre os três e geralmente o mais utilizado em estudos (PANFILI et al., 2002, p. 34). Os termos básicos que descrevem a estrutura morfológica do otólito são detalhados na Figura 2.

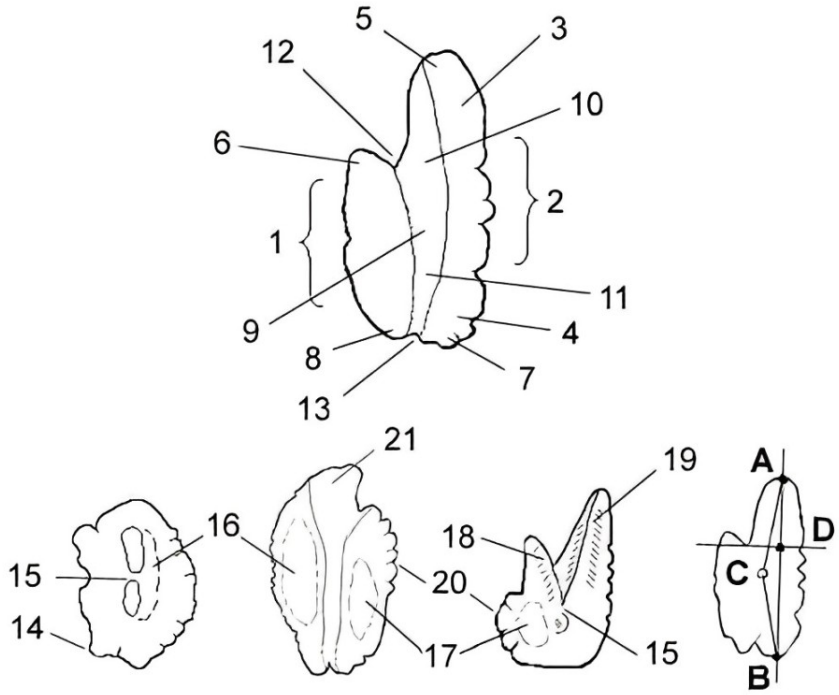


Figura 2 – Terminologia morfológica do otólito dentre os teleósteos. Interior: 1 – asa dorsal, 2 – asa ventral, 3 – parte anterior, 4 – parte posterior, 5 – rostro, 6 – antirostro, 7 – pós-rostro, 8 – pararostro, 9 – sulco, 10 – óstio, 11 – cauda, 12 – excisão maior, 13 – excisão menor, 14 – papila, 15 – arco do sulco (antepara), 16 – área ventral, 17 – área dorsal, 18 – cristas dorsais, 19 – cristas ventrais, 20 – margem com lóbulos (margem lobada), 21 – gancho (rostro tem pequena curvatura). Exterior: AB – comprimento do otólito, AD – comprimento do rostro (Rst), AC – raio rostral (R), CB – raio póstrostral (Z 1), C – núcleo. Fonte: Svetocheva, Stasenkova e Fuks (2007).

Em composição, são estruturas predominantemente compostas por aragonita, que é um polimorfo de carbonato de cálcio que surge como cristais biomineralizados dentro de um líquido de endolinfa, um fluido presente no ouvido interno dos peixes, morfologia, tamanho e organização cristalina desse mineral são reguladas por proteínas especializadas durante seu processo de crescimento (PAYAN et al., 1999). Esse processo de crescimento resulta em otólitos com características distintas que são usadas em estudos de idade, crescimento e análises ambientais que servem para compreender a dinâmica das populações e a saúde dos ecossistemas aquáticos (POPPER; RAMCHARITAR; CAMPANA, 2005).

O crescimento dos otólitos acontece pela deposição de seus elementos base, aragonita e proteínas, em camadas concêntricas e, tal crescimento é definido pelo metabolismo e pelo ambiente que influenciam as variações daqueles elementos base, em padrão diário, sazonal ou anual, de modo constante durante toda a vida do peixe. A primeira região formada no otólito, denominado núcleo, é onde existe a mais baixa densidade da estrutura e é de onde ele continua a crescer (PANFILI et al., 2002). A densidade da estrutura, por sua vez é definida pelas modificações nas formas cristalinas do carbonato de cálcio ($CaCO_3$), o que a torna uma propriedade relevante para o papel funcional dos otólitos (SCHULZ-MIRBACH et al., 2019).

Estudos biométricos de otólitos são importantes para a gestão de estoques pesqueiros (HÜSSY et al., 2016), uma vez que eles podem fornecer conhecimento sobre o metabolismo das espécies (DUARTE-NETO et al., 2014), conectividade entre populações (BRADBURY et al., 2011), identificação, discriminação e avaliação de tais estoques (HE et al., 2020; DUARTE-NETO et al., 2008; ADAMS; ROHLF; SLICE, 2004), resultando em práticas de manejo sustentável (GREEN et al., 2009; SCHULZ-MIRBACH et al., 2019) que se alinham com os Objetivos de Desenvolvimento Sustentável, Consumo e produção responsáveis e Uso sustentável dos recursos marinhos, propostos pela ONU em sua Agenda até 2030 (NJ, 2015; UN, 2015).

Quantificar densidade de otólitos como estrutura completa 3D ainda é um campo pouco investigado, porém recentemente uma nova área chamada eco-densitometria para explorar variações de densidade interna e externa foi sugerida por Vasconcelos-Filho et al. (2019) quando definiu uma série de parametrização ótima para a obtenção de imagens 3D de microtomografia computorizada de raios X (μCT), ao qual permitiu perceber em alta resolução variações de densidade na estrutura completa de otólitos. No entanto, imagens 3D de μCT são um conjunto de dados de alta dimensão ou *Big data* (no caso de imagens 3D, nuvem de pontos ou voxels), e os métodos que trabalham com grandes massas de dados requerem o desenvolvimento técnicas, ou a aplicação das já existente, que consigam processar os dados e esboçar informações em dimensões mais baixas.

Estudos e investigações da natureza de grandes massas de dados estão cada vez mais recorrentes devido à grande coleta e armazenamento de dados por várias áreas da ciência nos dias atuais (SINGH; MEMOLI; CARLSSON, 2007; MANOGARAN; LOPEZ; CHILAMKURTI, 2018). Uma proeminente área que trabalha com *Big data*, como nuvens de pontos, é a Análise Topológica de Dados, ou TDA da sigla em inglês, porque possui um rico conjunto de ferramentas que visam fornecer algoritmos matemáticos e estatísticos bem fundamentados para explorar, analisar e inferir estruturas topológicas e geométricas complexas, a fim de encontrar padrões ou formas nos dados, permitindo resultados a baixas dimensões (CHAZAL; MICHEL, 2021; WASSERMAN, 2018).

Uma das ferramentas da TDA amplamente utilizada em várias aplicações científicas, incluindo a disseminação do coronavírus (CHEN; VOLIĆ, 2021), dinâmica organização do cérebro humano (SAGGAR et al., 2018) e obtenção de inferências sobre dados complexos como a identificação de subconjuntos de pacientes com câncer de mama (LUM et al., 2013), é a técnica Mapper ou algoritmo Mapper (SINGH; MEMOLI; CARLSSON, 2007). Devido ao grande número de funções de filtro e parâmetros, bem como à dificuldade de ajustá-los para obter as interpretações desejadas por meio do Mapper, foi desenvolvida uma versão menos generalizada da técnica Mapper. Essa nova técnica, chamada Ball Mapper, ou simplesmente *BM*, não depende de funções de filtro, sendo dependente apenas da massa de dados e requer apenas um parâmetro de ajuste, otimizando tempo na obtenção das análises (DŁOTKO, 2019). Esse novo algoritmo tem sido aplicado no estudo da disseminação da COVID-19 no Reino Unido (DŁOTKO; RUDKIN, 2020), no entendimento da falência de empresas (QIU; RUDKIN; DŁOTKO, 2020) e na compreensão da tomada de decisões financeiras (DŁOTKO; QIU; RUDKIN, 2022).

Outra ferramenta tão utilizada quanto Mapper em TDA, na investigação científica relacionada a grandes massas de dados na forma de nuvem de pontos, é a Homologia Persistente (HP). Introduzida por Edelsbrunner, Letscher e Zomorodian (2002), ela é definida como uma ferramenta algébrica capaz de capturar características topológicas, número de componentes conectados, *loops* e vazios, em dados, formas e funções numa abordagem multiescala (EDELSBRUNNER; HARER et al., 2008). Teve aplicações em diversos domínios como: na Biologia para reconhecer a forma de biomoléculas (AMÉZ-QUITA et al., 2020) e entender o padrão de crescimento de plantas (LI et al., 2017); em Oncologia na análise da forma de dados subjacentes (LESNICK, 2013); em conexões com Geometria Computacional e tarefas de *Machine Learning* para redução de dimensionalidade (CARLSSON et al., 2012); em séries temporais para, detecção de periodicidade e entender mudanças topológicas ao longo do tempo (PEREA; HARER, 2015; PEREA, 2019), e análise de dados financeiros (MA, 2020); em descrição e classificação de formas (CARLSSON et al., 2004); dentre outras.

Um dos entraves em processar grandes massas de dados, naturalmente, são os elevados custos e tempos computacionais e a consequente demora na obtenção das análises. Pensando em reduzir custo e tempo computacional nas análises, uma alternativa a ser investigada para alcançar ganho computacional ao aplicar técnicas da TDA é explorar os efeitos das Técnicas de Redução de Dados bem fundamentadas na Inferência Estatística, o que pode tornar o processo menos demorado e possibilitaria explorar mais aplicações em conjuntos de dados de alta dimensão, como imagens 3D de μCT com alta resolução obtidas de otólitos de peixes.

Este trabalho de tese propõe uma redução de dimensionalidade, usando técnicas de amostragem probabilística, em dados de imagens 3D de alta dimensão buscando ganho em tempo e custo computacional na análise de densidade e forma de otólitos de peixes. Especificamente, seria avaliar os efeitos da resolução de imagens 3D de μCT como um passo preliminar para explorar variações da densidade óssea de otólitos de peixes, utilizando a técnica BM da TDA, a baixo custo computacional. Ademais, estudar HP como um novo classificador para a forma do otólito, uma vez que os estudos de Vasconcelos-Filho et al. (2019) demonstraram também que existem vazios no interior dos otólitos, o que os colocam, do ponto de vista geométrico, na condição de serem tratados como objetos topológicos.

Os resultados dessas aplicações da Topologia em otólitos podem representar ampliações do conhecimento na Biologia Marinha, uma vez que os estudos sobre variações de densidade otolítica são escassos, e a topologia pode ajudar na definição de tamanhos de amostras menores possibilitando outras investigações. Um exemplo, é a possibilidade de expor um novo classificador com base na forma 3D completa do otólito, o que seriam uma abordagem alternativa as metodologias existentes na literatura, dentre elas: Análise de Discriminante e Transformada de Fourier (SALIMI et al., 2016); Método de Clusterização *K-means* (LI et al., 2021); Parâmetros Geométricos e Dimensão Fractal (PIERA et al., 2005; DUARTE-NETO et al., 2014; CONDAL; GUIDA, 2020); Transformada Wavelet e Descritores Elípticos de Fourier (VAN; Q.T; V.D, 2022).

As metodologias supra mencionadas apoiam-se em análise de imagens de contornos 2D sobre uma variedade de métodos relacionados ou combinados a outros já estabelecidos (STRANSKY, 2014). É então intuitivo pensar que ao considerar a estrutura completa do otólito (imagens 3D), resultados mais acurados de classificação pela forma possam emergir. Sobre essa suposição, ainda não foi proposta uma ferramenta unificada, que opere a baixo custo computacional, para tornar tal hipótese realidade. Pela aplicabilidade da TDA em diversos contextos e pelo fato dos otólitos poderem ser considerados como objetos topológicos, a HP apresenta potencial para ser aplicada como tal ferramenta de classificação para a forma do otólito sob uma redução apropriada da massa de dados.

2 Fundamentos teóricos

Este capítulo estrutura objetivamente a base teórica dos principais objetos de estudo desta tese: **Otólitos**, estruturas calcificadas encontradas nos ouvidos de diversos organismos e **Análise Topológica de Dados** – TDA, abordagem inovadora na ciência de dados que revela padrões e relações subjacentes em conjuntos de dados complexos.

Nele a TDA é introduzida como um conjunto de ferramentas da análise de dados e oferecida como uma nova abordagem para compreender a complexidade geométrica e topológica de estruturas como nuvem de pontos extraídas a partir de dados reais, podendo proporcionar *insights* valiosos para a interpretação de padrões na ciência do otólito. Basicamente, este capítulo estrutura uma nova ponte ligando a biologia marinha e as técnicas avançadas atuais de análise de dados, ao explorar o potencial da TDA na revelação de informações fundamentais contidas nos otólitos por meio imagens tridimensionais provenientes de microtomografia computadorizada de raios X (μCT).

2.1 Otólitos

Otólitos são estruturas calcificadas biomineralizadas encontradas no ouvido interno de peixes ósseos responsáveis por audição e equilíbrio (POPPER; FAY, 1993).

A partir do aparelho auditivo interno (ou sistema labiríntico, ou labirinto interno, ou ainda sistema vestibular) que se localizada na região superior da cabeça do peixe, em número de três por ouvido os otólitos ficam imersos em endolinfa (PAYAN et al., 2004) dentro de câmaras labirínticas conhecidas por utrículo - *utricle* do inglês, *lagena* e sáculo - *sacculus*. Assim, os otólitos, também possuem nomes específicos a partir de cada câmara labiríntica, sendo respectivamente, *lapillus*, *asteriscus* e *sagitta*.

O utrículo é responsável pela detecção de acelerações lineares e pela orientação espacial do peixe. A lagena é uma estrutura sensorial alongada que faz parte do sistema auditivo, pois está associada à detecção de sons e à percepção auditiva nos peixes. O sáculo é outra câmara e desempenha um papel semelhante ao utrículo na detecção de acelerações lineares, mas também está envolvido na detecção da posição vertical do corpo em relação à gravidade (PANFILI et al., 2002).

A Figura 3 ilustra o aparelho auditivo para um peixe da espécie *Diapterus brevirostris*. Nela o sistema vestibular é destacado a partir de sua posição em relação a cabeça

do peixe (parte traseira superior) para observar a localização de uma das duas câmaras labirínticas e de seu respectivo trio de otólitos.

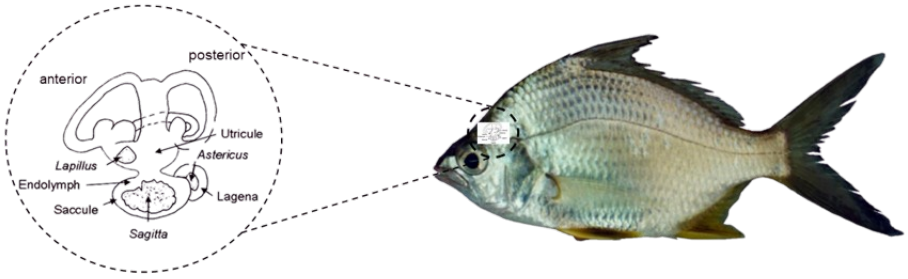


Figura 3 – Posição do sistema labiríntico em *Diapterus brevirostris* em relação a cabeça do peixe. Destacado está o diagrama dos canais semicirculares e a posição dos 3 otólitos, *lapilus*, *sagittae* e *asteriscus*, dentro de suas respectivas câmaras labirínticas *utricle*, *saccule* e *lagena*. Adaptado de Panfili et al. (2002, p.32) e Gallardo-Cabello et al. (2015).

A morfologia do otólito pode ser influenciada por fatores genéticos, ambientais e evolutivos (LOMBARTE; LLEONART, 1993; NOLF, 1995; TORRES; LOMBARTE; MORALES-NIN, 2000). Com isso variação de forma nos otólitos pode estar relacionada a adaptações específicas ao ambiente em que o peixe vive, como a profundidade da água, tipo de substrato, hábitos alimentares, entre outros. Por exemplo, otólitos de peixes que habitam águas profundas podem ter características diferentes daqueles de peixes que vivem em águas rasas. Na Figura 4 estão, em linha, cinco otólitos *sagittae* de cinco espécies de peixes, onde é possível observar uma significativa variação de tipos e formatos para otólitos de um mesmo tipo entre espécies diferentes.

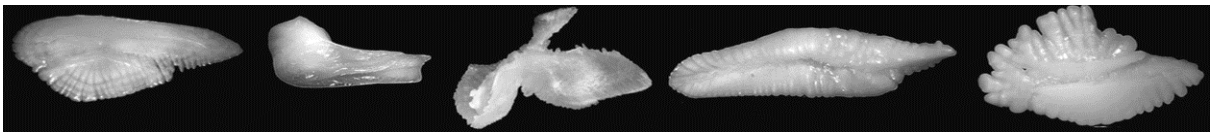


Figura 4 – Exemplos da diversidade de otólitos de diferentes formatos e tamanhos, presentes nos teleósteos. Em linha e sem escala estão alguns otólitos *sagittae* de *Merluccius bilinearis*, *Halargyreus johnsoni*, *Lampris guttatus*, *Urophycis tenuis* e *Lopholatilus chamaeleonticeps*. Pode-se perceber, a depender da espécie, que tamanho e forma de otólitos não possuem padrões definidos, podendo assumir projeções e invaginações específicas da espécie. Adaptado de Popper, Ramcharitar e Campana (2005).

Estudos biométricos sobre otólitos de peixes têm sido foco de inúmeras pesquisas científicas dentro da biologia marinha, em específico dentro da ictiologia, devido a diversidade de aplicações que eles permitem. Isso porque, esses biominerais versáteis oferecem

valiosas informações que permitem compreender aspectos sobre uma ampla gama de tópicos, incluindo crescimento de peixes, determinação de idade, avaliação de estoques, histórico ambiental, ecologia e até mesmo preferências alimentares. As pesquisas que apoiaram esses tópicos os tornaram abrangentes e consequentemente tornaram-se áreas de estudo sobre otólitos (GREEN et al., 2009), a saber: **microquímica ou microestrutura**, **formação, análise de idade e crescimento**, **morfotipo**, e **avaliação de estoques pesqueiros**. Esses campos de investigação ampliaram o conhecimento sobre otólitos apoiados no avanço tecnológico das **técnicas de processamento de imagens** (FISHER; HUNTER, 2018), perspectivando novas áreas de estudo dentro da ciência do otólito, a exemplo a **densidade** otolítica na sua forma estrutural.

- **Microquímica ou Microestrutura**

A estrutura dos otólitos é composta de aproximadamente 90% de carbonato de cálcio, na forma cristalina da aragonita, além de outros sais inorgânicos, que se desenvolvem sobre uma matriz proteica localizada no aparelho auditivo do peixe (SASAGAWA; MUGIYA, 1996; CAMPANA, 1999; CAMPANA; THORROLD, 2001; PANFILI et al., 2002). O carbonato de cálcio é difundido através de uma membrana celular de endolinfa e camadas de aragonita são permanentemente depositadas em incrementos discretos (WU et al., 2011). A microestrutura dos otólitos é essencial para examinar padrões diários de crescimento ao analisar os incrementos que por sua vez podem ser contados para fornecer informações sobre a vida diária do peixe (MORALES-NIN, 2000). Esses incrementos também podem ajudar a identificar mudanças ontogenéticas e estressores ambientais (KALISH, 1989; SPONAUGLE, 2010). A Figura 5 traz um exemplo de análise da composição química do estrôncio (Sr) na estrutura de um otólito, a partir da qual é possível observar a quantificação do elemento na imagem mais a direita.

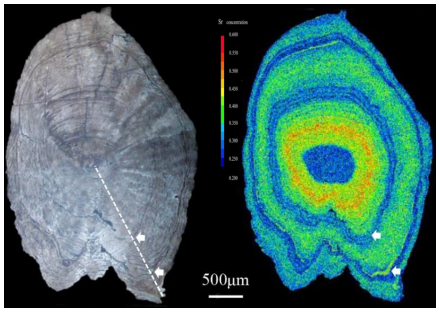


Figura 5 – Exemplo de análise microquímica do elemento *estrôncio* (Sr) para caracterização da estrutura de anéis de crescimento em otólitos, que por sua vez são úteis na determinação da idade de peixes. Geralmente este tipo de análise é feita partindo da região central do otólito a uma de suas extremidades, como revela a linha na figura da esquerda. Adaptado de Hu et al. (2022).

- **Formação, Idade e Crescimento**

Os otólitos são formados ao longo da vida do peixe através de um processo de deposição de carbonato de cálcio, que consiste na formação de anéis concêntricos, ou simplesmente anéis (GAULDIE; NELSON, 1990; MORALES-NIN, 2000). Tais anéis servem como registro da idade e do crescimento do peixe pela sua contagem (Figura 6). Esta característica fundamental, além de permitir aos cientistas estimar idade de peixes, fornece informações vitais para gestão pesqueira, esforços de conservação e outras investigações científicas decorrentes dentro da biologia e da ecologia marinha (GEFFEN, 1982; CAMPANA, 2001).

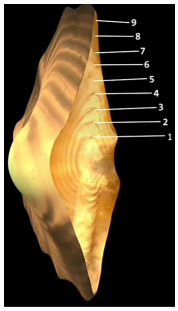


Figura 6 – Exemplo da contagem de anéis em um otólito seccionado transversalmente. Adaptado de NOAA's (2013).

- **Morfotipo**

A variação morfológica em otólitos, conhecida como morfotipos (Figura 7), tem sido reconhecida como uma ferramenta para distinguir espécies ou populações de peixes, devido ao seu formato específico entre espécies (de Almeida et al., 2023). Até mesmo mudanças morfológicas entre otólitos de indivíduos machos e fêmeas já são conhecidas a cerca de três décadas (SCHWARZHANS, 1994). A identificação de morfotipos ganhou importância, pois desempenha um papel crucial no estudo da biodiversidade e na compreensão dos processos ecológicos (DULVY et al., 2008; BOLLE et al., 2004). Além disso, por apresentar baixo teor de água, essas estruturas são úteis para estudos arqueológicos, paleontológicos e antropológicos, pela presença em restos fossilizados, o que ajuda na identificação e mapeamento de espécies por região (HAIMOVICI et al., 2023).

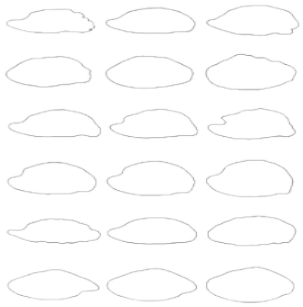


Figura 7 – Exemplo de morfotipos de otólitos analisados a partir da perspectiva do contorno. De cima para baixo otólitos *sagittae* de seis diferentes espécies. Ao centro, forma média do otólito da espécie, e aos extremos variação da forma a $\pm 2\sigma$. Adaptado de Pavlov (2021).

- **Avaliação de estoques pesqueiros**

A urgência em estabelecer um ciclo para tornar a pesca uma atividade sustentável é motivada pelo papel significativo que ela desempenha nos meios de subsistência em todo o mundo. Com o aumento constante do consumo de recursos naturais marinhos, surge uma preocupação crescente quanto à preservação desses recursos. Nesse contexto, torna-se imperativo estudar a gestão sustentável das pescas. A Figura 8 traz um esquema ilustrativo do papel dos envolvidos em uma gestão eficaz dos recursos pesqueiros, orientada a um consumo sustentável. Observe que o ciclo sustentável torna-se funcional pela colaboração conjunta de pescadores, colaboradores privados, cientistas e gestores públicos. Veja que a análise de dados, que podem ser otólitos, fornecidos por pescadores a pesquisadores para investigações, é essencial para o ciclo.

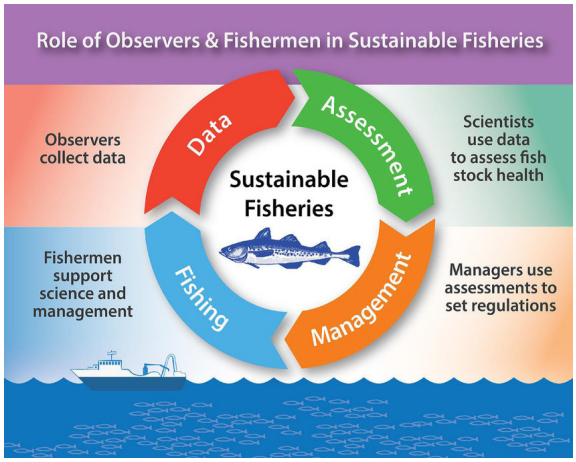


Figura 8 – Ilustração de um modelo de ciclo sustentável dos recursos pesqueiros. Nesse ciclo, proposto pela empresa NOAA Fisheries[a], observadores são contratados para avaliar e disponibilizar informações sobre o trabalho dos pescadores. Fonte: Wang e DiCosimo (2019).

[a] A NOAA Fisheries é uma agência dos Estados Unidos que faz parte da National Oceanic and Atmospheric Administration (NOAA). Seu foco principal é a gestão e conservação dos recursos pesqueiros e aquáticos marinhos.

Pode-se dizer, que a pesca sustentável está dentro dos objetivos 12 e 14 propostos pela ONU em sua agenda até 2030, pois eles buscam garantir padrões sustentáveis de consumo e produção e, conservar e utilizar de forma sustentável os oceanos, mares e recursos marinhos (UN, 2015). Para alcançar esses objetivos, os métodos científicos têm papel crucial ao permitirem uma compreensão da dinâmica das populações de peixes, logo, a implementação deles pode contribuir para um manejo mais eficaz dos recursos pesqueiros, visando a preservação dos ecossistemas marinhos e a garantia de que a pesca seja uma atividade sustentável a longo prazo (GEBREMEDHIN et al., 2021).

Por sua vez, otólitos podem servir como uma ferramenta inestimável para o endosso e gestão sustentável dos estoques pesqueiros. Isso porque ao integrar otólitos aos métodos científicos atuais, os cientistas podem identificar a estrutura genética das populações de peixes, o que pode ajudar a preservar a biodiversidade e apoiar uma gestão pesqueira eficaz (RUZZANTE et al., 2006; GEBREMEDHIN et al., 2021).

- **Densidade**

Os otólitos apresentam variações de densidade relacionadas à sua composição química. Compreender essas variações é crucial para refinar os métodos de determinação da idade (HU et al., 2022), além de abrir outras possibilidades de estudo, como avaliar se a densidade pode ser um indicador importante da flutuabilidade do peixe, o que forneceria informações sobre a distribuição vertical de peixes em colunas de água.

A densidade da estrutura física dos otólitos (densidade óssea), independentemente dos elementos químicos, pode não ser um tópico extensivamente estudado ou documentado na literatura científica. As pesquisa mais comuns tendem a focar nas características químicas dos otólitos, incluindo sua composição e variações na concentração de elementos, como viu-se nos tópicos anteriores acima. Recentemente uma nova área de estudo chamada eco-densitometria foi aberta por Vasconcelos-Filho et al. (2019) para quantificar variações de densidade óssea de otólitos de peixes ao revelar variações de densidade interna e externa de tal estrutura em uma perspectiva 3D (Figura 9).

Nesse sentido, a estrutura física dos otólitos, incluindo sua densidade, pode ser um aspecto relevante a ser explorado, especialmente considerando como diferentes propriedades físicas podem influenciar o comportamento mecânico e a função dos otólitos. Uma evolução ou uma pesquisa específica sobre a densidade da estrutura física como uma nova variável quantitativa do otólito pode começar a descrever diversos aspectos da biologia marinha ao associá-la a outras variáveis.

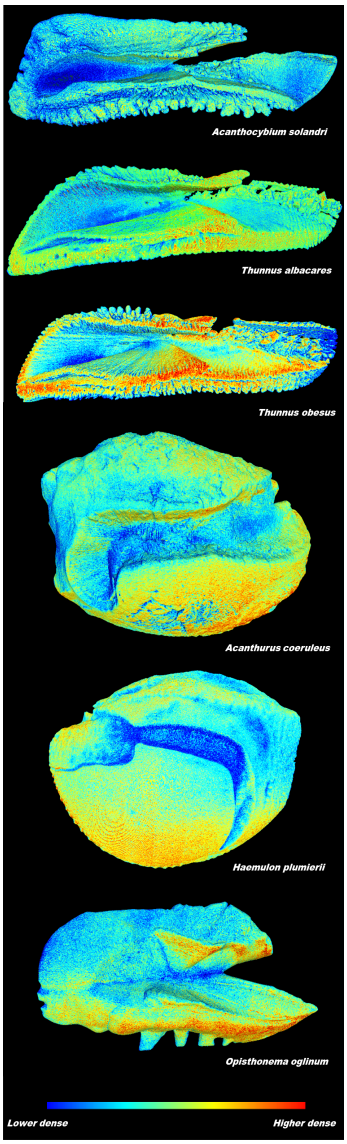


Figura 9 – Variações de densidade óssea em otólitos de várias espécies em uma perspectiva 3D. Imagens sem escala de tamanho. Adaptado de Vasconcelos-Filho et al. (2019).

2.2 Imagens Digitais para estudos de Otólitos

As técnicas de análise e processamento de imagens apresentam uma relevância dual, porque destacando-se por duas características fundamentais. Primeiramente, contribuem significativamente para a melhoria da qualidade visual das imagens, visando facilitar a interpretação humana. Em segundo lugar, possibilitam a extração eficiente de informações intrínsecas, as quais podem ser correlacionadas a variáveis específicas, propiciando uma expansão do conhecimento relacionado ao objeto fotografado (GONZALEZ; WOODS, 2018) – ver Figura 10 para exemplo de uma imagem digital obtida do crânio de um peixe. Nas últimas décadas, a introdução e aprimoramento dessa tecnologia conferiram robustez às metodologias de investigação otolítica, resultando em *insights* valiosos sobre a biologia, ecologia e dinâmica populacional de diversas espécies aquáticas (MENDOZA, 2006; STRANSKY, 2014; CAMPANA, 2005).

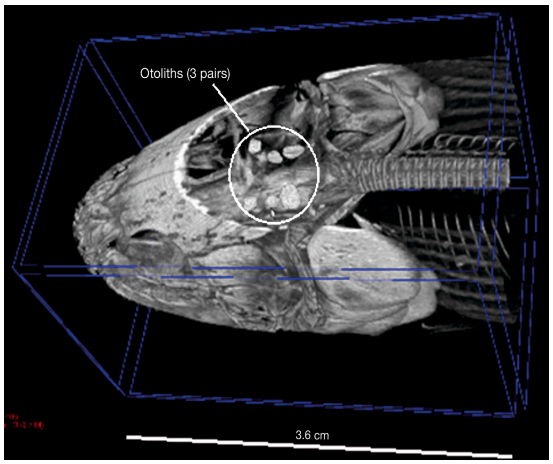


Figura 10 – Imagem digital do crânio de um peixe identificando os 3 pares de otólitos. Tal imagem enfatiza o nível de resolução atual da tecnologia de imagens digitais (FISHER; HUNTER, 2018).

Ferramentas de processamento de imagens, combinadas com algoritmos avançados de *machine learning*, facilitam a identificação e medição de características otolíticas de maneira mais eficiente (MOORE et al., 2019). Com isso, a capacidade de capturar imagens de alta resolução permite uma análise detalhada das características morfológicas dos otólitos, forma, tamanho e microestrutura (NAZIR; KHAN, 2021). Isso não apenas facilita a identificação de espécies, mas também possibilita estudos mais refinados sobre variações individuais e populacionais ao promover resultados mais consistentes e confiáveis, além de reduzir o tempo necessário para análises (TUCKEY et al., 2022).

É então apropriado dizer que, a Análise Digital de Imagens interdisciplinariza os campos de estudos sobre otólitos, abrangendo diversas áreas como anéis de crescimento, microestrutura, microquímica, estudo da forma, entre outros. A aplicação de técnicas de

análise e processamento de imagens digitais para análise de otólitos não apenas aprimora a eficiência dos processos, mas também facilita a integração de diferentes disciplinas, contribuindo para uma compreensão mais abrangente e precisa das características otolíticas.

Na análise de anéis de crescimento, a Análise Digital de Imagens possibilita uma contagem precisa e eficiente, fornecendo dados cruciais para estimar a idade dos peixes (NAVA et al., 2018). A microestrutura dos otólitos, incluindo detalhes finos muitas vezes imperceptíveis a olho nu, pode ser minuciosamente investigada por meio de imagens digitais de alta resolução (REY et al., 2016). No campo da microquímica, a análise digital contribui significativamente para a identificação e quantificação de elementos químicos presentes nos otólitos, auxiliando na compreensão das condições ambientais durante o histórico de crescimento do peixe (ELSDON et al., 2008). A análise da forma dos otólitos, essencial para identificação de espécies e estudos de ecologia, também se beneficia significativamente da precisão proporcionada pelas imagens digitais (CADRIN; FRIEDLAND, 1999).

Em suma, o uso de imagens digitais na análise de otólitos representa um avanço significativo na pesquisa ictiológica e na gestão de recursos aquáticos. Essa abordagem proporciona uma compreensão mais profunda da biologia e ecologia das espécies, com implicações diretas na conservação e no manejo sustentável dos ecossistemas aquáticos. O constante aprimoramento das técnicas digitais promete continuar a contribuir significativamente para a expansão do conhecimento sobre otólitos e seu papel vital na vida marinha.

Percebe-se então que, no avanço da tecnologia de imagens, destaca-se a utilização de raios X como um marco significativo no estudo de otólitos pela alta taxa de resoluções que apresentam (HABERTHÜR et al., 2023). A exemplo, essa técnica de captura de imagens tem a capacidade de analisar a densidade óssea, oferecendo assim uma nova perspectiva para compreender outras propriedades dessas estruturas.

2.3 Análise Topológica de Dados

Uma maneira simples de entender a Análise Topológica de Dados, ou *TDA* do inglês, é refletir sobre as palavras que compõem seu nome como três conceitos fundamentais, que seriam Análise, Topologia e Dados, sobre a busca por compreender características intrínsecas e relações subjacentes em conjuntos complexos de dados.

A Análise refere-se à investigação sistemática e ao entendimento profundo de padrões e informações contidas nos dados, mas também pode ser pensada na subárea da matemática "Análise", uma vez que esta compõe os fundamentos teóricos da topologia, também presente em TDA.

Topologia é o ramo da matemática que explora as propriedades do espaço e suas estruturas fundamentais tratando-os matematicamente como espaços topológicos. Esses espaços representam formas que são analisados em uma perspectiva de dobrar e esticar sem "rasgar" ou eliminar quaisquer furos, assim ela estuda as propriedades dos espaços que permanecem inalteradas sob transformações ou deformações contínuas. Essa característica intrínseca da Topologia dá a ela o poder de ser considerada como uma extensão da Geometria.

As propriedades das formas que são preservadas pela noção de flexibilidade na Topologia, como o número de furos que uma forma possui em cada dimensão, são chamadas de invariantes topológicos. Invariantes Topológicos são números que se atribui a dados, forma, variedade, ou complexo simplicial, para extrair informação ou aprender algo sobre a estrutura geral de tais estruturas.

E por fim, Dados, representa as informações a serem analisadas. A TDA lida com conjuntos de dados de alta dimensionalidade ou dados multidimensionais. Ela utiliza as propriedades da Topologia, na busca por identificar padrões complexos, capturar relações espaciais e estruturas, além de conexões e características fundamentais que podem não ser óbvias utilizando métodos tradicionais.

Nesse sentido, a Topologia tem conexões com áreas como Análise, Geometria e Topologia Algébrica (MUNKRES, 2000; DEO, 2018). Dentre essas conexões, a intersecção da Topologia com técnicas algébricas levou ao surgimento da Topologia Algébrica, que por sua vez abriu caminho em diversos domínios científicos orientados a dados, resultando no desenvolvimento da Análise Topológica de Dados (GHRIST, 2007; GHRIST, 2014). Assim, a TDA envolve Ciência de Dados, Computação, Matemática e Estatística na fundamentação e criação de algoritmos capazes de resumir informações de dados complexos e de alta dimensão (CARLSSON, 2009; EDELSBRUNNER; HARER, 2022).

A Topologia Algébrica serve como ponte que conecta a Topologia clássica com aplicações modernas envolvendo dados, ao entregar ferramentas algébricas, por meio de algoritmos, para investigar espaços topológicos, permitindo a identificação de características importantes como componentes conectados, *loops* e vazios. Além disso, ao atribuir estruturas algébricas, como grupos ou classes de homologia, a espaços topológicos, a Topologia Algébrica pode capturar e quantificar suas propriedades geométricas intrínsecas (KAMMEYER, 2022; MUNKRES, 2018).

Atualmente, o desenvolvimento da TDA é impulsionado pelas mesmas razões que levaram a sua criação, a crescente demanda por reconhecer padrões nos grandes conjuntos de dados (ou *Big data*) que são recolhidos pelo poderio tecnológico atual de vários campos da ciência, como a Biologia, a Neurociência, Mercados Financeiros, as Redes de Sensores,

dentre outros. Tais dados, por sua vez estão cada vez mais complexos e com dimensões mais altas, e como consequência os métodos tradicionais de análise de dados muitas vezes ficam aquém em capturar suas estruturas inerentes.

Nessa nova realidade da era dos dados, a TDA aproveita os conceitos e métodos da Topologia Algébrica para revelar propriedades topológicas dos dados através de ferramentas computacionais como Mapper, Ball Mapper (BM) e Homologia Persistente (HP), fornecendo uma lente através da qual conjuntos de dados complexos podem ser estudados, revelando características e relacionamentos essenciais. Essas técnicas enriqueceram a capacidade de extrair *insights* de dados complexos, oferecendo uma nova perspectiva sobre as estruturas e formas subjacentes "escondidas" em conjuntos de dados (BAAS et al., 2020).

Mapper é uma ferramenta da TDA que constrói uma representação gráfica dos dados na forma de um grafo, onde os nós correspondem a clusters de pontos de dados e as arestas indicam suas conexões. Ele oferece uma maneira perspicaz de simplificar e visualizar conjuntos de dados complexos e de alta dimensão em dimensões mais baixas, permitindo que os pesquisadores obtenham uma melhor compreensão da estrutura subjacente dos dados em um simples grafo, a partir do qual propriedades topológicas podem ser extraídas.

O BM, uma simplificação do Mapper, concentra-se na captura de estruturas locais usando bolas ou vizinhanças em torno de pontos dos dados. Essa técnica capaz de destacar detalhes refinados dos dados encontrou aplicações em áreas como biologia, mercado imobiliário e financeiro, ciência dos materiais e pandemia de COVID-19.

A HP, uma das técnicas mais renomadas em *TDA*, rastreia a evolução de características topológicas em diferentes escalas espaciais. Ao fazer isso, fornece uma ferramenta robusta e estável para mensurar características topológicas nos dados, o que pode ser particularmente valioso na compreensão da persistência de padrões estruturais em dados.

Por ser uma das primeiras técnicas para análise de dados desenvolvidas dentro da TDA, a HP é uma das mais utilizadas, o que resultou na criação de muitos algoritmos para sua computação, a fim de atender as diversas aplicações científicas. Nisso houve o interessante trabalho de Otter et al. (2017), em que o interesse foi conhecer qual ferramente ou software da TDA melhor desempenha a computação da topologia na análise de dados. Daquele ano em diante, atualizações surgiram, bem como a criação de novos softwares, como o `giotto-tda`, uma biblioteca em `Python` criada por colaboração e contribuições de diversos pesquisadores, a fim de fornecer ferramentas para a análise topológica de dados, incluindo mapper e a computação da homologia persistente (TAUZIN et al., 2021).

Como consequência dos significativos resultados alcançados, as técnicas de análise de dados Mapper e HP, além de se tornarem chave dentro da TDA, devido as suas grandes

capacidades em analisar e interpretar dados de alta dimensão, foram constantemente aperfeiçoadas e conectadas a ferramentas da área de *Machine Learning*, aumentando a gama de aplicações dentro da ciência de dados.

Em geral, em aplicações de TDA, os dados são considerados como espaços topológicos, essa perspectiva permite seu desenvolvimento e aprimoramento como metodologia para análise exploratória de dados multiescala, fazendo com que ela alcançasse aplicações significativas em ramos como Genômica e Evolução (RABADÁN; BLUMBERG, 2019), identificação de câncer de mama (NICOLAU; LEVINE; CARLSSON, 2011), reavaliação de índices financeiros e retornos de ações (DŁOTKO; QIU; RUDKIN, 2019), dentre outras.

Em suma, na jornada interdisciplinar da Topologia à Topologia Algébrica que levou ao desenvolvimento da Análise Topológica de Dados, abstrações matemáticas foram transformadas em algoritmos capazes de extrair *insights* significativos de dados em diferentes contextos. Esta introdução prepara caminho para um aprofundamento nos princípios e aplicações da *TDA*, destacando a sua capacidade de desvendar as estruturas topológicas complexas incorporadas nos dados, tornando-a uma surpreendente ferramenta em vários cenários científicos.

Nesta tese será proposta a introdução da Amostragem Probabilística na TDA, como ferramenta que pode ser útil na redução de tempo e custo computacional, propiciando análises e interpretações mais rápidas. Na TDA, o conjunto de dados, que representa o objeto de interesse, tratado como espaço topológico, muitas vezes é uma matriz de dados na forma de nuvem de pontos que serve de dados de entrada para os algoritmos. A ideia é amostrar as nuvens de pontos que representam o conjunto de dados, a ser analisado topologicamente, usando Amostragem Probabilística, uma vez que, fundamentalmente, ela é capaz de capturar as informações da populações dos dados.

Diante dessa motivação em alcançar economia e tempo computacional, além de esforço humano, o texto segue com os elementos base da Topologia Algébrica, essenciais para compreensão dos aspectos práticos das técnicas da TDA, necessários para desenvolver as aplicações propostas por esta tese.

Elementos da Topologia Algébrica para computação da topologia

A amostragem tem a capacidade de equilibrar custo de processamento com análise de grandes volumes de dados, porque fornece informações estatísticas similares às dos dados populacionais a partir de amostras menores. Assim a amostragem pode ser conceituada como um processo que seleciona aleatoriamente objetos de uma população, de tal forma que as informações da amostra possam ser generalizadas para a população, produzindo redução de tempo e custo computacional (KARTHIK; ABHISHEK, 2019, pp. 79–81).

Amostra aleatória

Definição 2.3.1. Um conjunto de variáveis aleatórias $X_1, ..., X_n$ é dito *amostra aleatória de tamanho n da população* $f(x)$, se $X_1, ..., X_n$ são variáveis aleatórias mutuamente independentes e as marginais fdp ou fmp de cada X_i possui a mesma função $f(x)$. Alternativamente, $X_1, ..., X_n$ são ditas *variáveis aleatórias independentes e identicamente distribuídas com fdp ou fmp* $f(x)$, comumente abreviadas por variáveis aleatórias iid (BERGER; CASELLA, 2001, p. 207).

Espaço Euclidiano

Definição 2.3.2. Um Espaço euclidiano n-dimensional, ou Espaço n-euclidiano ($\mathbb{E}^n$), é um conjunto de todas as n-tuplas $(x_1, x_2, ..., x_n)$ de números reais, com cada n-tupla sendo um ponto do espaço $\mathbb{R}^n$. Observando que o Espaço 1-euclidiano são os números reais ($\mathbb{R}$) e o Espaço 2-euclidiano é o plano ($\mathbb{R}^2$), os $\mathbb{E}^n$s são considerados generalizações dimensionais superiores dos números reais (KAHN, 1995, p. 16). A distância d entre dois pontos $(x_1, x_2, ..., x_n)$ e $(y_1, y_2, ..., y_n)$ em um $\mathbb{E}^n$ é dada por $d(x_i, y_i) = \left\{\sum_{i=1}^{n}(x_i - y_i)^2\right\}^{1/2}$, a qual denomina-se por distância euclidiana.

Nuvem de pontos

Definição 2.3.3. Uma nuvem de pontos é um subconjunto finito $\mathbb{X}$ do $\mathbb{R}^n$ (para algum $n \in \mathbb{N}$) no qual possa ser induzida a métrica da distância euclidiana.

Vizinhança

Definição 2.3.4. Uma vizinhança de um ponto p no Espaço Euclidiano é o conjunto de todos os pontos a uma distância r de p para algum $r \in \mathbb{R}$.

Espaço Topológico

Definição 2.3.5. Um Espaço Topológico é um conjunto abstrato E com uma atribuição não vazia de subconjuntos de E a cada elemento de E. Os elementos de E serão chamados pontos p, e os subconjuntos atribuídos a cada ponto p de E serão chamados de vizinhança de p. A atribuição de vizinhança a cada ponto de E deve ser sujeita às condições:

- Se U é uma vizinhança de p, então $p \in U$;
- Qualquer subconjunto de E que contenha uma vizinhança de p também é uma vizinhança de p;

- Se U e V são vizinhanças de p, então $U \cap V$ também é uma vizinhança de p;
- Se U é uma vizinhança de p, então existe uma vizinhança V de p tal que U é uma vizinhança de todo ponto de V.

O processo de atribuir vizinhanças a cada elemento de um conjunto E, transformando E em um espaço topológico, às vezes é chamado de dar uma topologia a E ou definir uma topologia em E (WALLACE, 2011, p.14).

Bola

Definição 2.3.6. Diz-se que ${B_r}^n(c)$ é uma bola aberta centrada em c em um $\mathbb{E}^n$, se

$$ {B_r}^n(c) = \{p \mid p \in \mathbb{E}^n,\ d(p,c) < r\}, $$

e fechada se

$$ {B_r}^n(c) = \{p \mid p \in \mathbb{E}^n,\ d(p,c) \leq r\}. $$

As notação $B(x,r)$ e $B_r(x)$ também são usadas para indicar uma bola de raio r centrada em x.

Cobertura

Definição 2.3.7. Dado um espaço topológico $\mathbb{X}$, uma coleção de conjuntos abertos $\mathcal{U} = U_i$ é chamada de cobertura de $\mathbb{X}$ se a união dos conjuntos em $\mathcal{U}$ cobre todo o espaço $\mathbb{X}$, ou simplesmente,

$$ \mathbb{X} = \bigcup_i U_i. $$

Espaço métrico

Definição 2.3.8. Um conjunto A com uma função distância

$$ d : X \times X \to \mathbb{R} $$

é dito ser um espaço métrico se os valores de d são todos não negativos e para todo $x, y, z \in X$

$$ d(x,y) = 0 \Leftrightarrow x = y, $$

$$ d(x,y) = d(y,x), $$

$$ d(x,z) = d(x,y) + d(y,z). $$

Nesse caso, d é referido como uma métrica ou função de distância entre pontos do espaço.

Espaços métricos são uma grande e importante classe dos espaços topológicos (Definição 2.3.5) apoiado pelos números reais e espaços euclidianos (Definição 2.3.2), em outras palavras, todo espaço métrico A pode ser considerado naturalmente como um espaço topológico, desde que se tome em A uma coleção de abertos definidos a partir de uma métrica em A (LIMA, 1970, p. 61).

Espaços métricos também estão apoiados sob o conceito de vizinhança (Definição 2.3.4), assim, em um espaço métrico A, um subconjunto $A' \subseteq A$ é dito ser vizinhança de um elemento x a uma distância δ em A se ele contiver uma bola $B(x, \delta)$ (Definição 2.3.6). Essa definição, informalmente, diz que uma vizinhança de x é um conjunto que contém todos os pontos suficientemente próximos de x (AHLFORS, 1979).

Grupo Abeliano

Definição 2.3.9. É uma classe de conjuntos A equipada com a operação binária adição $(+)$ satisfazendo as seguintes propriedades:
1. Associatividade: Para todos os elementos a, b, c em A, $(a + b) + c = a + (b + c)$;
2. Identidade aditiva: Existe um elemento 0 em A tal que, para todo a em A, $a + 0 = a$;
3. Inversos aditivos: Para cada a em A, existe um elemento $-a$ em A tal que $a + (-a) = 0$;
4. Comutatividade (propriedade abeliana): Para todos os elementos a, b em A, $a + b = b + a$.

Grupos de Homologia

Definição 2.3.10. Dado um espaço topológico $\mathbb{X}$, o n-ésimo grupo de homologia, denotado por $H_n(\mathbb{X})$, é um grupo abeliano que representa os n-buracos em $\mathbb{X}$. Esse grupo mede a presença e a conectividade de componentes n-dimensionais em $\mathbb{X}$. Os elementos de $H_n(\mathbb{X})$, chamados de classes de homologia, representam ciclos topológicos em $\mathbb{X}$, enquanto as operações de grupo refletem a sobreposição e anulação desses ciclos.

Os grupos de homologia no contexto da TDA são ferramentas matemáticas utilizadas para estudar as propriedades topológicas dos espaços. Eles oferecem uma maneira de quantificar "buracos" e "ciclos" em espaços topológicos, sendo essenciais para analisar as características topológicas de conjuntos de dados.

Os ciclos aos quais se refere a Definição 2.3.10 é um subconjunto de $\mathbb{X}$ que pode ser coberto por uma coleção finita de complexos simpliciais, que são os elementos básicos usados para reconstruir um espaço, forma, ou mesmo conjunto de dados.

Complexo simplicial

Complexos simpliciais podem ser entendidos como uma generalização de grafos em dimensões mais altas. Eles partem da abstração sobre a composição de conjuntos em Álgebra, e podem ser definidos do seguinte modo:

Definição 2.3.11. Um *complexo simplicial abstrato* $\mathbb{S}$ é um conjunto de conjuntos de modo que, se $t \in S$ tal que $S \subset \mathbb{S}$ então $t \subset \mathbb{S}$. Os elementos t, denominados *simplex* —menor unidade de um *complexo simplicial*—, formam os subconjuntos Ss de $\mathbb{S}$ denominados de *simplices*—plural de *simplex*—, que compõem o *complexo simplicial*.

Em termos práticos, complexos simpliciais são estruturas geométricas que descrevem a "forma" de um espaço topológico. Originalmente a construção de complexos simpliciais é baseada no processo de triangulação, sobre a ideia de que por este procedimento pode-se construir qualquer forma através da união de triângulos. A ideia é conectar pontos não coplanares para criar estruturas geométricas mais complexas. Esse padrão continua para dimensões superiores de modo que cada dimensão superior adiciona um novo vértice, criando simplices de ordem mais alta.

A Figura 11 traz um esboço do procedimento da construção de um complexo simplicial abstrato desde o seu elemento mais simples, o ponto.

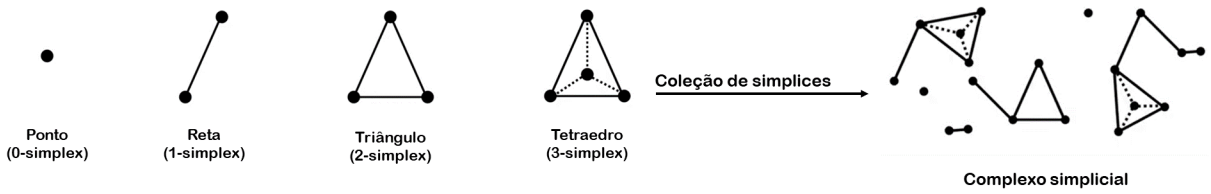


Figura 11 – A construção de um complexo simplicial começa com simples pontos e progride para estruturas mais complexas. Ponto (0-simplex): são pontos individuais denominados vértices, cada ponto é um 0-simplex. Pontos conectados formam arestas, denominadas 1-simplex (segmentos de reta). Grupos de três pontos não colineares formam triângulos, que são os 2-simplex, e grupos de quatro pontos não coplanares formam tetraedros, que são denominados 3-simplex. A união desses elementos formam o complexo simplicial. Fonte: Autor, 2023.

A triangulação, portanto, envolve a criação de complexos simpliciais adicionando essas estruturas mais altas, garantindo que elas respeitem a condição da Definição 2.3.11 e que cada face (k-simplex) tem uma fronteira formada por $(k-1)$-simplices. Essa construção é fundamental em Topologia Algébrica e TDA para representar a topologia de conjuntos de dados complexos.

O ponto de partida para a construção de um complexo simplicial é um conjunto de dados que se deseja conhecer informações através de características topológicas. Existem diferentes abordagens para a construção de complexos simpliciais, dentre elas os dois tipos mais comuns são, o Vietoris-Rips (Rips) e de Čech (Čech).

Enquanto o Rips forma arestas entre pontos que são próximos com base na distância euclidiana, o Čech considera discos em torno dos pontos e forma arestas se esses discos

não se sobrepõem a outros discos (situação em que dois discos não compartilham muitos pontos em comum). Em efeito prático o Rips forma arestas (1-simplex) diretamente entre pontos próximos, resultando em uma representação mais "densa" da conectividade e o Čech inclui arestas entre pontos que estão dentro de discos que não se sobrepõem a outros discos centrados em diferentes pontos, proporcionando uma representação mais "espaçada" da conectividade.

Portanto, a escolha entre eles depende das características específicas do conjunto de dados e do tipo de conectividade que se deseja capturar. A ilustração dos dois para uma mesma nuvem de pontos e para um mesmo raio de filtração pode ser vista na Figura 12.

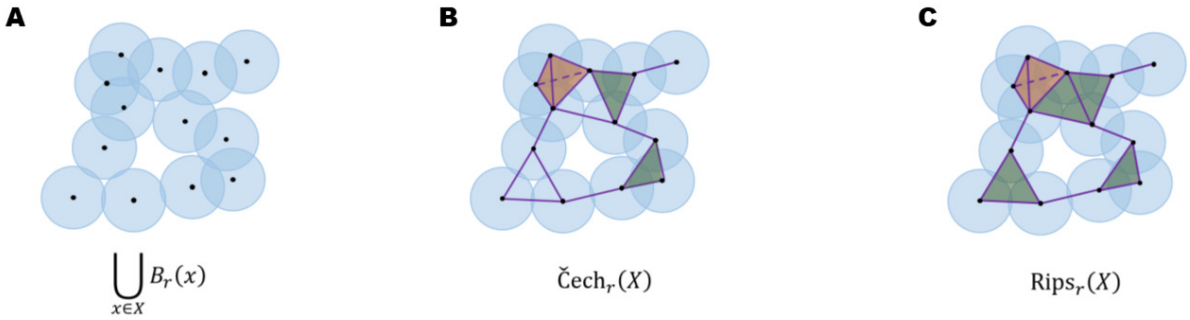


Figura 12 – (A) Nuvem de pontos $\mathbb{X}$ (Definição 2.3.7) coberta por bolas $B_r(x)$ (Definição 2.3.6). (B) Complexo simplicial de Čech - $\check{\text{C}}\text{ech}_r(\mathbb{X})$ e (C) Vietoris-Rips - $\text{Rips}_r(\mathbb{X})$ construídos sobre uma mesma nuvem de sete pontos $\mathbb{X}$ do plano $\mathbb{R}^2$ para um mesmo raio $r > 0$. No complexo de Čech faces de triângulos são obtidas sempre que três bolas se intersectam simultaneamente entre si, a formação de tetraedro acontece quando quatro bolas se conectam do mesmo modo. Para a formação destes mesmos elementos, triângulos e tetraedros, por parte do Rips, a intersecção não necessariamente deve ser simultânea. A partir da imagem pode-se perceber que existe uma relação de pertinência entre esses complexos definida por $\check{\text{C}}\text{ech}_r(\mathbb{X}) \subseteq \text{Rips}_r(\mathbb{X}) \subseteq \check{\text{C}}\text{ech}_{2r}(\mathbb{X})$. Observe ainda que esses complexos podem ser construídos em qualquer $\mathbb{X} \subseteq \mathbb{R}^d$. Adaptado de Reani e Bobrowski (2021).

A construção dos Complexos Simpliciais de Čech e Rips parte de um conjunto de dados fornecido como uma nuvem de pontos $\mathbb{X}$ (Definição 2.3.3) constituída pelos elementos $p_1, p_2, p_3, ..., p_n$. Em seguida são colocadas bolas B de raio r centradas em cada elemento p_i de $\mathbb{X}$ de modo a intersecção entre bolas deve ser não vazia.

Ao fazer r crescer de 0 a $+\infty$ vai-se produzindo complexos simpliciais que contêm os de raios menores, provocando a conhecida sequência aninhada de complexos, a que induz o isomorfismo dos grupos de homologia (Definição 2.3.10) nos dados. A evolução de características topológicas, tais como furos e componentes conectados, resultantes desse processo, chama-se filtração e define a homologia persistente da estrutura dos dados, de onde informações podem ser inferidas.

Filtração

Definição 2.3.12. Seja X um espaço topológico e $\mathcal{F} = \{X_t\}_{t \in I}$ uma família indexada por I de subconjuntos de X, onde I é um conjunto totalmente ordenado. Uma filtração é uma sequência de conjuntos:

$$X_0 \subseteq X_1 \subseteq \ldots \subseteq X_t \subseteq \ldots \subseteq X,$$

onde t é o parâmetro de filtração. Tal sequência de espaços topológicos juntamente com os homomorfismos de inclusão $\iota_t : X_t \hookrightarrow X_{t+1}$ para cada $t \in I$, são funções naturais de inclusão, satisfazendo $X_t \subseteq X_{t+1}$ para todo t.

Uma filtração é uma sequência crescente de conjuntos topológicos que capturam a evolução dos "buracos" em um complexo simplicial à medida que os elementos são adicionados. Essa sequência representa a adição gradual de elementos a X, formando grupos de homologia $H_k(X_i)$. Para representar a combinação direta de informações topológicas em diferentes níveis de uma filtração é usada a soma direta $\oplus$. Por exemplo, havendo grupos de homologia $H_k(X_i)$ associados a diferentes subconjuntos X_i na filtração, a soma direta é escrita como $H_k(X_1) \oplus H_k(X_2) \oplus \ldots \oplus H_k(X_n)$, e representa a computação das características topológicas do espaço topológico X_i.

A homologia persistente associa os grupos de homologia aos diferentes conjuntos da filtração e examina como os buracos topológicos evoluem à medida que os elementos são adicionados. A persistência é então capturada pelos intervalos de vida dos ciclos homológicos nas diferentes dimensões. Esses intervalos fornecem informações sobre a topologia que são invariantes em relação a pequenas perturbações nos dados. A contagem das características topológicas, "buracos", durante o processo de filtração é calculada pelos chamados Números de Betti.

Números de Betti

Definição 2.3.13. São uma classe de invariantes topológicos usados para distinção de espaços topológicos, tais como variedades compactas e outros conjuntos finitos, apoiado na conectividade de complexos simpliciais n-dimensionais.

Os números de Betti são invariantes topológicos que fornecem informações sobre a conectividade e a presença de buracos em espaços topológicos. Diz-se que o número de Betti β_j representa o j-ésimo grupo de homologia. Em termos práticos, a Homologia caracteriza conjuntos, tratados como espaços topológicos, atribuindo grupos abelianos a cada dimensão, de modo que cada grupo descreve características topológicas específicas, como componentes conectadas, ciclos ou buracos, que por sua vez são quantificados pelos

números de Betti, indicando o número de "buracos" em diferentes dimensões presentes no espaço. Deste modo, β_0 representa o número de componentes conectados ou componentes conexos, β_1 representa o número de furos unidimensionais, β_2 representa o número de furos bidimensionais e assim por diante

Como exemplo de números de Betti em um objeto geométrico, podemos pensar em um toro (objeto topológico tridimensional), Figura 13. Ele possui, um componente conectado β_0 que é sua própria estrutura, dois furos unidimensionais β_1 (um deles é correspondendo ao orifício central, ou o *loop* central do toro, onde se colocaria o dedo imaginando o toro como um anel - círculo vermelho maior na imagem. O outro é correspondente ao orifício ao redor da parte externa do toro - círculo vermelho menor) e um vazio, seu interior. Em resumo os números de Betti do toro são $\beta_0 = 1$, $\beta_1 = 2$ e $\beta_2 = 1$.

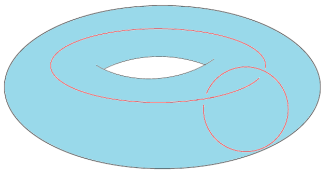


Figura 13 – Um toro possui números de Betti iguais a $\beta_0 = 1$, $\beta_1 = 2$ e $\beta_2 = 1$, uma vez que ele é um objeto simplesmente conectado, possui dois *loops* ou "buracos" de dimensão 1 (linhas vermelhas) e um vazio, "buraco" de dimensão 2, seu interior. Adaptado de Zomorodian e Carlsson (2005).

Grafo

Definição 2.3.14. Um grafo abstrato G é uma representação topológica de um conjunto S, discretizado em subconjuntos finitos X, na forma do par (V, E), onde V são vértices $(0 - simplex)$ e E são arestas $(1 - simplex)$ do grafo $G = (V, E)$ (GHRIST, 2014, p. 26).

Em outras palavras, um grafo pode ser considerado como uma coleção finita de 1-simplices unidos pelos seus vértices, o que o torna um complexo simplicial de dimensão 1 (KAHN, 1995, p. 184), deste modo os 1-simplices são as arestas que conectam os vértices 0-simpleces.

2.3.1 Homologia Persistente

Para compreender Homologia Persistente, antes é interessante entender o conceito de homologia. Homologia é uma área da Topologia com fundamentos na teoria de grupos capaz de caracterizar conjuntos de dados pela contagem das características topológicas *número de componentes conectados* e *furos* (HATCHER, 2002; FASY et al., 2014; EDELSBRUNNER; HARER, 2022).

Homologia Persistente refere-se a um método, dentro da teoria de homologia, que associa a cada espaço topológico uma série de grupos, ou de modo geral módulos, chamados grupos de homologia desse espaço, de tal maneira que espaços homeomorfos possuem grupos de homologia isomorfos (LIMA, 2021). Deste modo, homologia persistente trata-se de uma técnica dentro da matemática aplicada.

Desenvolvida a partir dos anos 2000, tem tido aplicações significativas em TDA desde então. A partir daquele ano, um dos primeiros estudos a apresentar a ideia de homologia persistente e a propor algoritmos eficientes para calcular *persistent homology* é atribuído ao trabalho de Silva, Morozov e Vejdemo-Johansson (2004). Antes, houve o trabalho de Robins (1999), o qual já buscava calcular características topológicas via homologia, porém apenas propondo um algoritmo para calcular tal tarefa sem aplicação prática. Anteriormente a esse, os trabalhos que forneceram apoio referencial relacionados a Topologia Algébrica, foram os estudos de Dey e Guha (1998) e Delfinado e Edelsbrunner (1995) em complexos simpliciais e números de *Betti*.

Aspectos práticos da Homologia Persistente

Em termos práticos, Homologia Persistente é uma abordagem dentro da teoria de homologia que mensura características topológicas em conjuntos de dados numa perspectiva multiescala (EDELSBRUNNER; LETSCHER; ZOMORODIAN, 2002; EDELSBRUNNER; HARER et al., 2008; EDELSBRUNNER; HARER, 2022).

Tal mensuração se dá pelo nascimento e desaparecimento de características topológicas quando há a variação de um parâmetros a partir de cada observação do conjunto, ditos componentes conectados. A medida que o valor do parâmetro cresce, componentes se fundem em componentes maiores, diz-se que alguns morrem e outros nascem, e nesses termos configura-se o "nascimento" e "morte" de características topológicas. O registro dessas características podem ser feitas, a critério de interpretação, em um *barcode* ou *diagrama de persistência*, veja Figura 14.

No diagrama de persistência (Figura 14F) são registadas pontos bidimensionais acima de uma diagonal (função identidade positiva), onde os x_is representam o momento de "nascimento" e os y_is representam o momento da "morte" de um componente. É evidente que pontos mais distantes da diagonal são os que possuem maior tempo de vida, ou seja, os que mais persistem.

Conhecidos os fundamentos teóricos de Homologia que levam à aplicação da Homologia Persistente, uma definição que combine uma explicação geral sobre ambos esses conceitos pode ser estabelecida.

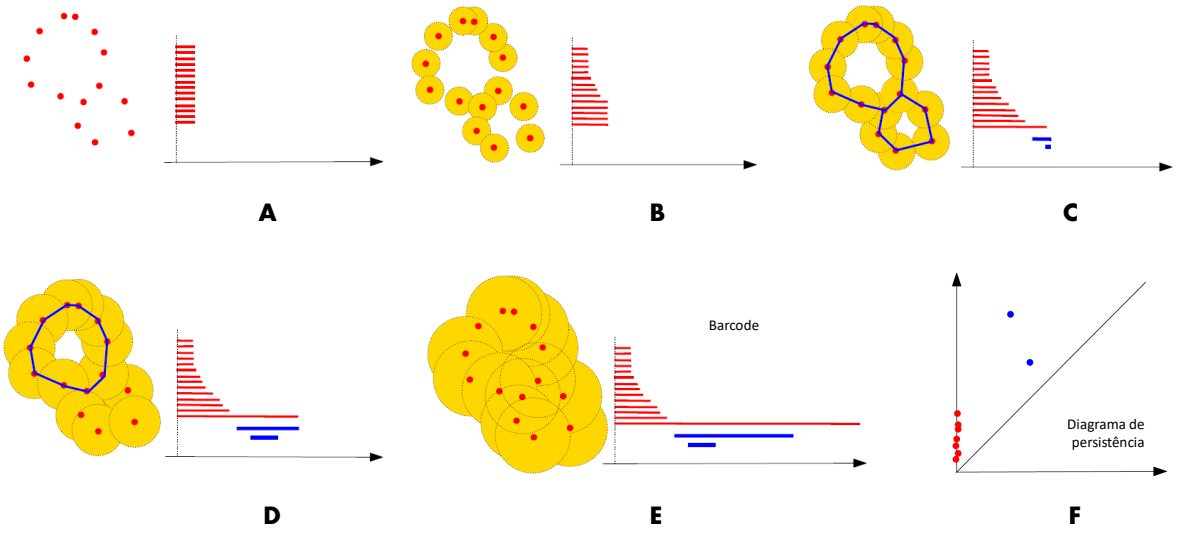


Figura 14 – Procedimento do cálculo da homologia persistente sobre uma nuvem de pontos bidimensional: (A) O cálculo da homologia persistente parte de uma nuvem de pontos, *point cloud*, no plano, no entanto esse procedimento pode ser realizado a qualquer dimensão n dos dados, e um gráfico com barras (*barcode*), com o número de barras inicial igual ao número de pontos ou número de componentes conectados inicial. (B) Um raio $r \in \mathbb{R}$ cresce de modo constante em torno de cada componente conectado, e a mesma medida as barras crescem no gráfico. Quando componentes conectados fundem-se devido a intersecção de bolas, é dito que eles morrem e nascem outros componentes maiores. Nesse momento algumas barras no *barcode* param de crescer (dos componentes mortos) e outras continuam a crescer (dos componentes que persistem). (C) A medida em que componentes fundem-se, há o surgimento das características topológicas furos, ou *loops* bidimensionais. (D) Essas novas características também vão sendo registrados em novas barras no gráfico, até fundirem-se, e portanto "morrerem". (E) Quando o raio r cresce suficiente para só haver um componente conectado, o procedimento termina e o *barcode* é encerrado. A partir dele propriedades topológicas podem ser contadas. No *barcode*, o comprimento das barras representam o tempo de persistência das característica topológicas pelos seus respectivos tempos de vida. (F) Um chamado diagrama de persistência pode ser gerado a partir do *barcode* para dar uma nova perspectiva na interpretação das características topológicas. Ele é uma gráfico 2D do primeiro quadrante com uma diagonal na posição da função identidade trivial. Nele, a abscissa x_i representa o momento do "nascimento" de uma característica e a ordenada y_i representa o momento da "morte". Característica topológicas mais distantes da diagonal são as que tiveram o maior tempo de vida. Na ilustração, há o surgimento de dois furos unidimensionais de tamanhos diferentes, representados por duas barras azuis de diferentes comprimento no *barcode* e por dois pontos azuis a diferentes distâncias da diagonal no diagrama de persistência. Adaptado de Chazal e Michel (2021).

Homologia e Homologia Persistente

Definição 2.3.15. Homologia é um método ou técnica usada para determinar invariantes topológicos em formas, como componentes conectados e furos. Homologia Persistente é a homologia capturada pela filtração de formas, por complexos simplicials construídos em um espaço métrico ao qual a forma está embutida. A evolução da filtração quantifica os invariantes topológicos da forma pelos números de Betti e os associa um tempo de vida.

A definição acima caracteriza o papel da homologia como uma ferramenta para capturar características topológicas de conjuntos e a homologia persistente como uma extensão que considera a variação dessas características ao longo de uma filtração.

Para exemplificar o cálculo de homologia persistente para um objeto topológico com números de Betti conhecidos, revisitemos o toro como caso ilustrativo. A Figura 15 apresenta a computação da homologia persistente para uma amostra de pontos independente e identicamente distribuída (iid) de um toro. O emprego das ferramentas visuais, como *barcode* e diagrama de persistência, oferecem uma representação detalhada da topologia intrínseca do objeto em questão, sendo possível observar como a homologia persistente capturou seus invariantes topológicos através dos números de Betti.

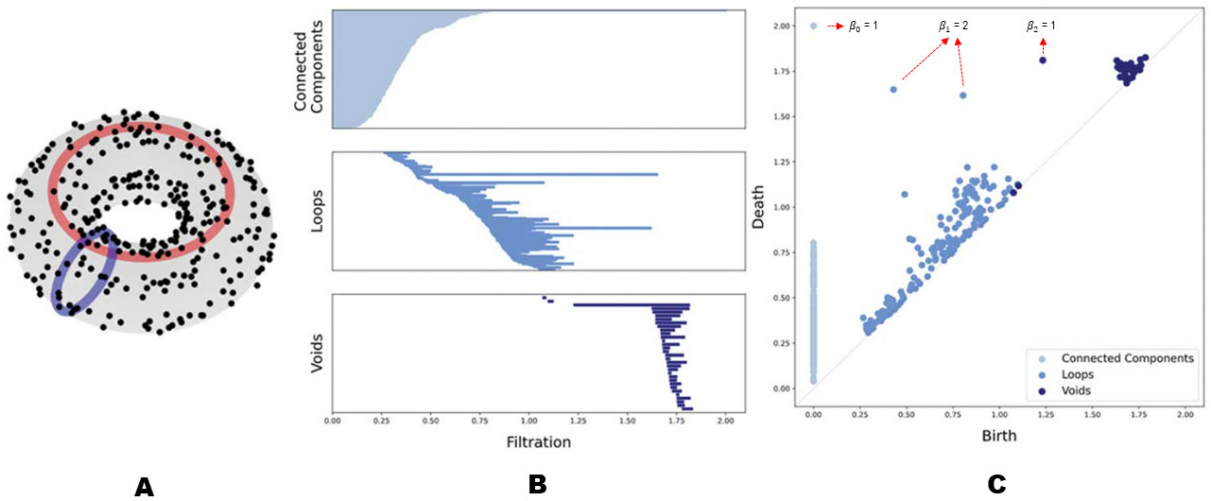


Figura 15 – Homologia persistente do toro. (A) Representação do toro reconstruído como uma nuvem de pontos iid. (B) *Barcodes* resultantes da computação da homologia, destacando o dos componentes conectados na parte superior, onde uma barra mais prolongada sobre a linha superior do gráfico indica $\beta_0 = 1$. No centro, o *barcode* dos *loops*, evidenciando $\beta_1 = 2$ com duas barras proeminentes em relação às demais. O *barcode* inferior representa os vazios, com apenas uma barra se sobressaindo em tamanho, indicando $\beta_2 = 1$. (C) Representação da homologia persistente por meio de um diagrama de persistência. As setas vermelhas identificam os números de Betti do toro, exibindo-os como pontos com o maior tempo de vida no do diagrama. Adaptado de Lazar e Ryu (2021).

A partir dos exemplos, percebe-se que Diagramas de Persistência oferecem uma representação concisa e informativa da evolução topológica dos conjuntos nivelados por uma função em um espaço topológico. Ele permite compreender a persistência de certas características ao longo de diferentes escalas e fornece uma ferramenta valiosa para comparação e classificação de espaços topológicos, uma vez que exibe as características topológicas intrínsecas dos dados.

2.3.2 Mapper

O Mapper é outra ferramenta amplamente empregada em TDA. Sua abordagem é apoiada na representação de um conjunto de dados, independentemente de sua dimensão original, por um grafo unidimensional ou, em algumas aplicações, tridimensional. Tal grafo é capaz de proporcionar uma representação estruturada que permite inferir informações topológicas significativas sobre o conjunto de dados considerado.

Fundamentalmente, na construção do Mapper, conjuntos de dados são mapeados para grafos de *Reeb* — apoiado na teoria de Morse, trata-se de um objeto matemático que reflete a evolução de conjuntos de níveis de uma função a valores reais em uma variedade (REEB, 1946) — usando o Teorema do Nervo da Topologia Algébrica com o objetivo de simplificar informação. A Figura 16 exibe uma representação de um toro por um grafo de Reeb, em vermelho.

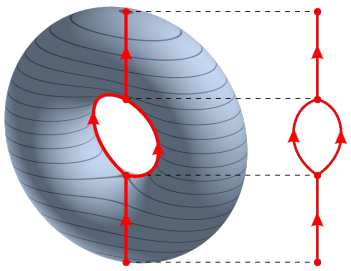


Figura 16 – À direita grafo de Reeb (em vermelho) do toro. Fonte: Wikipedia contributors (2024).

O Teorema do Nervo fornece uma maneira de construir um complexo simplicial que captura as relações de conectividade entre conjuntos abertos em uma variedade topológica, que é um espaço topológico que localmente se assemelha a um espaço euclidiano. O Mapper, apoiado no Teorema do Nervo, muitas vezes utiliza o complexo de Čech representando os conjuntos abertos através de interseções de bolas centrados nos dados, para construir o grafo. Para entender os aspectos matemáticos da construção do grafo Mapper alguns requisitos preliminares se fazem necessários.

Função de Homotopia

Definição 2.3.16. Dadas duas funções contínuas $f, g : X \to Y$ entre dois espaços topológicos, uma função de homotopia entre f e g é uma função contínua $H : X \times [0, 1] \to Y$

tal que $H(x,0) = f(x)$ e $H(x,1) = g(x)$ para todo $x \in X$. Isso indica a existência de uma deformação contínua entre f e g.

A Definição 2.3.16 introduz o conceito de uma função H que descreve uma deformação contínua entre duas funções f e g, deste modo f e g são ditas homotópicas. Em termos mais simples, duas funções são homotópicas se podem se "transformar suavemente" uma na outra, mantendo extremidades fixas durante a transformação.

Equivalência Homotópica

Definição 2.3.17. Duas variedades topológicas X e Y são homotopicamente equivalentes se existem funções contínuas $f : X \to Y$ e $g : Y \to X$ tal que as composições $g \circ f$ e $f \circ g$ são homotópicas à identidade em X e Y, respectivamente. Em outras palavras, existem funções contínuas $F : X \times [0,1] \to X$ e $G : Y \times [0,1] \to Y$ satisfazendo:

1. $F(x,0) = x$ para todo $x \in X$ e $F(x,1) = g(f(x))$ para todo $x \in X$,
2. $G(y,0) = y$ para todo $y \in Y$ e $G(y,1) = f(g(y))$ para todo $y \in Y$.

A Definição 2.3.17 emprega o conceito de funções de homotopia para estabelecer a equivalência homotópica entre duas variedades topológicas. A equivalência homotópica entre espaços topológicos é importante porque conhecer se um conjunto é homotopicamente equivalente a outro significa dizer que eles compartilhamento muitas informações topológicas, a saber a mesma homologia, o que contribui para abordagens computacionais na manipulação e interpretações de estruturas complexas.

Variedade Contrátil

Definição 2.3.18. Uma variedade topológica $\mathbb{X}$ é contrátil se existe uma função de homotopia $F : \mathbb{X} \times [0,1] \to \mathbb{X}$ tal que $F(x,0) = x$ para todo $x \in \mathbb{X}$ e $F(x,1)$ é constante para algum ponto fixo $x_0 \in \mathbb{X}$.

Em outros termos, a definição 2.3.18 diz que um espaço topológico $\mathbb{X}$ pode ser representado por um único ponto se ele é equivalente homotopicamente a tal ponto (LIMA, 2003, p. 11). Exemplos de variedade contráteis são bolas e conjuntos convexos do $\mathbb{R}^n$.

Teorema do Nervo sobre Variedades Contráteis e Equivalência Homotópica

Teorema 2.3.1 (Teorema do Nervo)**.** Seja $\mathbb{X}$ uma variedade topológica e $\mathcal{U}$ uma cobertura aberta de $\mathbb{X}$. Suponha que, para todo subconjunto finito $\mathcal{V} \subseteq \mathcal{U}$, a interseção $\bigcap_{V \in \mathcal{V}} V$ seja não vazia e contrátil. Então, a união dos conjuntos em $\mathcal{U}$, denotada por $N(\mathcal{U}) = \bigcup_{U \in \mathcal{U}} U$, é também uma variedade contrátil e é homotopicamente equivalente a $\mathbb{X}$.

O Teorema do Nervo estabelece que um conjunto $\mathbb{X}$ pode ser representado de maneira compacta por meio de pontos. Essa representação é alcançada ao associar cada subdivisão, representada por um elemento U_i da cobertura de $\mathbb{X}$, a um vértice no complexo de Čech $C(\mathcal{U})$, o qual é, por sua vez, representado pelo nervo $\mathcal{U}$, denotado por $N(\mathcal{U})$. Essa abordagem possibilita a representação de uma variedade $\mathbb{X} \subseteq \mathbb{R}^n$ por um simples grafo unidimensional, que por sua vez pode ser o grafo Mapper. Apoiado na ideia central de fornecer um sumário unidimensional para um conjunto de dados $\mathbb{X} \subseteq \mathbb{R}^d$ $(d \geq 1)$ e com os conceitos necessários à mão, o Algoritmo Mapper pode ser representado de modo ilustrativo, ver Figura 17.

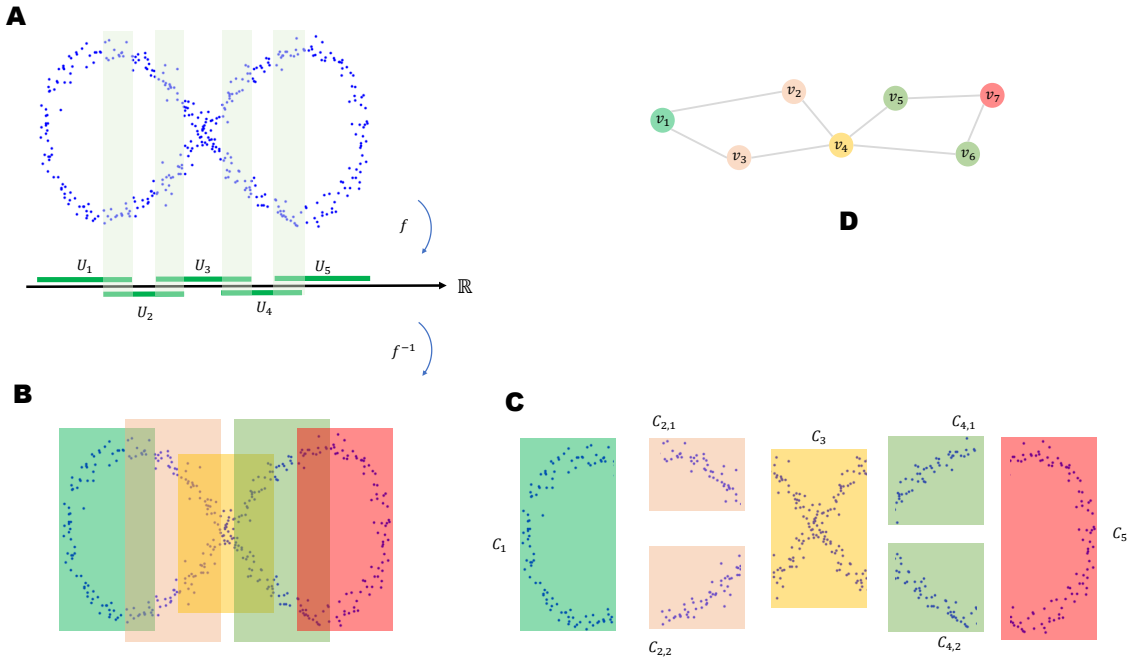


Figura 17 – Ilustração do Algoritmo Mapper sobre uma lemniscata formada por 400 pontos. (A) A lemniscata é um conjunto $\mathbb{X} \subset \mathbb{R}^2$. Uma função de filtro f é usada para representar $\mathbb{X}$ em $\mathbb{R}$ pelos intervalos $U_1, ..., U_4$. (B) Pela função f^{-1} (função de filtro inversa) o conjunto $\mathbb{X}$ é reconstruídos por intervalos sobrepostos U_is. (C) Os intervalos são divididos em clusters C_is contráteis a um ponto. (D) Como os clusters são contráteis a um ponto eles podem ser representados por vértices v_is interligados entre os clusters que compartilham características, ou seja, cuja intersecção é não vazia, formando um grafo abstrato que representa àqueles 400 pontos do conjunto inicial com apenas 7 vértices e 8 arestas. Configurações adicionais podem ser estabelecidas para melhorar a interpretação das informações. Por exemplo a cor dos vértices pode ser feita por uma característica de interesse e o tamanho dos vértices podem ser proporcionais ao números de informações (dados) que eles representam. As duas características topológicas principais dos dados foram capturadas pelo grafo, os dois furos unidimensionais. Fonte: Autor, 2023.

Para obter o grafo representado na Figura 17D por meio da abordagem do Mapper, diversos passos são necessários, sendo crucial que o usuário os manipule adequadamente para garantir uma representação topológica fiel dos dados. A escolha de parâmetros e etapas do processo requer uma consideração cuidadosa, uma vez que influenciam diretamente a qualidade do resumo topológico obtido. A representação por grafos é especialmente atrativa para expressar a topologia dos dados, uma vez que os grafos não possuem uma forma intrínseca. Essa característica harmoniza-se bem com os princípios da topologia, em que um objeto topológico pode ser deformado em outro, contanto que suas propriedades topológicas fundamentais sejam preservadas. No exemplo apresentado, ambos os *loops* foram preservados, ilustrando a capacidade do Mapper em capturar características topológicas essenciais dos dados.

No exemplo da *lemniscata* (Figura 17A), o conhecimento prévio da forma dos dados facilitou a interpretação do Mapper. Contudo, as manipulações dos parâmetros do Mapper podem impactar significativamente no grafo final podendo levar a interpretações inconsistentes caso não haja conhecimento prévio dos dados. A escolha cuidadosa desses parâmetros demanda esforço e cautela, pois, como dito, grafos não possuem forma intrínseca. Isso nos diz que, caso o interesse seja capturar a forma dos dados pelo grafo, o algoritmo pode proporcionar uma representação gráfica divergente da forma original dos dados, mesmo que o resumo topológico esteja consistente para os parâmetros configurados.

Pensando em oferecer um rigoroso descritor para a forma dos dados, outra ferramenta topológica inspirada no Mapper foi proposta. Desenvolvida por Dłotko (2019), o Algoritmo Ball Mapper (BM), surge como uma alternativa na análise exploratória de dados através de grafos.

A principal diferença na construção do BM em relação ao Mapper está no modo como a cobertura dos dados é construída. Enquanto que no Mapper a cobertura é feita por funções de filtro, o BM usa a Definição 2.3.6 de bola conjuntamente com a Definição 2.3.7 de cobertura para recobrir o conjunto de dados considerado. Isso faz com que o BM precise apenas de um único parâmetro de configuração, que seria o raio da bola.

Deste modo, a ideia central dessa nova abordagem é uma constante $\varepsilon > 0$ usada para encontrar uma coleção de pontos $C \subset \mathbb{X}$ tal que a coleção de bolas $B(C) = \bigcup_{x \in C} B(x, \varepsilon)$ cubra $\mathbb{X}$ completamente. Com isso uma rede é formada de modo que cada bola seja agora um cluster representado por um nervo contrátil. Assim, cada bola representa um vértice que são conectados se possuírem compartilhamento de dados. Assim como no Mapper, seus vértices podem configurados quanto a coloração e tamanho.

Com esse princípio de criação, é possível perceber que o BM depende exclusivamente da densidade dos dados, e isso o torna robusto na captura da forma dos dados ao mesmo

tempo que fornece o sumário topológico. Como o BM será o algoritmo adotado para a aplicação nas imagens dos otólitos nesta tese, seu esboço juntamente com os fundamentos matemáticos serão deixados para o capítulo da metodologia, onde será ilustrado sobre uma amostra dos próprios dados das imagens usadas neste estudo.

Em TDA, Mapper e BM têm como objetivo fornecer resumos topológicos de conjuntos de dados complexos, e suas diferenças se dão sobre como definem e conectam regiões no espaço de baixa dimensão, logo suas particularidades podem ser sintetizadas em termos de abordagem e flexibilidade, do seguinte modo:

- O Algoritmo Mapper
    - Abordagem: O algoritmo Mapper mapeia o conjunto de dados para um espaço de baixa dimensão usando funções de projeção e, em seguida, divide esse espaço em regiões sobrepostas chamadas "bins" (caixas). Cada bin é associado a um grupo de pontos no conjunto de dados original. Construção do Grafo: Conecta-se bins adjacentes no espaço de baixa dimensão, formando um grafo que captura a topologia do conjunto de dados.
    - Flexibilidade: Pode ser adaptado para diferentes tipos de dados e métricas.
- O Algoritmo Ball Mapper
    - Abordagem: O Ball Mapper também reduz a dimensionalidade, mas, em vez de bins, utiliza "bolas" e a noção de vizinhança para cobrir o conjunto de dados original. Cada bola cobre um conjunto de pontos no espaço de alta dimensão. Construção do Grafo: A construção do grafo ocorre de maneira semelhante ao Mapper, conectando bolas que se intersectam, mas com o uso de vizinhança local, o que pode levar a representações mais detalhada da forma e das características topológicas dos dados.
    - Flexibilidade: Pode ser particularmente eficaz quando há agrupamentos ou estruturas em escalas diferentes no conjunto de dados, pois ele é escalável.

Quanto a opção por qual das técnicas pode ser mais adequada, pode-se dizer que esta escolha está pautada em subjetividade, uma vez que tal escolha pode depender da natureza específica do conjunto de dados e dos padrões topológicos que se deseja capturar. No caso desta tese, a escolha pelo BM está alicerçada em duas questões: A primeira é que, em um dos estudos feitos, variações de densidade estrutural do otólito, pode ser mais interessante se avaliada na perspectiva da forma dos otólitos. Como a forma dos dados, otólitos, é conhecida, este estudo também pode servir como teste verificativo do sumário

topológico produzido pelo BM, ou seja, se a saída do BM gera de fato um grafo com uma representação similar à forma do otólito, como proposto por tal abordagem. A segunda é devido ao fato de o BM precisar de apenas um parâmetro, o que gera economia de tempo nas análises quando se tem diversos conjuntos com diferentes tamanhos a serem analisados.

2.3.3 Aplicações a Imagens Digitais

O estudo que abriu as aplicações da Análise Topológica em imagens foi o mesmo que introduziu a técnica mapper no início dos anos 2000's. Dirigida por Gurnar Carlsson e apoiada pela empresa *Ayasdi Inc.*[1], a pesquisa intitulada *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition* (SINGH; MEMOLI; CARLSSON, 2007) propôs o algoritmo mapper como método computacional capaz de descrever e extrair importante informações de dados com alta dimensão.

As imagens as quais Singh, Memoli e Carlsson (2007) usaram para aplicar o algoritmo consistiu em uma imagem de um conjunto de dados de diabetes (ANDREWS; HERZBERG, 1985) reconstruída com a técnica *pursuit projection* (HUBER, 1985), na aplicação para dados reais, e imagens tridimensionais de cavalos, cabeça humana e elefante em aplicações de teste e validação do algoritmo.

Uma significativa aplicação do mapper se deu em imagens de ressonância magnética funcional (*functional magnetic resonance imaging*, *fMRI* do inglês), permitindo uma perspectiva visual da dinâmica organizacional do cérebro humano. Nesse estudo, Saggar et al. (2018) mapearam o cérebro de pessoas, por imagens de *fMRI*, enquanto elas realizavam diversas atividades, desde estado de repouso, assistir vídeos, a resolver problemas matemáticos. A ideia foi observar a atividade neural para entender a organização dinâmica do cérebro durante diferentes atividades humanas. A Figura 18 traz, à esquerda, uma sequência de imagens 3D obtidas por *fMRI* do cérebro de um participante da pesquisa, ao centro o grafo mapper construído a partir dos dados das imagens, e à direita, um destaque da região de maior concentração de informações no grafo, revelando conexões durante as transições entre tarefas. Essa aplicação nos revela a contribuição da TDA para mapear a atividade cerebral mediante mudanças no fluxo sanguíneo associadas à atividade neural.

A TDA é uma ferramenta flexível que pode ser aplicada a diversos tipos de dados topológicos, incluindo dados provenientes de imagens digitais em uma variedade de contextos. A exemplo, na biologia Li et al. (2018) aplicaram TDA, em particular a Homologia Persistente, como uma abordagem morfométrica para caracterizar e demarcar um espaço morfológico de folhas. Nesse estudo, o uso da TDA foi capaz de permitir uma

[1] Empresa de tecnologia com foco em TDA e *Machine Learning*. Fundada em 2008 por Gunnar Carlsson, Gurjeet Singh, e Harlan Sexton.

descrição robusta e rica das características morfológicas, ajudando a entender a variação e a organização das formas das folhas aos organizá-las em espaços morfológicos.

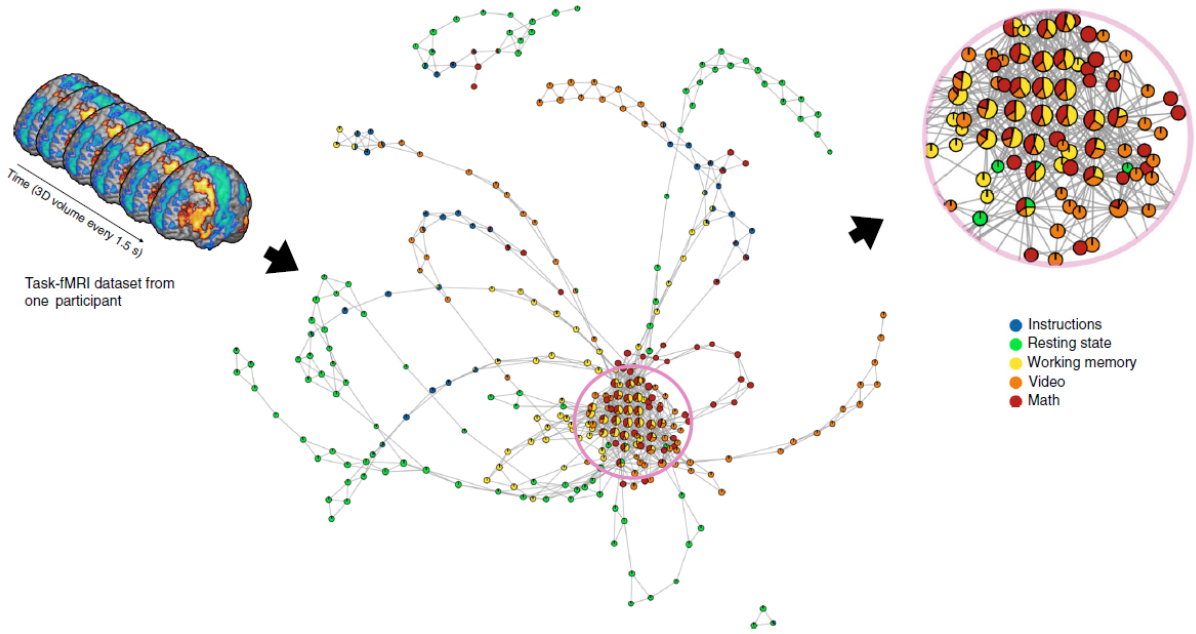


Figura 18 – Sequência de imagens obtidas por ressonância magnética funcional (à esquerda). Grafo mapper construído sobre as imagens (ao centro). Destaque para a região mais conectada do grafo (à direita). Adaptado de Saggar et al. (2018).

Aplicações em imagens usando Homologia se deram no trabalho de Li, Ovsjanikov e Chazal (2014), onde os autores investigaram como a homologia persistente pode ser empregada para o reconhecimento de estruturas específicas em imagens, utilizando-a como uma ferramenta para extrair informações topológicas e identificar padrões importantes. Já o trabalho de Singh et al. (2014) usou homologia persistente como descritor topológico de arranjos celulares no tecido tumoral capturados através de imagens histológicas 2D, avaliando seu desempenho na tarefa de classificar subtipos de câncer de mama.

Percebe-se que, imagens digitais oferecem um verdadeiro palco para as técnicas de TDA, e com isso suas aplicações podem atuar em outras áreas extra neurociência cognitiva e pesquisa clínica, em potencial a biologia. Isso porque outras técnicas de aquisição imagem, além de *fMRI*, têm permitido muitas aplicações de TDA em biologia. Uma técnica de destaque são imagens obtidas por tomografia computadorizada de raios X, a qual aplicações são encontradas nos trabalhos de Li et al. (2017), Chitwood et al. (2019) e Amézquita et al. (2020).

Na análise de imagens tridimensionais, a contribuição do artigo de Carrière, Oudot e Ovsjanikov (2015) aborda a criação de assinaturas topológicas para pontos em formas 3D, com ênfase na estabilidade dessas assinaturas para garantir robustez nas análises topológicas. Isso é valioso, especialmente em aplicações práticas onde os dados podem ser suscetíveis a pequenas perturbações. O trabalho *"Persistence-based pooling for shape pose recognition"*, atribuído a Bonis et al. (2016), aborda questões relacionadas ao reconhecimento de pose em formas, usando técnicas de *pooling*[2] baseadas em persistência. Essa é uma aplicação útil de TDA, uma vez que a integração de conceitos de persistência e *Machine Learning* podem melhorar a capacidade de reconhecimento em aplicações relacionadas a formas tridimensionais.

O trabalho de Chazal et al. (2009) aborda a representação e comparação estável de formas geométricas em situações de perturbações ou deformações. Os autores destacam características topológicas persistentes em várias escalas, oferecendo uma abordagem robusta para analisar formas complexas. Essa proposta revela-se promissora na análise de imagens com ruído, uma vez que busca representar e comparar formas de maneira estável sob perturbações e deformações, fornecendo uma perspectiva valiosa para a análise de formas em condições desafiadoras.

Outro trabalho que fortalece a ideia de aplicar TDA em imagens digitais 3D, sem a preocupações da possibilidade de ruídos, é o artigo de Reininghaus et al. (2015). Nele os autores combinam TDA com *Machine Learning* para classificar objetos, onde o foco principal é a aplicação de técnicas topológicas no contexto do aprendizado de máquina, proporcionando robustez e capacidade de generalização aos modelos em face de perturbações nos dados.

Outros trabalhos que aplicam TDA em imagens para identificação de padrões e classificação de objetos podem ser encontrados na literatura. Boa parte deles enfatizam o uso de persistência na análise estrutural para fornecer uma representação robusta e informativa das características topológicas presentes nos dados, o que pode ser útil em diversas aplicações, como reconhecimento de objetos, processamento de imagem, entre outros.

[2] *Pooling* refere-se a uma técnica utilizada em aprendizado de máquina e visão computacional para reduzir a dimensionalidade ou agregar informações de uma determinada região de interesse. No contexto específico do reconhecimento de forma e pose em 3D, ela é aplicado a características topológicas persistentes, contribuindo para uma representação eficaz no processo de reconhecimento de pose.

3 Objetivos

3.1 Objetivo Geral

Extrair, reconhecer, e identificar características de densidade e de forma na estrutura tridimensional de otólitos através da análise topológica de imagens 3D de tomografia computadorizada.

3.2 Objetivos Específicos

- Reduzir tempo e custo computacional sobre o processamento e análise de dados de alta dimensão usando amostragem probabilística;
- Descrever variações da densidade óssea tridimensional de otólitos usando grafos e invariantes topológicos;
- Propor um novo e eficiente classificador/descritor para otólitos com base em sua estrutura 3D, aplicando ferramentas da Análise de Dados Topológica, Homologia Persistente e métricas topológicas, combinadas a um modelo de *Machine Learning*;
- Avaliar correlações entre variáveis do peixe, idade, comprimento e densidade óssea, e características topológicas entropias de persistência;
- Ampliar o conhecimento de métodos biométricos aplicados a otólitos de peixes marinhos, visando enriquecer a compreensão de estudos biológicos e ecológicos.

Os objetivos dessa tese visam fortalecer os estudos biológicos possibilitando a compreensão detalhada da densidade otolítica de diversos peixes marinhos e avaliando um novo classificador para a forma do otólito, podendo seus resultados contribuírem diretamente para estratégias eficazes de conservação e gestão sustentável dos recursos pesqueiros. Além disso, a aplicação prática dessas descobertas serve como indicador valioso em estudos ecológicos, fornecendo um entendimento mais profundo das interações entre os peixes marinhos e seus ambientes, particularmente diante de desafios ambientais contemporâneos, como as mudanças climáticas.

4 Material e Métodos

4.1 Amostra e aquisição das imagens

A amostra é composta por 21 otólitos *sagittae* (imagens 3D). A distribuição dos indivíduos por espécies pode ser vista na Tabela 1.

Tabela 1 – Amostra de estudo: espécies e quantidade de indivíduos por espécie

Espécie	
Opisthonema ogrinum	2
Acanthocybium solandri	3
Thunnus albacares	1
Thunnus obesus	5
Acanthurus coeruleus	8
Haemulum plumierii	2
Total	21

A opção por uma amostra com indivíduos de espécies distintas e em diferentes idades, se deu pelo fato do estudo ser para fins de comparação e comprovação. Sabendo que as diferenças ecológicas, como habitats distintos e hábitos alimentares, definem os otólitos, possibilitando estudos de composição, estrutura, classificação e identificação, uma variedade taxonômica da amostra pode ser interessante para generalização dos resultados.

As imagens de μCT foram feitas no Laboratório de Tomografia Computadorizada de Raios X do Departamento de Energia Nuclear da UFPE - Universidade Federal de Pernambuco, Brasil. Os otólitos *sagittae* foram colocados em um suporte de isopor e escaneados por um scanner modelo XT H 225 ST da Nikon Metrology, com parâmetros definidos em 80 kV, 220 μA e filtro de alumínio de 0,5 mm. As imagens pertencem a uma coleção própria obtidas com auxílio do projeto de pesquisa AIMMO - Analysis of 3D IMages and growth Modelling of Otoliths project (Processo CNPq: 457387/2014-9 e FACEPE: APQ-0178-108/2014), dedicado a promover estudos e pesquisas sobre modelagem matemática, estatística e computacional de fenômenos biológicos sobre otólitos. Na Figura 19 estão 4 imagens 3D originais para quatro das espécies do estudo.

No processo de aquisição, o aparelho de μCT produziu Z fatias (imagens 2D) transversais para cada otólito, as quais reconstruíram as imagens 3D, com informações internas e externas de densidade ajustadas para a escala de unidade *Hounsfield* (HU).

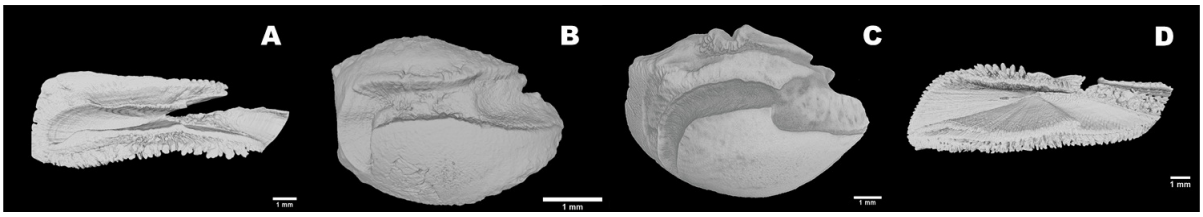


Figura 19 – Em sequência, algumas imagens originais obtidas pelo aparelho de μCT das espécies (A) *Acanthocybium solandri*, (B) *Acanthurus coeruleus*, (C) *Haemulon plumierii*, (D) *Thunnus obesus*.

Essa escala HU transforma o coeficiente de atenuação linear de uma material em um valor adimensional, definido como radiodensidade do material (BUZUG, 2011). Ela foi preferida por ser capaz de quantificar os tons de cinza característicos do imageamento por raios X, possibilitando reconhecer informações diferentes de densidade por uma maior diferenciação de cores. Para extrair os voxels (coordenadas X, Y, Z) com suas respectivas HUs das imagens, um *script* em linguagem *R* foi usado (Apêndice A). As imagens 3D de cada otólito foram reconstruídas obtendo-se os voxels a partir das imagens como uma nuvem de pontos 3D na forma de uma matriz de dados com 4 colunas X, Y, Z e HU, e portanto de dimensão igual ao número de voxels vezes 4. Deste modo cada voxel (X, Y, Z) define a localização espacial de um valor de radiodensidade HU na vizinhança[1] de cada ponto (X, Y, Z) do otólito.

4.2 Ferramentas da Análise Topológica de Dados

4.2.1 O Algoritmo *Ball Mapper*

O funcionamento do algoritmo Ball Mapper (BM) parte de uma nuvem de pontos X e uma constante positiva $\varepsilon > 0$. Em seguida, é selecionado um subconjunto de pontos (marcos) $\mathrm{N} \subset X$, de modo que para cada $x \in X$, exista $\mathrm{n} \in \mathrm{N}$ tal que $d(x, \mathrm{n}) \leq \varepsilon$ (N é chamado $\varepsilon - net$). Como $X \in \bigcup_{\mathrm{n}\in\mathrm{N}} B(\mathrm{n}, \varepsilon)$, todo X é coberto por bolas $B(\mathrm{n}, \varepsilon)$. Um grafo abstrato é gerado, em que vértices correspondem as bolas $B(\mathrm{n}, \varepsilon)$ cujos tamanhos são proporcionais a quantidade de dados (pontos) contidos por elas, e as arestas entre vértices correspondem a interseções não vazias de bolas. Geralmente, os vértices são coloridos de acordo com um atributo dos dados, neste estudo a radiodensidade HU. Em resumo, o algoritmo BM exibe um grafo de uma estrutura n-dimensional. As etapas do algoritmo é ilustrada, sobre uma foto 2D ou slice Z de uma amostra com 1% dos dados, extraída usando Strat de um dos otólitos da pesquisa, na Figura 20.

[1] A matemática por traz dessa palavra, Definição 2.3.4, nos diz que a densidade média a uma certa proximidade de um voxel é igual a própria densidade no voxel.

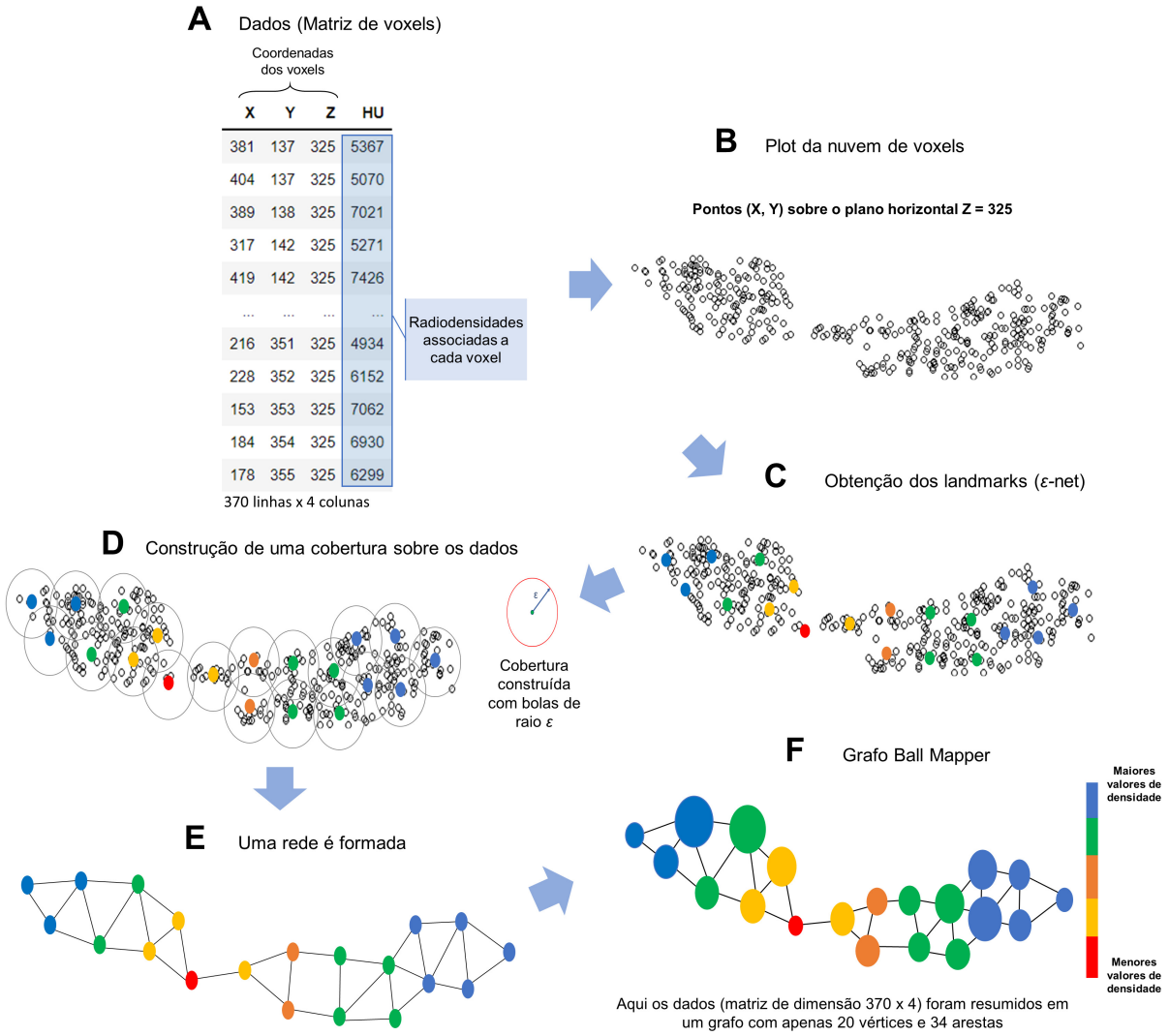


Figura 20 – (A) Matriz de entrada X, fatia $Z = 325$ extraído de uma amostra a 1% de um otólito de *Thunnus obesus*. Ela possui 370 voxels aos quais estão associados valores de radiodensidade em Unidades Hounsfield (HU). (B) Plot dos pixels da matriz de dados. (C) Definindo o valor do parâmetro $\varepsilon > 0$, um subconjunto de pontos (marcos ou landmarks) é selecionado e uma rede chamada $\varepsilon - net$ é construída. (D) Todo o conjunto de dados X é coberto por bolas de raio ε (bola vermelha à direita) centradas em cada marco. (E) Uma rede ou grafo abstrato é gerado, cujos vértices correspondem às bolas e as arestas correspondem às interseções não vazias entre elas. (F) Grafo Ball Mapper de X obtido com raio ε, a partir do qual propriedades topológicas podem ser inferidas. O tamanho dos vértices é ponderado de acordo com o número de voxels que suas respectivas bolas encapsulam, eles podem ser coloridos de acordo com uma variável de interesse nos dados, no caso desse estudo a variável densidade HU. Uma paleta de cores, aqui com apenas 5 cores, ajuda a interpretar ilustradamente as variações de densidade nessa fatia (imagem 2D) do otólito através do grafo. Fonte: Autor, 2022.

4.2.2 Invariantes Topológicos sobre grafos

Para mensurar o grau de similaridade entre amostras a 5% e Pop quanto à topologia de rede dos grafos que as representam, avaliou-se alguns *Invariantes Topológicos* sobre tais grafos. Em teoria dos grafos e ciência de redes, os *Invariantes Topológicos* (ITs) descrevem propriedades estruturais de um grafo que não mudam sob isomorfismos de grafos, ou seja, propriedades são estruturais que permanecem inalteradas independentemente de como o grafo é desenhado. Eles descrevem aspectos fundamentais da conectividade da rede. Avaliar a topologia da rede que constitui um grafo é importante para conhecer a forma como as arestas e vértices estão dispostos e se relacionam entre si. Os ITs adotados nesse trabalho para comparação entre grafos foram escolhidas por fornecerem informações gerais da topologia a partir da estrutura global das redes que compõem os grafos, são eles:

Agrupamento: Fornece o grau médio de agrupamento de nós no grafo. Em termos locais se trata da densidade de ligações na vizinhança de um nó; ***Grau do nó:*** é a média da conectividade de todos os nós do grafo, ou seja, o número médio de arestas que incide nos nós; ***Assortatividade:*** é o grau de similaridade das conexões no grafo em relação ao grau do nó, ou seja, mede o nível de relação dos nós com outros de mesmo grau; *Comprimento médio do* ***caminho mais curto:*** é o caminho com o menor número de arestas entre dois nós.

A *Eficiência* de um par de nós em um grafo é o inverso multiplicativo da distância do caminho mais curto entre os nós, ela é uma medida da capacidade de um vértice se comunicar eficientemente com outros vértices na rede; ***Eficiência global*** de um grafo é a média da eficiência de todos os pares de nós. ***Eficiência local*** de um nó é a eficiência global média do subgrafo induzido pelos vizinhos do nó, ela mede a capacidade de um vértice se comunicar com seus vizinhos imediatos. ***Densidade do grafo:*** em um grafo não direcionado, é a relação entre o número de arestas presentes no grafo e o número total de arestas que o grafo pode ter para seu conjunto de nós. É definido pela equação D = 2 |E| / |V| (|V| - 1), onde |V| é o número de vértices e |E| é o número de arestas. Tipicamente varia no intervalo [0,1], sendo 0 para um grafo sem arestas e 1 para um grafo completo;

Transitividade: é a fração de todos os triângulos possíveis presentes no grafo; ***Conectividade de aresta:*** é o número mínimo de arestas removidas para eliminar todos os caminhos possíveis entre um par de nós; ***Número de nós do grafo***; ***Número de aresta do grafo***; **Característica de Euler** χ de um grafo G(V,E) conexo é definida por χ = |V|-|E|, onde |V| e |E| denotam o número de vértices e de arestas do grafo respectivamente, é uma invariante importante em quaisquer dados fundamentados em natureza topológica e geométrica (ŁAWNICZAK et al., 2021; FAROOQ et al., 2022).

4.2.3 Homologia Persistente

Para a aplicação da homologia persistente nos dados dos otólitos, empregou-se o pacote computacional `giotto-tda`[2], desenvolvido na linguagem de programação `Python`. Esta seção expõe os métodos da homologia adotados para o problema de classificação, iniciando pela definição de espaço vetorial, base para as demais. Optou-se por manter a mesma simbologia da documentação do pacote a fim de evitar excessos.

Espaço Vetorial sobre o Corpo $\mathbb{R}$

Definição 4.2.1. *Dado o corpo dos números reais $\mathbb{R}$, um conjunto V é chamado de espaço vetorial sobre o corpo $\mathbb{R}$ se as seguintes condições são satisfeitas:*

1. *Adição Vetorial: Existe uma operação chamada adição vetorial, denotada por $+$, que associa a cada par de vetores $u, v \in V$ um vetor $u + v \in V$. Essa operação deve satisfazer as seguintes propriedades:*

   a) *Comutatividade: Para todos $u, v \in V$, $u + v = v + u$.*

   b) *Associatividade: Para todos $u, v, w \in V$, $(u + v) + w = u + (v + w)$.*

   c) *Existência do Elemento Nulo: Existe um vetor chamado vetor nulo, denotado por 0, tal que para todo $u \in V$, $u + 0 = u$.*

   d) *Existência do Inverso Aditivo: Para cada $u \in V$, existe um vetor chamado inverso aditivo de u, denotado por $-u$, tal que $u + (-u) = 0$.*

2. *Multiplicação por Escalar: Existe uma operação chamada multiplicação por escalar, denotada por $\cdot$ ou simplesmente concatenação, que associa a cada vetor $u \in V$ e cada escalar $\alpha \in \mathbb{R}$ um vetor $\alpha \cdot u \in V$. Essa operação deve satisfazer as seguintes propriedades:*

   a) *Compatibilidade com a Multiplicação do Corpo: Para todos $u \in V$ e $\alpha, \beta \in \mathbb{R}$, $\alpha \cdot (\beta \cdot u) = (\alpha \cdot \beta) \cdot u$.*

   b) *Identidade Multiplicativa: Para todo $u \in V$, $1 \cdot u = u$, onde 1 é o elemento neutro da multiplicação no corpo $\mathbb{R}$.*

   c) *Distributividade em Relação à Adição Vetorial: Para todos $u, v \in V$ e $\alpha \in \mathbb{R}$, $\alpha \cdot (u + v) = \alpha \cdot u + \alpha \cdot v$.*

   d) *Distributividade em Relação à Adição Escalar: Para todos $u \in V$ e $\alpha, \beta \in \mathbb{R}$, $(\alpha + \beta) \cdot u = \alpha \cdot u + \beta \cdot u$.*

[2] Biblioteca de código aberto e comunidade ativa que integra TDA e ML (TAUZIN et al., 2021). Diponível em: https://giotto-ai.github.io/gtda-docs/latest/index.html

Módulo de Persistência

Definição 4.2.2. *É um conjunto de espaços vetoriais $V(s)$, $s \in \mathbb{R}$, sobre um campo $\mathbb{K}$. Cada $V(s)$ está associado a uma escala s durante uma filtragem topológica. Além disso, são definidos mapas lineares $f_{st} : V(s) \to V(t)$ para $s \leq t$ em $\mathbb{K}$. Esses mapas, conhecidos como "mapas estruturais", satisfazem a seguinte propriedade: se $r \leq s \leq t$, então $f_{rt} = f_{st} \circ f_{rs}$. Notavelmente, a composição de mapas estruturais segue a ordem transitiva das escalas. A condição de que todos, exceto um número finito de mapas estruturais, são isomorfismos, implica que a estrutura topológica persiste ao longo da filtragem, com características que nascem e morrem em tempos específicos.*

Diagrama de Persistência

Definição 4.2.3. *Um diagrama de persistência é um multiconjunto multiescala de pontos definido em $\mathbb{R} \times (\mathbb{R} \cup \{+\infty\})$. Ele associa um módulo de persistência com a seguinte condição: para cada par s, t, o número contado com multiplicidade de pontos (b, d) no multiconjunto, satisfazendo $b \leq s \leq t < d$, é igual à classificação de f_{st}. Uma propriedade importante é que existe um isomorfismo entre dois módulos de persistência se e somente se seus diagramas de persistência forem iguais (ZOMORODIAN; CARLSSON, 2005).*

Espaço normado

Definição 4.2.4. *É um espaço vetorial V munido da função*

$$|| - || : V \to \mathbb{R}$$

restrita aos valores não negativos de $|| - ||$, e que para cada u, $v \in V$ e um $a \in \mathbb{R}$ tem-se

$$||u|| = 0 \Leftrightarrow u = 0$$

$$||au|| = |a| \ ||u||$$

$$||u + v|| = ||u|| + ||v||$$

A função $|| - ||$ pode ser referida como uma norma do espaço vetorial V, que naturalmente define uma espaço métrico com

$$d(u, v) = ||u - v||$$

como função distância.

Norma L^p

Definição 4.2.5. *Uma norma L^p é um espaço normado de funções p-integráveis*

$$f \mapsto \left(\int_U |f(x)|^p dx \right)^{1/p}$$

com atribuição $||-||_p$. Uma função f é dita p-integrável se

$$\int_U |f(x)|^p dx$$

é finito, sendo f definida em $C(U, \mathbb{R})$ (conjunto das funções contínuas para valores reais em U tal que $U \subseteq \mathbb{R}^n$).
Quando $p = 2$ tem-se a única norma L^p que induz o produto interno

$$\langle f, g \rangle = \left(\int_U |f(x) - g(x)|^2 dx \right)^{1/2}$$

Imagem de Persistência

Definição 4.2.6. *A Imagem de persistência refere-se à distância L^p entre diagramas de persistência com suavização gaussiana calculada a partir dos eixos de nascimento (ADAMS et al., 2017).*

Distâncias de Wasserstein e Bottleneck

Definição 4.2.7. *Kerber, Morozov e Nigmetov (2017) define a distância de p-Wasserstein como uma distância ou métrica entre dois diagramas de persistência D_1 e D_2 calculada pelo o ínfimo sobre todas as bijeções $\gamma : D_1 \cup \Delta \to D_2 \cup \Delta$ de*

$$\left(\sum_{x \in D_1 \cup \Delta} ||x - \gamma(x)||_\infty^p \right)^{1/p},$$

onde $||-||_{+\infty}$ é definido para $(x, y) \in \mathbb{R}^2$ por $\max\{|x|, |y|\}$ e Δ é um multiconjunto $\{(s, s) | s \in \mathbb{R}\}$ com multiplicidade $(s, s) \mapsto +\infty$. O limite $p \to +\infty$ define a distância de Bottleneck. Mais explicitamente, é o ínfimo sobre o mesmo conjunto de bijeções do valor

$$\sup_{x \in D_1 \cup \Delta} ||x - \gamma(x)||_\infty$$

Traduzindo para termos de uma distância δ entre D_1 e D_2, seria

$$\delta_\infty(D_1, D_2) = \inf_\gamma \sup_{x \in D_1} ||x - \gamma(x)||_\infty.$$

Pensando na sobreposição de D_1 e D_2, essa última equação nos diz o quanto se deveria ajustar para que D_2 e D_1 sejam iguais.

Um conjunto de diagramas de persistência conjuntamente com qualquer noção de distância torna-se um espaço métrico.

Persistência landscape

Definição 4.2.8. *A persistência landscape de um diagrama de persistência* $\{(b_i, d_i)\}_{i\in I}$ *é o conjunto* $\{\lambda_k\}_{k\in\mathbb{N}}$ *de funções*

$$\lambda_k : \mathbb{R} \to \overline{\mathbb{R}},$$

sendo $\lambda_k(t)$ *o k-ésimo maior valor do conjunto* $\{\Lambda_i(t)\}_{i\in I}$ *em que*

$$\Lambda_i(t) = [\min\{t - b_i, t - d_i\}]_+$$

e $[c]_+ := \max(c, 0)$. *A função* λ_k *é denominada como camada k da persistência landscape. A construção da persistência landscape estabelece uma vetorização dos diagramas de persistência tendo o espaço vetorial de funções reais definido no* $\mathbb{N} \times \mathbb{R}$ *como seu contradomínio. Para qualquer* $p = 1, 2, .., \infty$ *podemos considerar diagramas de persistência D cuja persistência landscape* λ *associada é p-integrável, ou seja,*

$$||\lambda||_p = \left(\sum_{i\in\mathbb{N}} ||\lambda_i||_p^p\right)^{1/p},$$

onde

$$||\lambda_i||_p = \left(\int_{\mathbb{R}} |\lambda_i^p(x)|dx\right)^{1/p}$$

é finito.
A equação norma $p-$ *landscape para* $p = 2$, *emprestada da norma* L^p *quando* $p = 2$, *define o valor kernel landscape em dois diagramas de persistência como*

$$\langle\lambda, \mu\rangle = \left(\sum_{i\in\mathbb{N}} \int_{\mathbb{R}} |\lambda_i(x) - \mu_i(x)|^2 dx\right)^{1/2},$$

onde λ *e* μ *são as persistências landscape associadas a eles (BUBENIK et al., 2015).*

Silhueta ponderada

Definição 4.2.9. *Sendo* $D = \{(b_i, d_i)\}_{i\in I}$ *um diagrama de persistência e* $w = \{w_i\}_{i\in I}$ *um conjunto de reais positivos, a Silhueta de D ponderada por w é um função* $\phi : \mathbb{R} \to \mathbb{R}$ *dada por*

$$\phi(t) = \frac{\sum_{i\in I} w_i\Lambda_i(t)}{\sum_{i\in I} w_i},$$

onde

$$\Lambda_i(t) = [\min\{t - b_i, t - d_i\}]_+$$

e $[c]_+ := \max(c, 0)$. *Para o caso onde* $w_i = |d_i - b_i|^p$ *para* $0 < p \leq \infty$ *a* ϕ *é referida como poder p da silhueta ponderada de D. O processo de construir uma Silhueta define uma vetorização do conjunto dos Ds com alvo no espaço vetorial de funções contínuas a valores reais (CHAZAL et al., 2014).*

Vetorizações Heat

Definição 4.2.10. *Os pontos de um diagrama de persistência como suporte dos* `deltas de Dirac`*, torna possível construir, para qualquer $t > 0$, duas vetorizações do conjunto que compõe o diagrama de persistência para o conjunto das funções contínuas a valores reais definidas no quadrante $\mathbb{R}^2_{>0}$. Desta forma, a vetorização "heat" de simetria é construída para cada diagrama de persistência D resolvendo a equação do calor*

$$\begin{aligned} \Delta_x(u) = \partial_t u \quad & \textit{definida em} \quad \Omega \times \mathbb{R}_{>0} \\ u = 0 \quad & \textit{definida em} \quad \{x_1 = x_2\} \times \mathbb{R}_{\geq 0} \\ u = \sum_{p \in D} \delta_p \quad & \textit{definida em} \quad \Omega \times 0, \end{aligned}$$

onde $\Omega = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \leq x_2\}$, e após isso resolvendo essa mesma equação, para a mesma escolha de t, sob a mudança $(x_1, x_2) \mapsto (x_2, x_1)$, a imagem de D fica então definida pelas diferença entre as duas soluções (REININGHAUS et al., 2015; ADAMS et al., 2017).

A solução da equação do calor com a condição inicial dada pelos `deltas de Dirac`, definida para $p \in \mathbb{R}^2$, é

$$\frac{1}{4\pi t} \exp\left(-\frac{\|p - x\|^2}{4t}\right),$$

onde mudança de variável $\sigma = \sqrt{2t}$ permite conexão com variáveis aleatórias normalmente distribuídas.

Curva de *Betti*

Definição 4.2.11. *Partindo de um diagrama de persistência D, a Curva de Betti é a função*

$$\beta_D : \mathbb{R} \to \mathbb{N},$$

a qual os valores do domínio $s \in \mathbb{R}$ são os números de pontos em $D = \{(b_i, d_i)\}$, contados em cada dimensão de homologia $H_{0,1,2,\ldots}$, sob a restrição $b_i \leq s < d_i$.

Entropia de Persistência

Definição 4.2.12. *No contexto da Homologia Persistente, em um diagrama de persistência $D = \{(b_i, d_i)\}_{i \in I}$, com $d_i < +\infty$, a entropia de seus pontos em cada gerador de persistência $H_{0,1,2,\ldots}$, é medida por*

$$E(D) = -\sum_{i \in I} p_i \log p_i,$$

onde $p_i = (d_i - b_i)/L_D$ e $L_D = \sum_i (d_i - b_i)$ (RUCCO et al., 2016).

Considerando as três primeiras dimensões de homologia $H_{0,1,2}$, a Entropia de Persistência é uma forma de representar nuvens de pontos (objetos topológicos) por coordenadas tridimensionais, uma para cada dimensão de homologia, podendo assim fornecer gráficos de dispersão ao qual cada objeto é representado por um único ponto 3D. Tais gráficos são chamados de *matrizes de características*, os quais permitem fazer comparações, identificação e classificação entre objetos ou conjuntos de dados.

Outro modo de visualizar a matriz de característica com o objetivo de auxiliar na interpretação das informações, é a normalização da *entropia de persistência*. Segundo Myers, Munch e Khasawneh (2019) ela é estabelecida ao dividir a Entropia de Persistência (Definição 4.2.12) pelo logaritmo natural da soma dos tempos de vida de todos os pontos do diagrama de persistência.

Entropia de Persistência Normalizada

Definição 4.2.13. *Em um diagrama de persistência* $D = \{(b_i, d_i)\}_{i \in I}$*, a normalização da entropia E(D) é dada por*

$$E'(D) = \frac{E(D)}{log(L(D))}.$$

Random Forest

Definição 4.2.14. *Introduzido por Breiman (2001), Random Forest é uma ferramenta baseada em árvores de decisão para aprendizado de conjunto, aleatoriedade criteriosa, capaz de produzir modelos preditivos incrivelmente precisos sobre um banco de dados, visando uma compreensão minuciosa. Assim, pode ser simplesmente definido como uma combinação de árvores de predição onde cada árvore depende dos valores de um vetor iid para todas as árvores da floresta, a fim de melhorar a precisão em tarefas de classificação. Ainda conta com a propriedade de contornar o overfitting do conjunto de treinamento além da capacidade de ajuste e validação ao mesmo tempo em que é treinado, sendo assim eficiente em tarefas de classificação (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).*

OOB score

Definição 4.2.15. *Out of bag (OOB) score é um modo de avaliação e validação do modelo de aprendizado de máquina Random Forest usado como uma ferramenta interna associada ao conjunto de dados (BREIMAN, 2001). Ao treinar um modelo de Random Forest, cada árvore de decisão é treinada em uma amostra aleatória (obtida via bootstrap) dos dados, uma parte do conjunto de dados não é considerada no treinamento, tal parte é denominada como "fora da bolsa" ou "out of bag" do inglês. Mais precisamente o OOB score mede o desempenho do modelo sobre o conjunto de dados que não foram incluídos, ou seja, àqueles "out of bag".*

4.3 Sistemática da Análise

Com o objetivo de reduzir o custo computacional, necessário para aplicar os métodos BM e HP da TDA nos dados brutos das imagens 3D de μCT, foi explorado um caminho alternativo utilizando amostragem probabilística para reduzir a resolução das imagens, a fim de verificar se imagens reduzidas forneceriam resultados de densidade similares aos dos dados brutos, e se seria possível utilizar o mesmo procedimento para classificar otólitos com base na sua estrutura 3D. Um diagrama ilustrativo da sistemática dos métodos é apresentado na Figura 21, e a descrição desses procedimentos segue em sequência.

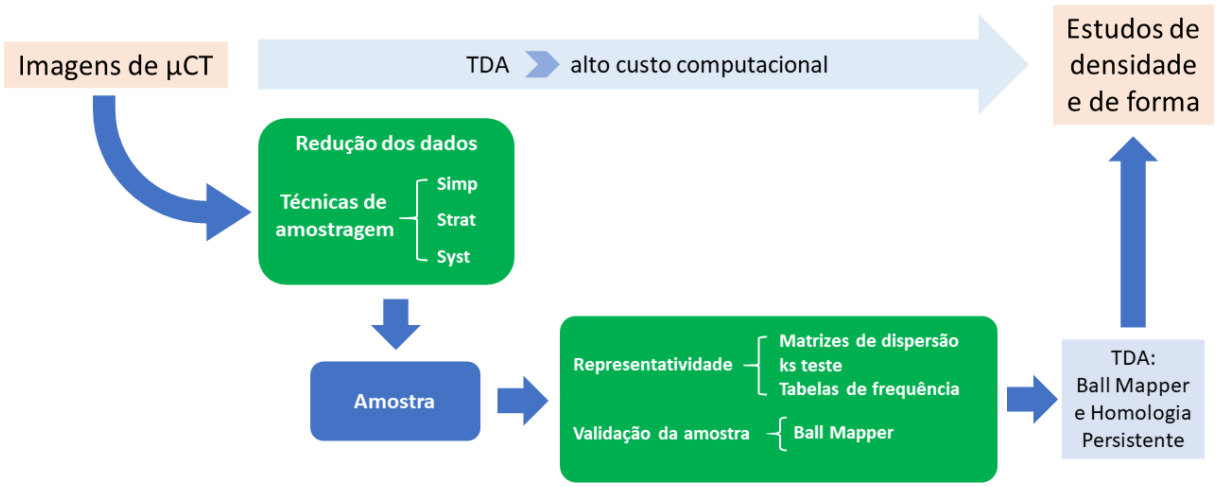


Figura 21 – Sistemática dos métodos. Uma vez que aplicar TDA diretamente nos dados brutos das imagens requer alto custo computacional, foi explorado um caminho alternativo para estudar padrões de densidade e de forma nos otólitos. Tal caminho consiste em reduzir os dados utilizando amostragem probabilística, validar as amostras e realizar o estudo com TDA sobre as amostras reduzidas.

A primeira etapa consiste em reduzir os dados brutos (voxels) das imagens (considerados como "População", Pop, para abreviar) utilizando os três procedimentos de amostragem probabilísticas mais conhecidos, Amostragem aleatória simples (Simp), Amostragem aleatória estratificada proporcional (Strat) e Amostragem aleatória sistemática (Syst), todas sem reposição, a fim de se obter uma amostra (Samp) reduzida de cada imagem dos otólitos. Essas técnicas foram escolhidas porque desempenham um papel fundamental na redução de dados, ao passo que estatísticas amostrais podem sumarizar informações através da criação de um conjunto de dados gerenciável, a fim de modelar e resumir estatísticas populacionais, com objetivo de economia de custos computacionais quando a população é grande (KARTHIK; ABHISHEK, 2019, pp. 80; 123).

Definindo n como tamanho de Samp (número de voxels na Samp) e N como a quantidade total de voxels na Pop, na Simp cada voxel é selecionado aleatoriamente da população dos N voxels disponíveis em cada imagem, até atingir o tamanho n da Samp. Aqui cada elemento tem igual probabilidade $\frac{n}{N}$ de ser sorteado, o que define essa técnica de amostragem como *equiprobabilística*, portanto, qualquer uma das $\binom{N}{n}$ combinações pode ser a amostra sorteada (COCHRAN, 1977). A Strat teve os slices ou coordenadas Z como variável de estratificação. Nessa técnica a Samp de tamanho n é composta amostrando o mesmo percentual ou proporção de voxels n/N de cada slice Z por Simp (KALTON, 1983; KISH, 1995). Na Syst, caso particular de amostra estratificada (SILVA, 1998), os voxels foram ordenados pela coordenada Z e cada Pop dividida em n partes de comprimentos N/n, onde um voxel foi amostrado de cada uma dessas partes, até compor a Samp, sistematicamente segundo a expressão $[k + (i - 1)N/n]$, onde k é uma constante escolhida por Simp do conjunto $\{1, 2, ..., N/n\}$ e $i = 1, 2, 3, ..., n$ (KALTON, 1983; KISH, 1995).

A representatividade das amostras (Samps) extraídas foi verificada por análise estatística descritiva e não paramétrica. A fim de perceber que as amostras sorteadas possuem as mesmas distribuições populacionais, matrizes de dispersão permitiram observar o comportamento das distribuições amostrais sobre as coordenadas dos voxels X, Y, Z e da HU. Como critério quantitativo, uma vez que a distribuição de HU é não conhecida, o teste de *Kolmogorov-smirnov* (KOLMOGOROV, 1933; SMIRNOV, 1939) com significância estatística em 5% foi usado para verificar diferenças significativas entre as distribuições de HU, comparação entre Pop e Samps. Estatísticas descritivas básicas (*média, desvio padrão (sd), erro padrão (se) e coeficiente de variação (cv)*) calculados sobre HU permitiram comparar parâmetros amostrais com seus valores populacionais correspondentes, além de estudar a variabilidade da média de HU para as amostras. Tabelas de frequência ajudaram a perceber a representatividade dos voxels por cada slice Z (grupo de fotos 2D) entre Pop e Samps.

Como a distribuição dos dados de HU é não normal, um procedimento metodológico, utilizando o *BM* foi estabelecido para ajudar a determinar o tamanho n de amostra, resolução mínima, capaz de entregar informações de HU semelhantes às da Pop. Com tamanho amostral estabelecido, grafos BM de Pop e Samps reduzidas em 95%, pelos diferentes sistemas de amostragem, foram comparados como avaliação qualitativa das variações de densidade HU. Para comparação quantitativa entre grafos de Pop e amostras, Invariantes Topológicos (ITs) sobre grafos (descritas na Subseção 4.2.2) foram calculadas sobre suas respectivas redes. Em otólitos em que o BM detectou sujeiras e/ou pedaços quebrados, algoritmos de tratamento e filtragem de dados ajudaram a eliminar tais impurezas afim de não comprometer as análises topológicas.

Quanto a expectativa da TDA, especificamente Homologia Persistente, como classificador para a forma do otólito com base na estrutura 3D, foi utilizado a divisão de homologia do pacote *giotto-tda* (TAUZIN et al., 2021) para calcular a homologia persistente das imagens dos otólitos. Para visualizar qualitativamente a separação dos otólitos foram utilizadas matrizes de características sobre as entropias de persistência normalizadas e não normalizadas. As entropias de persistência foram calculadas sobre as dimensões de homologia H_0, H_1 e H_2 a partir de imagens de 5 mil voxels, extraídas dos 21 otólitos adotados neste estudo, usando Simp. Para mensurar quantitativamente a classificação entre os otólitos, as entropias alimentaram o modelo de *machine learning*, *Random Forest* (BREIMAN, 2001), do pacote *scikit-learn* (PEDREGOSA et al., 2011), acoplado a um *OOB score* (BREIMAN, 2001). Depois, características topológicas (descritas na Subseção 4.2.3 - Invariantes Topológicos sobre a Homologia Persistente) foram incorporadas na alimentação do modelo junto às entropias, afim de melhorar a acurácia da classificação ampliando as possibilidades de se ter mais suposições sobre as interpretações dos resultados. Os diagramas na Figura 22 resumem as análises desempenhas por parte do BM e da HP.

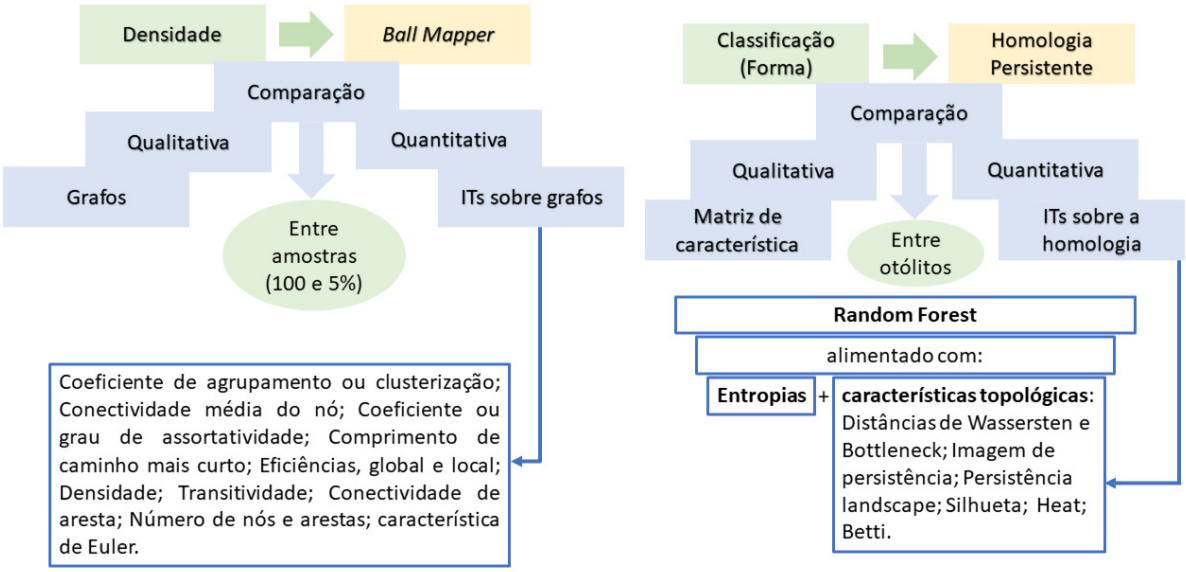


Figura 22 – Sistemática das análises. Descrição das análises sobre o estudo da radiodensidade HU usando Ball Mapper (à esquerda) e do estudo de classificação pela forma dos otólitos usando Homologia Persistente (à direita).

A escolha pela Simp, para a classificação pela forma dos otólitos, se deu porque ela não tem restrições de seleção da amostra, ou seja, ela detém mais aleatoriedade que as outras técnicas exploradas nesse trabalho, o que é interessante para preservar a estrutura 3D do otólito. Como este estudo não envolve uma variável de interesse, ou seja, apenas os voxels X, Y e Z, essa amostra é dispensada da avaliação da representatividade, sendo que sua validação e creditação é dada pela qualidade dos resultados da classificação.

O tamanho de amostra em 5 mil voxels escolhidos por Simp para cada otólito, se deu de modo empírico, pois como os cálculos de homologia requerem grande capacidade computacional, maior até que àquela exigida para os estudos de densidade com o BM, foram realizados testes para verificar com quantos voxels a capacidade computacional disponível a este estudo conseguiria processar. O primeiro teste foi feito sobre uma Samp a 5% do AS1, menor otólito, cerca de 489.388 voxels. Baixando a quantidade em cada 100 mil voxels, observou-se que o dispositivo adotado conseguiria processar em torno de 100 mil voxels, então retirou-se 5 mil voxels de cada um dos 21 otólitos, totalizando 105 mil voxels, e o dispositivo conseguiu computar tal quantidade.

Após suposições feitas sobre o comportamento das entropias, foi verificado seus relacionamentos com as variáveis do peixe idade, comprimento e radiodensidade (HU) usando Análise de Regressão Linear Simples com um nível de significância estatística de 5%. Como medida pré-diagnóstica, a análise se iniciou por uma matriz de dispersão e correlação utilizando o coeficiente de correlação de Spearman (ρ) (SPEARMAN, 1961), uma vez que não é conhecido o tipo de relação entre as variáveis, e a linha *spline*, que é usada para modelar relações complexas entre variáveis que não são facilmente capturadas por regressão linear simples, facilitando a visualização de padrões gerais na relação entre as variáveis. Os níveis de correlação adotados foram: muito fraca, se $0 \leq |\rho| \leq 0,19$; fraca, se $0,20 \leq |\rho| \leq 0,39$; moderada se $0,40 \leq |\rho| \leq 0,59$; forte, se $0,60 \leq |\rho| \leq 0,79$; muito forte, se $0,80 \leq |\rho| \leq 1$ (ANASTASI, 1976).

Foram testados os modelos linear $\mathrm{Y} = \beta_0 + \beta_1 * \mathrm{X}$ e exponencial linearizado $\mathrm{Y} = \beta_0 * exp(\beta_1 * \mathrm{X}) \Leftrightarrow ln\mathrm{Y} = ln\beta_0 + \beta_1 * \mathrm{X}$, permitindo comparações com o linear no âmbito da análise de regressão desempenhada. A escolha por esses modelos se deve a primeira tentativa de entender a natureza desses relacionamentos, se linear ou exponencial. O R^2-ajustado e o Critério de Informação de Akaike - AIC ajudaram na discussão sobre a escolha dos modelos que melhor estimam as variáveis pelas entropias, e em cima dos considerados melhores modelos foi avaliada a análise gráfica da regressão aliada a avaliação dos resíduos.

Todo o estudo foi realizado em um laptop *gaming* com sistema operacional Windows 11, processador Intel(R) Core(TM) i7-10750H, CPU 2.60GHz, e memória RAM de 32GB.

5 Resultados e Discussões

A Tabela 2 amplia as informações da amostra dos 21 otólitos (Tabela 1) utilizada nesta tese, fornecendo informações dos indivíduos, espécie, nomenclatura, idade e comprimento, e das imagens 3D, quantidade de fotos 2D por imagem 3D e valores de média e desvio padrão para a densidade HU dos otólitos. Como primeiro resultado, as duas últimas colunas contêm informações de Samps após uma redução de 95% dos voxels por Strat, acompanhada de seus respectivos valores de média e desvio padrão para HU (coluna 9), as quais são estimativas próximas às suas respectivas estatísticas populacionais (coluna 7).

A opção por uma amostra rica quanto a variedade de idades e espécies, se apoia na generalização científica dos procedimentos metodológicos aplicados na solução do problema de tese, que é estudar a radiodensidade otolítica e propor um novo modo de classificar otólitos pela forma, usando topologia. Pode-se dizer que a amostra adotada ajuda a creditar mais confiabilidade a demonstração e validação deste estudo. Além disso, durante as discussões dos procedimentos experimentais podem surgir *insights* de associação das informações encontradas, como as invariantes topológicas, às características dos otólitos, tais como a radiodensidade, idade, espécie, etc, ampliando as discussões e percepções que possam sugerir estudos adicionais e futuros que sirvam para quaisquer outras espécies.

A amostra de 21 otólitos adotada pode parecer pequena, porém como cada imagem é um *big data* de voxels, imagem de alta resolução, é requerido uma capacidade computacional significativa para realização deste estudo. As discussões e demonstrações serão feitas sobre o menor otólito da amostra, o AS1. O fato dele ser o otólito de menor imagem (menos voxels) dentre todos, faz dele o ideal para demonstrar a validade da redução da dimensionalidade proposta nesta tese para os dados de imagens. Quando necessário comparações, serão expostos resultados de otólitos de outras espécies.

A discussão na próxima seção se preocupa com a redução da incerteza acerca do uso da amostragem probabilística na redução de dados de imagens como procedimento válido para a redução de custo computacional ao aplicar métodos da TDA para se estudar imagens digitais. Deste modo discutiremos qual o tamanho mínimo de resolução das imagem retêm as informações de HU similares às da Pop dos dados.

A demonstração de que é possível estudar radiodensidade HU sobre uma redução de 95% dos dados das imagens, é feita analisando tamanhos de amostra em 1, 5 e 10%, como medida que rotula o tamanho mínimo adequado para tal objetivo.

Tabela 2 – Amostra de estudo. Informações dos peixes (colunas 1 — 4): espécies, organizadas por média de idade, descrição, idade em anos e comprimento em centímetros (cm) (Standard length – SL, Fork length – FL e Total length – TL). Dados das imagens 3D dos otólitos (colunas 5 — 9): Informações para os dados populacionais e de amostras estratificadas com tamanho em 5%: quantidade de imagens 2D (slices ou coordenadas Z) que compõem as imagens 3D (Coluna 5); tamanho populacional e amostral (colunas 6 e 8); valores de média e desvio padrão populacionais (coluna 7) e suas respectivas estimativas em uma Strat (coluna 9) para a variável de interesse, a radiodensidade em unidades hounsfield (HU).

Espécie	Otólito	Idade (anos)	Comprimento (cm)	Quantidade de slices (imagens 2D)	População (POP)		Strat a 5%	
					Número de voxels N	HU $\mu \pm \sigma$	n	$\overline{HU} \pm s$
Opisthonema oglinum	OO1	0,52*	14,8 SL	1.309	56.990.590	8.606,434 ±844,4046	2.849.528	8.606,159 ±843,8825
	OO2	0,68*	18 SL	1.202	52.656.855	8.627,553 ±792,4461	2.632.842	8.627,275 ±791,7629
Acanthocybium solandri	AS1	1,47*	118 TL	1.256	9.787.784	6.866,915 ±901,1399	489.388	6.866,787 ±901,9428
	AS2	2,89*	139,7 TL	1.335	24.744.274	6.743,157 ±840,9455	1.237.213	6.742,644 ±840,9860
	AS3	4,73*	155 TL	1.357	18.149.726	6.839,974 ±823,4242	907.486	6.840,635 ±823,7727
Thunnus albacares	TA	3*	120 FL	1.330	18.305.517	6.682,777 ±979,584	915.109	8.683,088 ±980,3563
Thunnus obesus	TO1	3,22*	105,8 FL	1.202	16.182.138	6.826,865 ±748,8098	809.109	6.827,991 ±746,9690
	TO2	4,49*	127,3 FL	1.334	24.458.013	6.661,356 ±694,4558	1.222.901	6.662,404 ±694,1256
	TO3	4,58*	128,7 FL	1.338	22.943.675	6.648,497 ±746,3680	1.147.184	6.648,578 ±746,8720
	TO4	5,29*	138,7 FL	1.404	31.022.482	6.606,955 ±714,1529	1.551.124	6.606,799 ±715,0979
	TO5	6,85*	157,3 FL	1.378	24.459.710	6.544,137 ±716,8538	1.222.985	6.545,014 ±717,0441
Acanthurus coeruleus	AC1	1	10,9 TL	1.191	74.834.084	8.655,782 ±834,1708	3.741.703	8.656,488 ±833,8575
	AC2	4	22 TL	1.154	115.795.160	7.882,942 ±919,2358	5.789.757	7.882,287 ±919,0292
	AC3	4	24,9 TL	1.347	124.569.189	7.904,421 ±919,6347	6.228.459	7.904,296 ±919,4072
	AC4	4	26,2 TL	1.130	133.765.004	7.802,880 ±936,4313	6.688.449	7.802,924 ±936,2038
	AC5	5	24 TL	1.096	129.537.288	7.876,248 ±924,7067	6.476.864	7.876,462 ±924,6298
	AC6	6	25,8 TL	927	103.168.048	7.846,072 ±1.028,5680	5.158.401	7.846,637 ±1.028,4680
	AC7	8	30,1 TL	1.279	130.376.491	7.662,073 ±899,9138	6.518.824	7.662,395 ±899,8054
	AC8	15	32 TL	973	148.381.738	7.510,870 ±957,4740	7.419.085	7.510,805 ±957,4737
Haemulon plumierii	HP1	8	18,3 TL	1.319	127.207.688	8.240,227 ±886,4483	6.360.383	8.240,037 ±886,9335
	HP2	14	26,6 TL	1.370	159.717.508	8.380,981 ±896,9024	7.985.875	8.381,199 ±896,9980

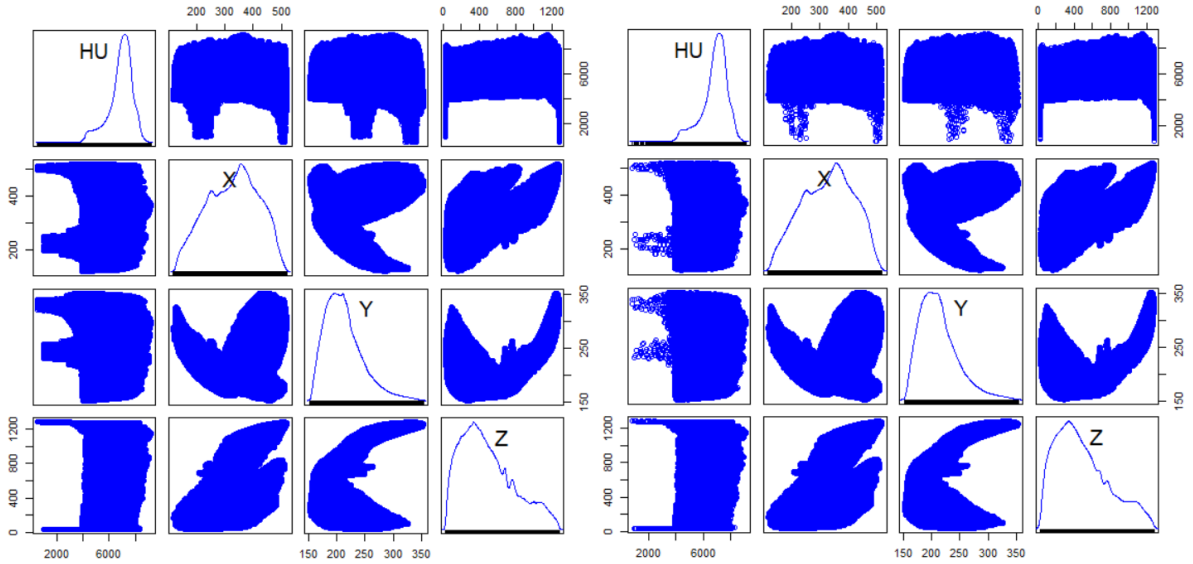
*Idades estimadas com base nas curvas de crescimento dos autores Stéquert, Panfili e Dean (1996), Lessa et al. (2008), McBride, Richardson e Maki (2008) e Duarte-Neto, Higa e Lessa (2012).

5.1 Imagens de otólitos reduzidas por amostragem probabilística

Após a extração de uma amostra de voxels a partir de cada imagem (parcela da resolução da imagem), deve-se verificar se as informações da amostra retirada refletem as informações dos dados populacionais, ou seja, se a amostra é representativa da população. O primeiro passo rumo a representatividade é verificar se as distribuições das variáveis da amostra são iguais as suas correspondentes na população, caso sejam, a Definição 2.3.1 garante a amostra retirada como uma amostra aleatória da população, refletindo o primeiro sinal de representatividade da população através da amostra.

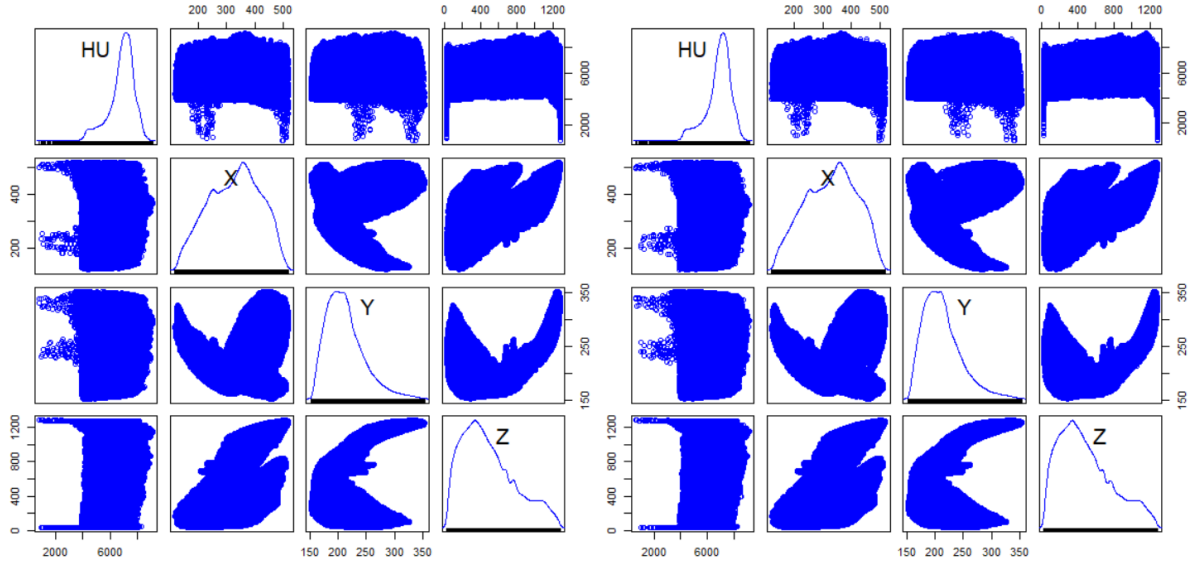
As matrizes de dispersão na Figura 23 fornecem visivelmente as distribuições de probabilidade dos voxels X,Y,Z e HU, para dados do otólito AS1 tanto para os dados populacionais (100% da resolução das imagens – Figura 23(b)) quanto para amostras, a 5% de resolução, extraídas pelos três sistemas de amostragem adotados (Figuras 23(b), 23(c) e 23(d)). Além disso as matrizes também oferecem uma visão em perspectiva 2D do otólito a cada par de eixo, que nada mais é que a distribuição dos voxels, como uma sombra projetada nos planos XY, XZ e YZ.

A partir das matrizes de dispersões da Figura 23 é possível observar qualitativamente a representatividade em distribuição das amostras reduzidas pela amostragem aleatória, e que independente da técnica de amostragem, dentre as adotadas, ao comparar matrizes de dispersão dos dados amostrais, Figuras 23(b), 23(c) e 23(d)), com a matriz de dispersão proveniente dos dados brutos, Figura 23(a), observa-se que as densidades amostrais são similares às suas correspondentes populacionais.



(a) 100% dos dados do AS1

(b) Strat a 5% do AS1



(c) Syst a 5% do AS1

(d) Simp a 5% do AS1

Figura 23 – Matrizes de dispersão: (a) dos dados brutos, ou populacionais (Pop) do AS1 e (b-d) das amostras a 5% proveniente dos diferentes sistemas de amostragem empregados para reduzir os dados, Strat, Syst e Simp respectivamente. Em (a), podemos observar que a HU e as coordenadas dos voxels têm densidades de probabilidade diferentes entre si, indicando que os voxels que representam a imagem de μCT do otólito AS1 é uma amostra aleatória da estrutura do otólito. Nas matrizes de dispersão das amostras reduzidas a 5% em (b), (c) e (d), vemos, independentemente do sistema de amostragem empregado, as mesmas densidades de probabilidade quando comparadas com suas correspondentes na matriz de dispersão dos dados brutos. Isso indica que amostras a 5% retêm informações das distribuições da Pop. A dispersão dos dados permite ainda ter uma visão em perspectiva 2D do otólito AS1 a cada par de coordenadas XY, XZ e YZ.

Independentemente do método de amostragem, amostras com 1% de resolução também retêm a densidades de probabilidade da Pop, ou seja, suas distribuições de probabilidade também não mudam em relação às suas respectivas nos dados brutos. Tal fato pode ser observado no Apêndice B (Figura 41) para dados das imagens do otólito TO3. Mesmo com a representatividade observada para 1% de resolução das imagens, na seção 5.2 a Topologia demonstra que em alguns otólitos da amostra não é apropriado usar resoluções menores que 5% para estudos de variações de radiodensidade otolítica, em HU, e que 5% é adotado aqui por ser a resolução mínima capaz de fornecer informações representativas de HU dos dados brutos para todos os otólitos da amostra do estudo.

Como as distribuições dos voxels de HU não são conhecidas, o teste estatístico não paramétrico de *Kolmogorov-Smirnov* (ks teste), apoiado sobre a hipótese nula de iguais densidades, comprova quantitativamente a representatividade das amostras reduzidas quanto aos dados de HU, comparando Pop e Samp. A Tabela 3 mostra o resultado do ks teste não significativo ($p - valor > 0.05$) para os tamanhos de Strat em 1, 5 e 10% comparados aos dados da Pop em relação à variável HU. Quantitativamente, o teste corrobora as informações fornecidas pelas matrizes de dispersão exibidas na Figura 23, que Pop e Strat possuem mesmas distribuições de probabilidade.

Tabela 3 – Teste de Kolmogorov-Smirnov, bilateral, para duas amostras com 95% de confiança sobre a variável de interesse HU para o otólito AS1. Comparação entre Pop e diferentes tamanhos de Strat.

Comparação entre	Estatística de teste D	$p - valor$
1% e 100%	0.0029121	0.3838
5% e 100%	0.0005964	0.9964
10% e 100%	0.0010816	0.2489

Apesar de se estar buscando evidências para garantir que a resolução mínima ideal para estudos da densidade otolítica é a partir de 5%, isso não deve ser apenas apoiado nesse $p - valor$ mais alto do resultado da Tabela 3. Na verdade, não encontrou-se nenhuma evidência por trás deste $p - valor$ mais alto, o que observou-se é que ele muda a depender da amostra extraída e do otólito. Resultados similares de significância estatística foram observados para o ks teste desempenhado em amostras a 1, 5 e 10%, provenientes dos demais sistemas de amostragem, porém os omitimos para simplificar.

Observada a representatividade por parte da distribuição dos dados, o próximo passo é perceber como se comporta os valores de HU nas amostras reduzidas. A Tabela 4 traz as principais descritivas básicas para a variável de interesse HU em tamanhos de amostra em 1, 5 e 10%, a fim de avaliar e comparar a variabilidade da média de

HU com a média dos dados brutos, Pop. A partir de 5% de resolução, podemos ver estimativas amostrais para HU relativamente próximas às da Pop, permitindo observar a representatividade estatísticas da amostra quanto a variável de interesse, HU.

Tabela 4 – Principais estatísticas descritivas de HU para diferentes tamanhos de Strats. Comparação das estimavas amostrais com seus respectivos valores calculados a partir dos dados da Pop sobre o otólito AS1.
sd – desvio-padrão; *se* – erro-padrão; *cv* – coeficiente de variação.

Estatística sobre HU	AS1/Pop	Amostra a		
		1%	5%	10%
média	6,866.915	6,867.338	6,866.787	6,866.008
sd	901.1399	899.6937	901.9428	901.0005
se (*média*)	0.2880381	2.875757	1.289295	0.9107071
cv	0.1312292	0.1310105	0.1313486	0.1312263

A representatividade do número de voxels ao longo das imagens 2D (fatias ou slices Z) é visualizada usando uma tabela de frequência. Através delas é possível perceber que, dos três sistemas de amostragem aplicados para redução dos dados, Simp foi a que mais perdeu voxels ao longo das fatias Z. Strat e Syst apresentaram proporções semelhantes às da Pop por intervalo de classe. A ocorrência da perda de proporções sobre a Simp também foi constatada para amostras extraídas por este método em outros otólitos. A Tabela 5 traz esse resultado para dados do otólito AS1 e a Tabela 12 no Apêndice C para dados do otólito TO3.

Tabela 5 – Tabelas de frequência mostrando a distribuição agrupada dos voxels ao longo das fatias Zs para Pop, Strat, Syst e Simp em % para o otólito AS1. No geral, todas as amostras de 5% representaram a Pop (proporção final em 100%), o detalhe a ser observado é que, a Simp apresentou alterações nas proporções de voxels ao longo das fatias Zs.

Intervalo de	Pop		Strat		Syst		Simp	
classe dos Zs	voxels	%	voxels	%	voxels	%	voxels	%
0 ⊣ 50	71433	0.73	3574	0.73	3571	0.73	3626	0.74
50 ⊣ 100	398437	4.07	19924	4.07	19922	4.07	19697	4.08
100 ⊣ 150	538683	5.50	26934	5.50	26934	5.50	26847	5.49
150 ⊣ 200	606080	6.19	30308	6.19	30304	6.19	30594	6.25
200 ⊣ 250	654431	6.69	32722	6.69	32722	6.69	32458	6.63
250 ⊣ 300	690210	7.05	34504	7.05	34510	7.05	34363	7.02
300 ⊣ 350	728850	7.45	36443	7.45	36443	7.45	36450	7.45

Tabela 5 – Continuação

Intervalo de	Pop		Strat		Syst		Simp	
classe dos Zs	voxels	%	voxels	%	voxels	%	voxels	%
350 ⊣ 400	710342	7.26	35520	7.26	35517	7.26	35455	7.24
400 ⊣ 450	666199	6.81	33308	6.81	33310	6.81	33366	6.82
450 ⊣ 500	605913	6.19	30297	6.19	30295	6.19	30524	6.24
500 ⊣ 550	554487	5.67	27727	5.67	27725	5.67	27683	5.66
550 ⊣ 600	500995	5.12	25049	5.12	25050	5.12	25244	5.12
600 ⊣ 650	423352	4.33	21167	4.33	21167	4.33	20800	4.25
650 ⊣ 700	407003	4.16	20353	4.16	20350	4.16	20547	4.20
700 ⊣ 750	294364	3.01	14718	3.01	14718	3.01	14728	3.01
750 ⊣ 800	325598	3.33	16280	3.33	16280	3.33	16181	3.31
800 ⊣ 850	239002	2.44	11951	2.44	11950	2.44	12066	2.47
850 ⊣ 900	219648	2.24	10983	2.24	10983	2.24	11079	2.26
900 ⊣ 950	192969	1.97	9651	1.97	9648	1.97	9635	1.97
950 ⊣ 1000	184268	1.88	9216	1.88	9214	1.88	9087	1.86
1000 ⊣ 1050	186208	1.90	9310	1.90	9310	1.90	9201	1.88
1050 ⊣ 1100	186630	1.91	9334	1.91	9332	1.91	9242	1.89
1100 ⊣ 1150	163751	1.67	8187	1.67	8187	1.67	8240	1.68
1150 ⊣ 1200	122500	1.25	6123	1.25	6125	1.25	6227	1.27
1200 ⊣ 1250	87032	0.89	4353	0.89	4352	0.89	4310	0.88
1250 ⊣ 1300	29399	0.30	1470	0.30	1470	0.30	1469	0.30
Total	9787784	100.0	489406	100.0	489387	100.0	489389	100.0

5.2 Validação topológica das imagens reduzidas

Validação Topológica de Amostra — O VTA

O objetivo desse procedimento é decidir sobre a resolução mínima capaz de fornecer informações de radiodensidade HU similares às da população dos dados. Em suma, ele deve indicar o quanto a resolução das imagens 3D pode ser reduzida de modo que as características da variável de interesse, HU, sejam adequadamente preservadas na amostra, avaliando topologicamente distribuições de HU de imagens 2D que compõem o otólito.

O método consiste em pegar algum slice da classe modal, por ser a região com maior número de voxels no otólito e, então, topologizá-la com parâmetro ε do BM variando de 15 a 60 em intervalos de 5. Neste estudo adotou-se o slice central da classe modal. Os

grafos BM resultantes devem capturar as informações de HU da fatia de modo sempre conexo, ou seja, sem apresentar vértices desconectados.

Na Figura 24 temos o resultado desse procedimento para a Strat do AS1, onde para amostras a partir de 5%, as informações de HU são codificadas por grafos conexos. Para a amostra a 1% é possível observar grafos desconexos para os valores dos parâmetros ε em 15, 20 e 35.

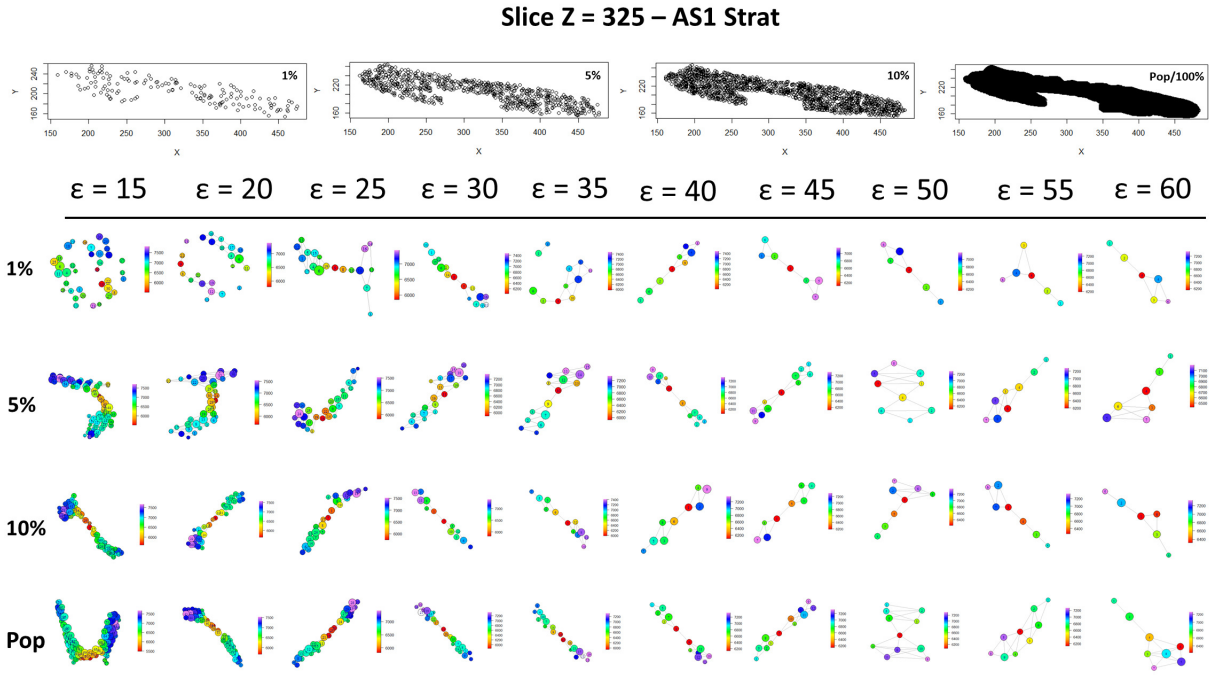


Figura 24 – Procedimento de Validação Topológica da Amosta (VTA) para diferentes tamanhos de Strat a partir da mesma fatia do otólito AS1, espécie *Acanthocybium solandri*. No topo da figura temos os plots da fatia Z=325 (foto 2D) extraída dos tamanhos amostrais 1, 5, 10 e 100%. Abaixo das fatias, grafos calculados sobre àquelas resoluções para o intervalo $[15, 60]$ do parâmetro do Ball Mapper ε. Observa-se variações de densidade semelhantes às da Pop a partir de 5% de resolução. Para valores de parâmetros ε fora do intervalo $[15, 60]$ informações de radiodensidade HU e forma não são codificadas de modo similar às da Pop.

A discussão desse procedimento se torna mais interessante quando feita em otólitos com formas irregulares, na amostra deste estudo os pertencentes a família *scombridae*, *Acanthocybium solandri*, *Thunnus albacares* e *Thunnus obesus* possuem essa particularidade. Nesses otólitos, sua parte mais profunda, região do sulco acústico, estão mais próximas de suas faces internas, provocando uma parte fina do otólito que requer atenção quando se trabalha com imagens reduzidas. Isso porque grandes reduções dessas imagens

podem ocasionar em perdas significativas de voxels dessa região. Essa é uma região de muito interesse no otólito, visto que ela é onde estão concentradas as mais baixas densidades.

O VTA mostra-se como método válido para diminuir a incerteza quanto ao tamanho de amostra adequado para estudo de variações de densidade otolítica com imagens reduzidas, uma vez que as conexões decorrentes do grafo BM aliados as imagens visuais 2D são capazes de indicar se voxels de certas regiões do otólito foram amostrados. Abaixo é descrito o passo a passo do teste VTA feito em um otólito da espécie *Thunnus obesus*.

Passo a passo do procedimento sobre o otólito TO3

Para cada otólito, a partir da resolução que se deseje validar, seleciona-se a fatia central da classe modal da distribuição dos voxels por grupos de Z. No caso do otólito TO3 tal faixa compreende ao intervalo (300; 350), assim, Z = 325 deve ser o slice escolhido para o procedimento. Na Figura 25 está o histograma da distribuição dos voxels na Figura 25A e a fatia Z = 325 colorida pela HU na Figura 25B.

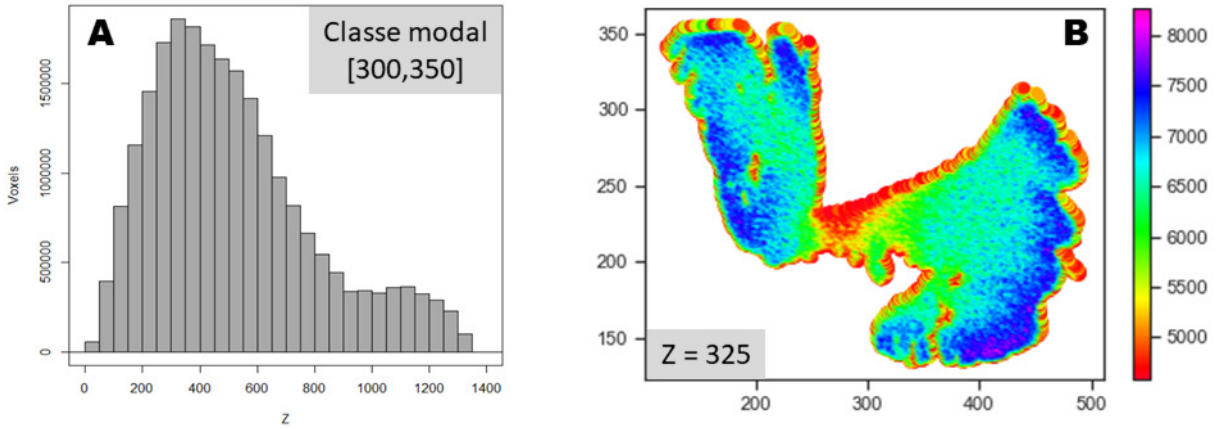


Figura 25 – (A) Distribuição dos voxels por slices Z do otólito TO3. Para o procedimento de VTA colhe-se o slice médio do intervalo de classe com o maior número de voxels por slice, classe modal. Observa-se que, neste caso, o slice escolhido deve ser o Z = 325. Em (B) plot do slice Z=325 colorido pela variável HU, com escala de cores variando do vermelho para valores mais baixos ao roxo representando os valores mais altos. Tal slice foi extraído a partir de 100% de resolução da imagem do TO3.

Nota: Embora o slice Z = 325 escolhido para o otólito TO3 coincida com o mesmo slice escolhido para o otólito AS1 neste estudo, a escolhida do slice para o VTA pode variar a depender do otólito e da espécie, isso pode ser visto no VTA das outras espécies no Apêndice D (Figuras 42 a 46).

A Figura 26 traz o procedimento para o otólito TO3. Ao ajustar o parâmetro ε do BM no intervalo de 15 a 60, observa-se que para 1% dos dados, as informações sobre a forma e densidade da fatia não são capturadas para valores ε de 15, 20 e 25. Como observado para o otólito AS1, para amostras a partir de 5%, a topologia encapsula as informações de densidade e forma similares às informações da população.

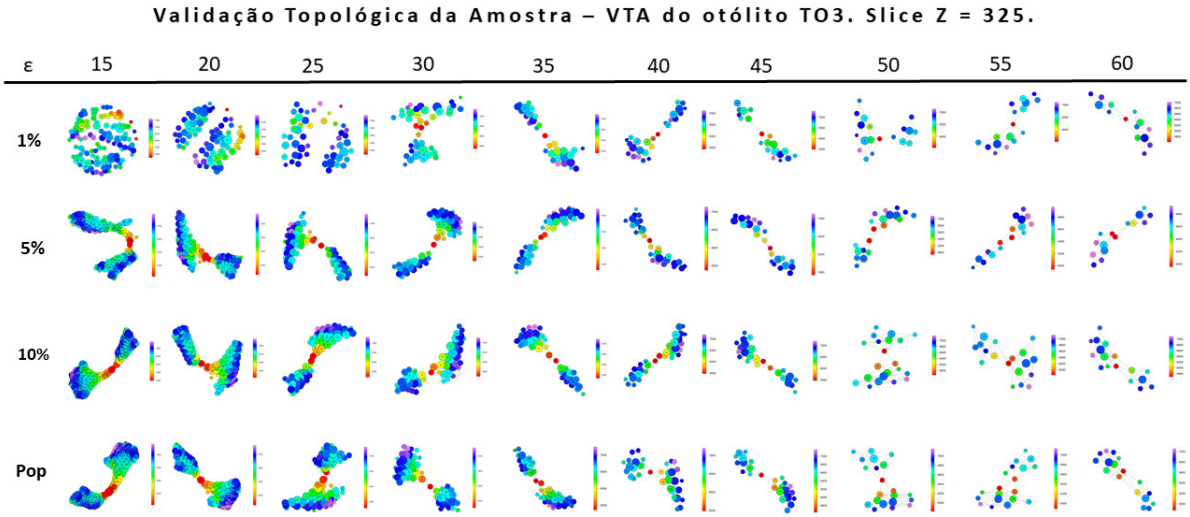


Figura 26 – Procedimento de Validação Topológica (VTA) da Amostra para o otólito TO3, *Thunnus obesus*. Grafos Ball Mapper (BM) da fatia Z = 325 para vários valores do parâmetro de ajuste ε (parâmetro do BM) em comparação entre Strat a 1, 5 e 10% e Pop. Observa-se que o BM para os valores dos parâmetros ε iguais a 15, 20 e 25 a 1% dos dados não fornecem informações interpretáveis sobre a densidade da fatia ensaiada. A partir de 5% dos dados, pode-se observar variações na densidade óssea semelhantes às da Pop para todos os valores do parâmetro ε. Esse fato sugere uma resolução mínima em 5% dos dados brutos como ideial para estudos de densidade otolítica com dados reduzidos de imagens. Grafos fora do intervalo [15; 60] para ε têm informações sobre forma e densidade óssea dissimilares aos da Pop.

Nota: O valor escolhido de Z para o VTA não é o valor médio da distribuição dos voxels pelos Zs, Figura 25A, mas sim a fatia intermediária da classe modal (poderia ser qualquer um dentro dessa classe) e neste caso específico para a imagem do TO3 tal classe é a faixa de intervalo [300, 350].

Na Figura 27 está a fatia Z=325 em diferentes resoluções para melhor explicar a presença de grafos desconexos para os valores do ε em 15, 20 e 25 com resolução da imagem em 1%. Para aquela fatia amostrada a 1%, não é possível fazer comparações com a mesma fatia da Pop, pois é difícil identificar as mesmas regiões ao longo das bordas.

Além disso, em 1% o conjunto de dados é dividido em duas partes, revelando que os voxels da parte mais fina não foram amostrados. Isto leva declarar que, ao empilhar as diversas fotos 2D para compor o otólito inteiro por uma imagem 3D com 1% de resolução, poderão existir regiões do otólito com voxels não amostrados. Isso sugere cautela na redução dos dados, especialmente em otólitos com formas irregulares como os das espécies *A. Solandri*, *T. Albacares* e *T. Obesus*, e justifica a utilização do procedimento VTA proposto.

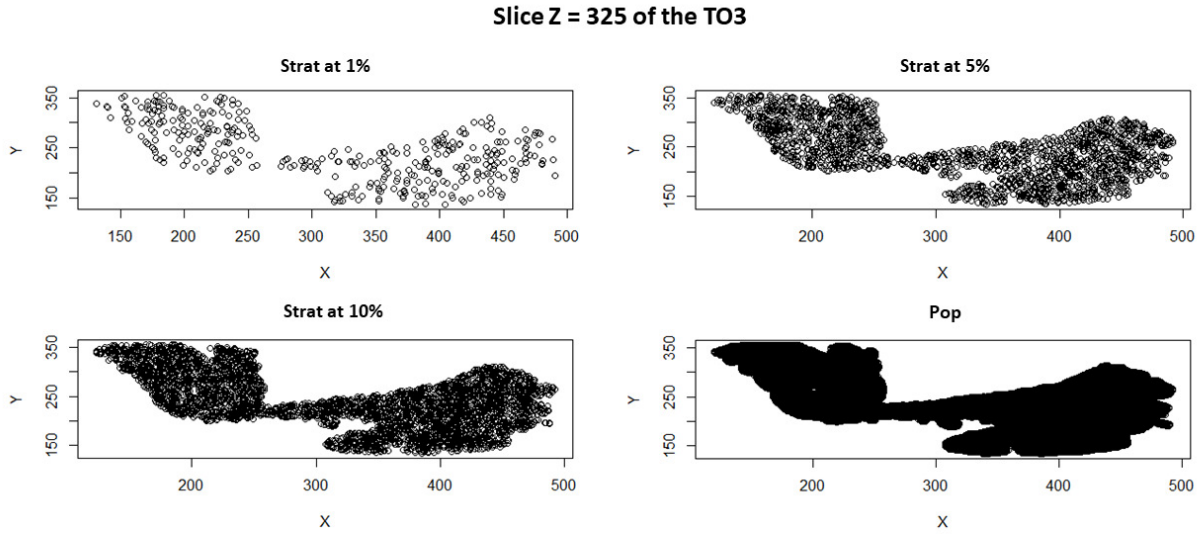


Figura 27 – Plots do mesmo slice Z=325 (slice central extraído da classe modal da distribuição dos Zs) com resoluções fixadas em 1, 5, e 10% extraídos por Strat, comparadas com o mesmo slice extraído da Pop (100% da resolução), dos dados do otólito TO3. Com 1% de solução torna-se difícil distinguir regiões e cavidades na borda da fatia, além disso, o conjunto de pontos parece ter sido dividido em duas partes, indicando que voxels da parte mais fina da imagem não foram amostrados, exprimindo a possibilidade de prejudicar uma análise cujo interesse seja densidade.

O critério adotado para a escolha do Z para o VTA está sustentado no fato dele ser de uma região do otólito com maior número médio de voxels por fatia. Pelas imagens de μCT, a região do núcleo foram as que mais careceram de voxels para serem representadas (isso será demonstrado na Seção 5.3 seguinte), logo escolher uma fatia sob estas considerações, é provável que ela contenha voxels do núcleo do otólito, como ele é a região de mais baixa densidade, implica que a fatia escolhida possivelmente terá valores de densidade em todo o domínio da mesma, algo útil para comparação visual por grafos e análise das variações de densidade.

Outra circunstância que também se mostrou útil na escolha de tais Zs para a VTA e fundamental para a validação das amostra, é a de o Z escolhido ser de uma

região que contenha o *sulcus acusticus*. Na amostra adotada neste estudo, com essa característica anatômica com sulcos mais profundos têm-se os otólitos da família *scombridae*. Devido a proximidade da parte mais baixa do sulcos ser próxima da região oposta do otólito, fatias com essa depressão são suscetíveis a possíveis perdas de valores de voxels e consequentemente de densidade após um procedimento de amostragem, como visto para a amostra em 1% do TO3 na Figura 27.

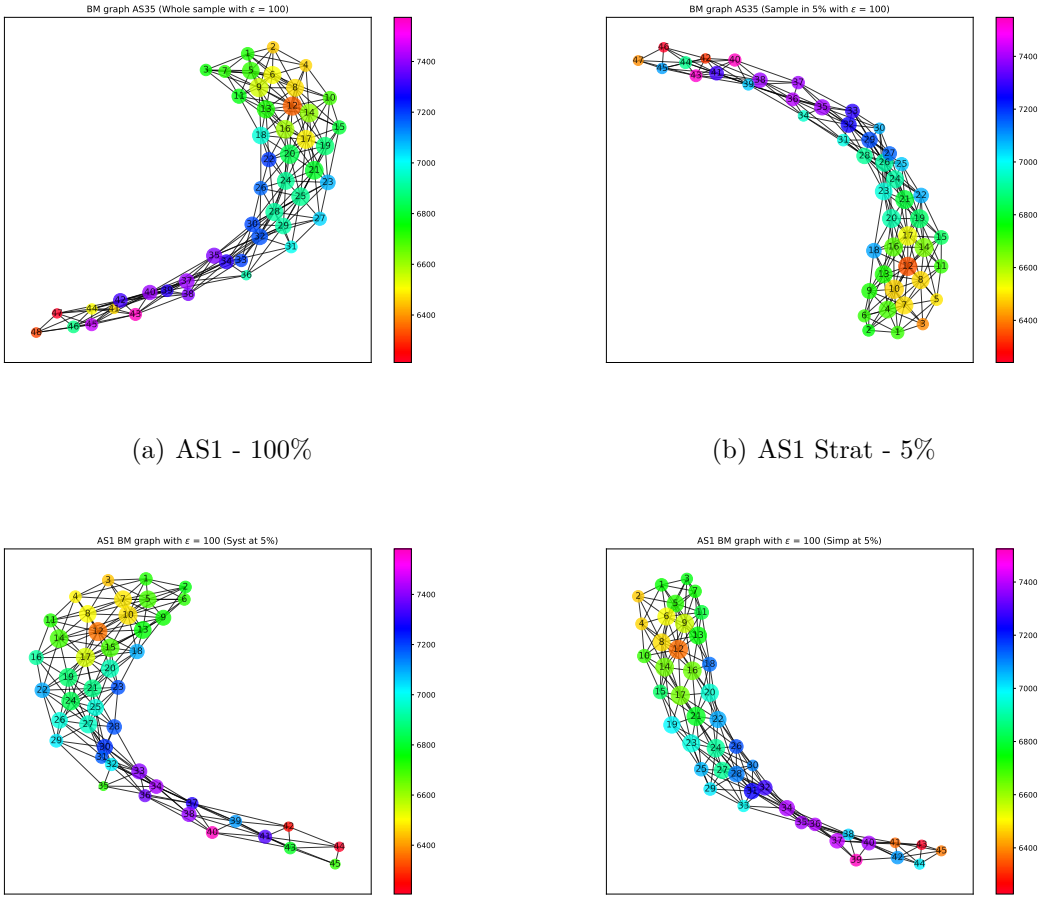
Pode-se dizer que, tal escolha auxilia na validação adequada da amostra, uma vez que, em baixas resoluções os grafos gerados a partir de slices de regiões que contenham o sulco são mais prováveis de serem desconexos. Portanto, com a validação realizada em slices sob essas suposições, é uma garantia que haverá a reprodutibilidade do procedimento nos demais slices das imagens.

Nos resultados obtidos até então, análises estatísticas revelaram representatividade mesmo em baixas resoluções das imagens tomográficas. Dos sistemas de amostragens aplicados na redução dos dados apenas Simp apresentou incertezas quanto ao seu uso para estudos de HU. Isso pode ter sido porque esse mecanismo de amostragem possui aleatoriedade máxima no processo de extração dos dados e, uma vez que imagens tomográficas 3D são um empilhamento de imagens 2D, impor a restrição de dividir os dados considerando cada foto 2D ou slice Z como estrato, caso da Strat e Syst, demonstrou melhor efeito de preservação da variável HU em até 95% de redução dos dados brutos das imagens da amostra.

Sobre o tamanho da resolução adequada para explorar variações de HU, o procedimento VTA se mostrou necessário nessa escolha. Sobre a possibilidade de redução de 95% dos dados das imagens para estudos de HU, será discutido os efeitos dessa redução feita pelos três tipos de Samps nas variações de HU usando o BM. Para tal, as semelhanças qualitativas serão feitas comparando a topologia representada por grafos gerados a partir dos dados brutos (100% ou Pops) e de Samps a 5% (Seção 5.3). Para a avaliação quantitativa será utilizado invariantes topológicos sobre as redes dos grafos (Seção 5.5). Antes é aberta a Seção 5.4 para discutir sujeiras nos otólitos captadas pelo BM.

5.3 Topologia de otólitos como perspectiva para explorar densidade

Na Figura 28, BMs com $\varepsilon = 100$ ilustram a variação da densidade óssea fornecida pelos dados brutos, Pop, e Samps a 5% para os diferentes tipos de amostragem sobre dados da imagem do otólito AS1. Todos os BMs obtidos a partir dos sistemas de amostragem empregados, Strat (Figure 28(b)), Syst (Figure 28(c)) e Simp Figure (28(d)) tiveram variações de densidade HU semelhantes às fornecidas pelo BM da Pop (Figura 28(a)).



(a) AS1 - 100%

(b) AS1 Strat - 5%

(c) AS1 Syst - 5%

(d) AS1 Simp - 5%

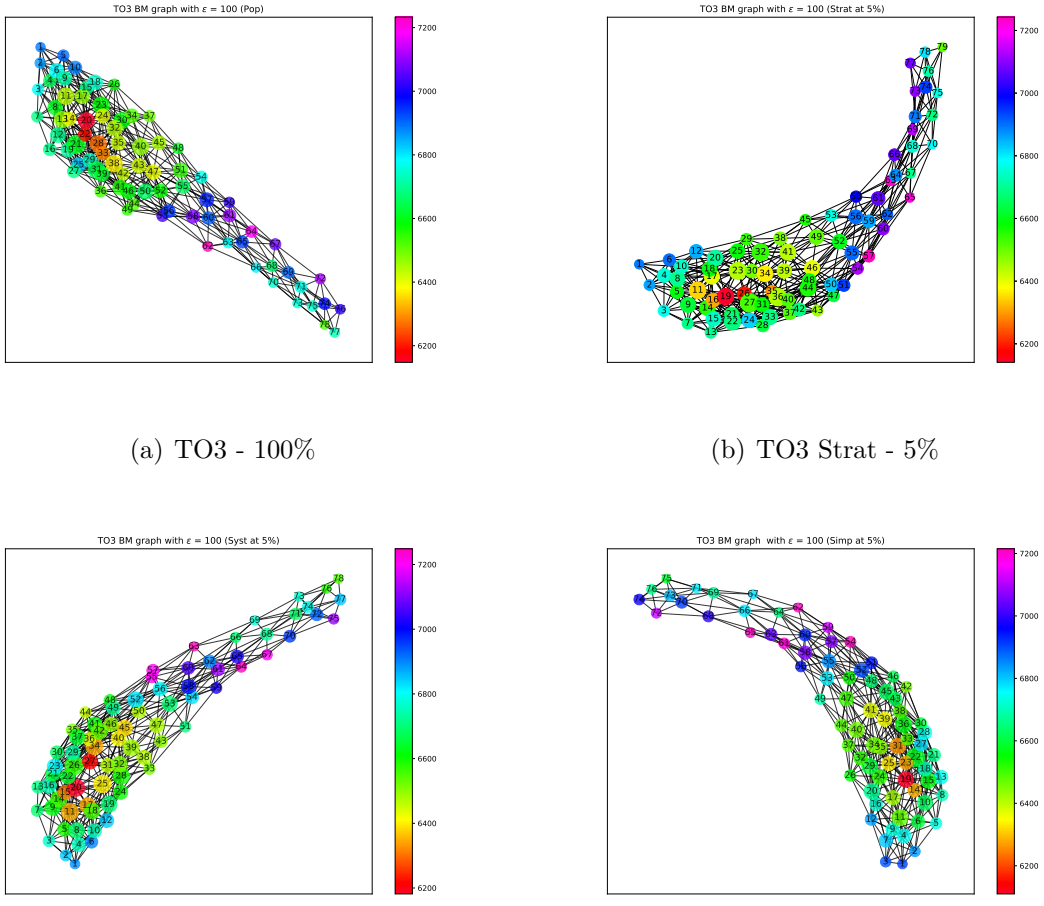
Figura 28 – Grafos Ball Mapper (BMs) com $\varepsilon = 100$ para o AS1 sobre os (a) dados brutos (Pop) e amostras (Samps) a 5% (b) Strat, (c) Syst e (d) Simp. Pode-se observar que os três sistemas de amostragens aplicados na redução dos dados (c–d) forneceram variações de densidade similares àquela revelada pelo grafo BM gerado sobre os dados brutos ou Pop (a). É possível observar ainda que as variações de HU na escala de cores foi reproduzida, indicando que a resolução fixada em 5% representa os dados brutos das imagens quanto a densidade otolítica, HU, de modo fidedigno.

Uma quantidade menor de nós nos grafos provenientes da amostragem pode ser observada. Foram 47 para a Strat e 45 para Syst e Simp contra os 48 nós do BM que representa a topologia da Pop. Pode-se dizer que essa diferença é mínima considerando uma redução de 95% nos dados. As variações de densidade, que é o mais importante, se mantiveram similares nos grafos provenientes dos dados reduzidos.

No BM da Figura 28(a), a bola 12 que encapsulou os voxels do núcleo na Pop foi reproduzida em todos os grafos das amostras com o mesmo número. Porém voxels

de menor densidade que os do núcleo foram observados na região caudal do grafo, bola 47 na Pop, 48 na Strat, 44 na Syst e 43 na Simp, sugerindo que algum tecido, possa ter permanecido naquela região após a limpeza, ou simplesmente a cauda desse otólito possui voxels de baixa densidade.

Apesar do método de amostragem Simp ter apresentado perdas de voxels por slices Zs (Tabela 5), os grafos da Figura 28(a) revelam que, mesmo para uma amostra extraída da menor imagem, otólito AS1, a variação de densidade foi semelhante a da Pop, indicando que as variações de densidade podem não ser afetadas pelo redução de dados via amostragem probabilística via Simp. No entanto, este resultado não se reproduziu para a amostra extraída por este método Simp a partir do otólito TO3. Ver Figura 29.



(a) TO3 - 100%

(b) TO3 Strat - 5%

(c) TO3 Syst - 5%

(d) TO3 Simp - 5%

Figura 29 – Grafos BMs com $\varepsilon = 100$ para o TO3 considerando os (a) dados brutos ou Pop, (b) Strat, (a) Syst e (d) Simp. Para o grafo feito sobre a amostra reduzida por Simp a região do núcleo foi representada por apenas um vértice, bola vermelha 19, enquanto os demais, similarmente ao da Pop, com dois.

Em um modo geral, pelos grafos BM obtidos a partir do TO3, pode-se declarar que a variação da densidade óssea foi representada pela topologia nos três sistemas de amostragem aplicados, Figura 29(b–c), fluindo dos valores de densidade mais baixos, bolas vermelhas, na região do núcleo até seus valores mais altos na região da cauda (bolas roxas), comparando com o grafo da Pop. Porém, o grafo BM obtido de Simp, Figura 29(d), representa o núcleo do otólito com apenas uma bola, vértice vermelho 19, enquanto os demais com duas bolas, Figuras 29(b) e 29(c), a princípio isso parece razoável. Essa observação pode ter sido efeito das diferenças nas proporções de voxels por Z observadas em amostras extraídas via Simp, como demonstrado para AS1 na Tabela 5, e para este otólito TO3 no Apêndice C (Tabela 12), já que essa perda de voxels leva embora os valores de HU e, consequentemente, podem afetar a variação da densidade revelada pelo grafo.

Uma análise mais aprofundada sobre os grafos do TO3 revela que a bola vermelha 19 do grafo da Simp codifica 94.927 voxels, isso representa 3,1% dos voxels encapsulados pelas bolas 20 e 22 do grafo da Pop. Para uma redução de 95% nos dados brutos isso também parece razoável, no entanto as bolas vermelhas 19 e 26 dos grafos da Strat e as bolas 20 e 27 do grafo da Syst codificam 5,4 e 5,8%, respectivamente, dos voxels encapsulados pelas bolas vermelhas 20 e 22 da Pop, o que é uma representação mais próxima da resolução de 5% estabelecida na configuração do tamanho das amostras durante o procedimento de amostragem. Esse resultado provoca incertezas quanto ao uso da Simp para estudos de densidade otolítica com imagens reduzidas por esse método.

Uma consequência da investigação quanto a aplicação da TDA em imagens de otólitos, foi a detecção de anomalias e sujeiras pelo o BM através da estrutura de alguns otólitos. Sobre essa descoberta, avaliou-se o comportamento das anomalias sobre dados reduzidos por Strat em relação aos dados brutos. Para fazer a limpeza foi usado simples algoritmos de tratamento de dados. Tal descoberta é exposta na seção que segue.

5.4 *Ball Mapper* como ferramenta de segmentação para otólitos

A extração e a limpeza de otólitos de peixes é um desafio uma vez que eles são pequenos e delicados. Tal característica pode ocasionar quebra que deixam pedaços, ou alguma sujeira na superfície dos otólitos, que podem passar despercebidas quando eles vão para análise em algum aparelho. Nisso, a quebra de otólitos durante o processo de extração pode comprometer os dados e dificultar a interpretação dos resultados, e por sua vez a sujidade pode comprometer a análise da composição. Devido a esses fatos, existem estudos dedicados a fornecer métodos e técnicas para extração e para avaliação de otólitos danificados (MYERS et al., 2020; BARDARSON et al., 2014).

Aqui, a preocupação com peças partidas ou sujidade retida na superfície de otólitos será diminuída, pois a topologia, através da técnica BM, codifica estas peças, ou como se diz em TDA, anomalias, em partes desconexas dos grafos BM, após identificados os dados de sujeira no grafo, eles podem ser simplesmente removidos a partir dos próprios dados por algoritmos simples.

Após identificar as anomalias no grafo, a função *"bm.points_covered_by_landmarks"* do próprio BM ajuda a localizar os voxels que correspondem aos valores anômalos (sujeiras ou pedaços de otólitos). A remoção desses voxels pode ser feita diretamente nos dados da imagem tomográfica (no formato de nuvem de pontos) utilizando um algoritmo de filtragem de linhas, aqui foram utilizadas as funções *"`drop`"* e *"`isin`"* do pacote `Pandas` do `Python`. Após limpar o conjunto de dados, basta reprocessar o algoritmo BM com os dados limpos. Esse procedimento ajuda a melhorar a precisão e confiabilidade na análise de dados a partir dos grafos BM.

Na Figura 30A é observado um fragmento destacado do otólito AS3 a partir de sua imagem 3D de μCT original. Tal fragmento foi detectado através do grafo BM pela bola 2, Figura 30B. Na Figura 30C, ver-se o grafo BM processado após a remoção dos voxels referentes àquele pedaço quebrada, deixando-o livre de quaisquer dados inconsistentes que poderiam comprometer a análise dos dados.

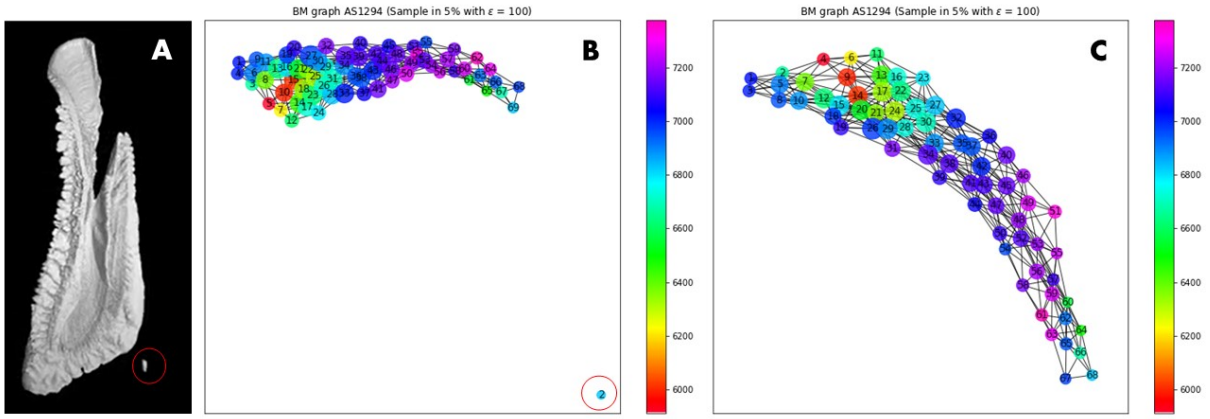


Figura 30 – (A) Imagem 3D de μCT sem escala do otólito AS3 revelando um fragmento, circulado em vermelho. (B) Grafo *BM* do AS3 no qual a bola 2, circulada em vermelho, desconectada da rede principal do grafo, encapsulou os voxels referentes ao fragmento observado. (C) Grafo BM do AS3 após a remoção dos voxels do fragmento.

Na Figura 31A, ver-se o grafo BM para dados brutos do otólito AC8, o qual apresentou um vértice (bola 262) detectado como anomalia. Ao avaliar tal vértice, percebeu-se que ele encapsulava apenas um voxel. O fato de ser apenas um único voxel, em um

universo de centenas de milhares, indica que tal voxel pode não ser decorrente de sujeira, mas de alguma instabilidade instrumental ocorrida durante o procedimento de aquisição da imagem deste otólito a partir do aparelho de μCT, o que sugere que TDA pode ser eficaz também na detecção de instabilidade instrumental durante a coleta de informações de dados de imagens. Na Figura 31B está o grafo dos dados brutos do otólito AC8 após a limpeza do voxel anômalo. Na Figura 31C o grafo BM para uma Strat a 5% do AC8, revelando a similaridade do grafo e da variação de densidade óssea em relação ao grafo obtido a partir dos dados brutos.

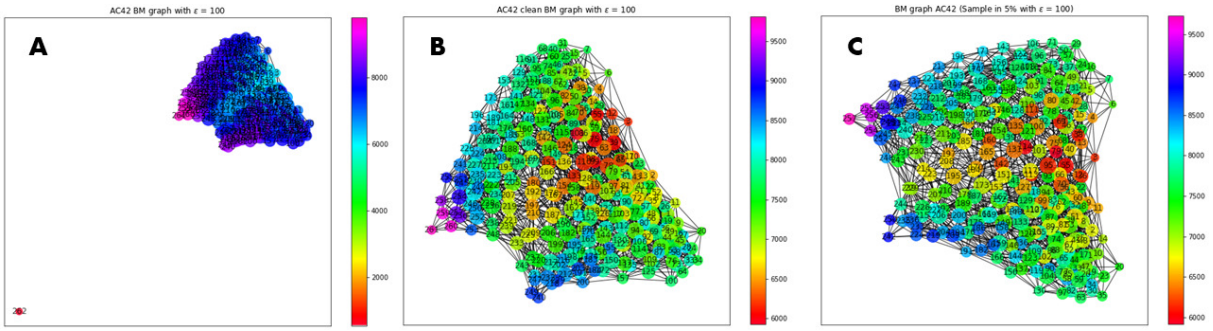


Figura 31 – (A) Grafos BM do otólito AC8 a partir de seus dados brutos. A bola 262 detectou apenas um voxel anômalo, sugerindo que pode ter sido instabilidade instrumental no processo de aquisição da imagem pelo aparelho de μCT. (B) Grafo BM a partir dos dados brutos após limpeza do voxel anômalo. (C) Grafo BM de uma Strat a 5%, revelando similaridade na variação de densidade em relação àquela observada a partir do grafo obtido dos dados brutos.

Nas Figuras 32A e 32B, nota-se sujeira detectada pelo grafo BM a partir dos dados do otólito AS2 em 100% e em 5% da resolução. Observe que a sujeira tem densidade muito menor do que o restante do otólito, sugerindo que algum tecido orgânico ou líquido pode ter permanecido em sua superfície após o procedimento de limpeza manual. Após de identificar os voxels que caracterizam a sujeira através do *BM*, novamente o algoritmo de tratamento de linhas foi aplicado e o grafo *BM* foi refeito. Pós procedimento de limpeza, a Figura 32C mostra como ficou o grafo BM para a Pop e a Figura 32D para a Strat em 5%. Neles é possível observar a eficiência da amostragem probabilística na redução da resolução das imagens mesmo com sujeira, pois é possível ver que a amostragem capturou a parcela proporcional da sujeira que residia nos dados brutos do otólito, demonstrando que até mesmo em otólitos sujos não haverá perda de informações quando for necessário reduzir imagens tomográficas com anomalias a fim de ganhos computacionais (Figura 32B).

Grafos, pós procedimento de limpeza e resultantes de Strat a 5%, revelando as variações de radiodensidade HU para os 16 demais otólitos, que não estão no escopo dos resultados, podem ser vistos no Apêndice E (Figura 47(a–p)).

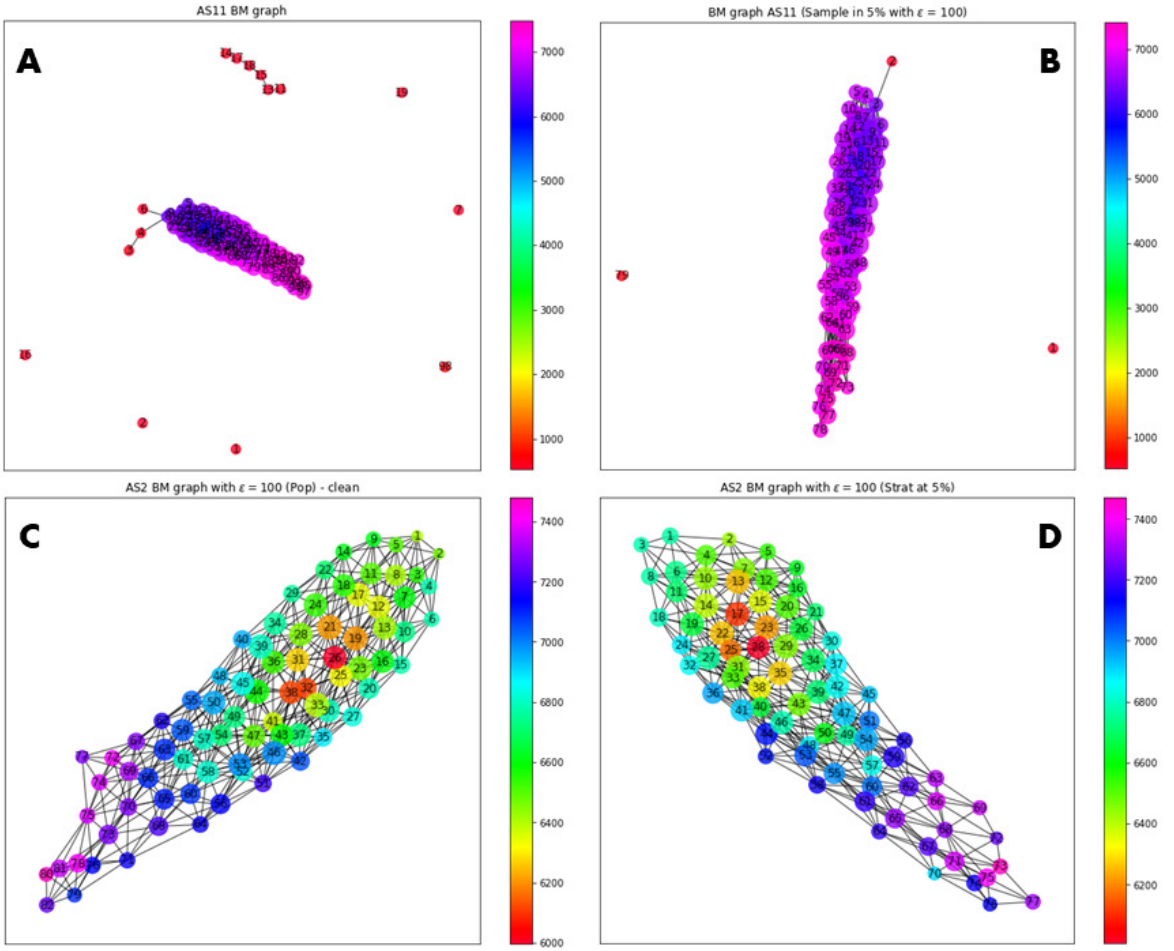


Figura 32 – Grafos BM com $\varepsilon = 100$ para o AS2. (A) com sujeira detectada para 100% da resolução (Pop). (B) com sujeira detectada pelas bolas vermelhas 1, 2 e 79 para uma Strat em 5%, revelando a eficiência da amostragem ao capturar a parcela proporcional da sujeira. (C) Grafo BM da Pop pós limpeza. (D) Grafo BM da Strat a 5% pós limpeza.

5.5 Invariantes Topológicos como ferramenta para comparação entre grafos

A Tabela 6 traz Invariantes Topológicos (ITs) calculados sobre as redes dos grafos BM para alguns otólitos da amostra, para Pop e Strat a 5%. Ao lado de cada IT amostral, está o percentual que ela representa em relação a seu valor na Pop. Observa-se que todos os valores, Pop/Strat, são relativamente próximos, exceto para a ***Assortatividade*** do AS1 em 5%, que deu apenas 61.8% do valor da Pop. Não encontrou-se evidências que justifique tal diferença. O que pode ser declarado, com base no valor e na definição dessa IT, é que vértices de graus semelhantes tendem se conectarem mais entre si na amostra reduzida. Por outro lado, pode ter sido simplesmente um caso isolado para este otólito.

Tabela 6 – Invariantes topológicos (ITs) para as espécies com maior abundância na amostra, calculados em redes obtidas a partir de grafos *BM* com parâmetro $\varepsilon = 100$. As respectivas representações percentuais revelam valores relativamente próximos para as ITs, exceto para a Assortatividade do AS1. Para este caso, não foram encontradas evidências que justificasse essa diferença, o que pode ter sido um caso isolado para o otólito AS1. Esse resultado simplesmente permite declarar que o gráfico da Strat em 5% amostrado de AS1 tem agora mais nós de grau semelhante conectados em comparação ao de sua Pop. Conclui-se, no geral, que a topologia de rede dos dados populacionais foi relativamente próxima da calculada sobre os dados reduzidos em 95%.

IT	AS1		AS2		AS3		TO3		AC8	
	Pop	Strat em 5%	Pop	Strat em 5%	Pop	Strat em 5%	Pop	Strat em 5%	Pop	Strat em 5%
Agrupamento	0.667	0.657 (98.5%)	0.616	0.623 (98.9%)	0.640	0.638 (99.7%)	0.635	0.634 (99.8%)	0.570	0.565 (99.1%)
Grau do nó	5.849	5.919 (98.8%)	9.629	8.634 (89.7%)	7.879	7.562 (96.0%)	9.126	9.202 (99.2%)	19.079	18.314 (96.0%)
Assortatividade	0.144	0.233 (61.8%)	0.429	0.459 (93.5%)	0.425	0.394 (92.7%)	0.426	0.362 (84.9%)	0.307	0.309 (99,3%)
Caminho mais curto	2.952	2.910 (98.6%)	3.014	3.020 (99.8%)	3.078	3.101 (99.3%)	3.008	2.998 (99.7%)	3.225	3.248 (99.3%)
Eficiência global	0.477	0.479 (99.6%)	0.448	0.449 (99.8%)	0.453	0.451 (99.6%)	0.460	0.459 (99.8%)	0.387	0.384 (99.2%)
Eficiência local	0.831	0.823 (99.0%)	0.808	0.804 (99.5%)	0.817	0.816 (99.9%)	0.814	0.814 (100%)	0.782	0.779 (99.6%)
Densidade	0.210	0.208 (99.0%)	0.163	0.164 (99.4%)	0.178	0.178 (100%)	0.178	0.174 (97.7%)	0.089	0.088 (98.9%)
Transitividade	0.609	0.599 (98.4%)	0.577	0.573 (99.3%)	0.595	0.590 (99.2%)	0.587	0.583 (99.3%)	0.525	0.519 (98.9%)
Conectividade	4	4	5	4	3	4	5	4	6	7
Número de nós	48	47 (97.9%)	82	77 (93.9%)	69	68 (98.5%)	78	79 (98.7%)	261	257 (98.5%)
Número de arestas	237	225 (94.9%)	543	480 (88.4%)	418	405 (96.8%)	534	537 (99.4%)	3051	2883 (94.5%)
Caract. de Euler (χ)	-189	-178 (94.2%)	-461	-403 (87.4%)	-349	-337 (96.6%)	-456	-458 (99,6%)	-2790	-2626 (94.1%)

A ideia central dos ITs é usá-los para perceber o quanto a topologia dos dados populacionais foi afetada após uma redução de 95% dos voxels das imagens tomográficas. Sob essa consideração, as aproximações dos ITs permitem declarar que mesmo com uma redução de 95% nas imagens foi possível observar topologias similares nas amostras a 5%. Essa representatividade topológica significativa somada com a representatividade estatística (Seção 5.1) são evidências indicativas de que estudos de variações da densidade otolítica podem ser desempenhados com apenas 5% da resolução das imagens originais.

Na Teoria de Grafos, dois grafos são ditos equivalentes ("simplesmente para não dizer iguais") quando existe isomorfismo[1] entre eles. O isomorfismo entre grafos é garantido se os ITs ***conectividade*** *(de arestas)*, ***número de nós*** e ***número de arestas*** forem iguais (SOMKUNWAR; VAZE, 2017). Isso não é observado em nenhum par dos grafos avaliados na Tabela 6, assim, apesar de útil, a diferença percentual nas ITs, oferece apenas medida "grosseira" de avaliar o erro% associado da representatividade entre os grafos.

Determinar se dois grafos são isomórficos é um problema computacionalmente desafiador, e essa seção oferece apenas uma abordagem mais flexível para avaliar a similaridade estrutural entre grafos comparando ITs. Vale lembrar que o problema de isomorfismo entre grafos é NP-completo, o que significa que não há algoritmo conhecido para resolvê-lo em tempo polinomial para todos os casos. Entretanto, para muitos grafos, alguns algoritmos existentes conseguem determinar o isomorfismo entre eles de maneira eficiente. O algoritmo testado (`NetworkX`, biblioteca `Python`, (HAGBERG; CONWAY, 2020)) nos grafos obtidos nesta tese não conseguiu concluir a tarefa, portanto não há declarações quanto a isomorfismo entre os grafos de Pop e Samps obtidos aqui.

A Tabela 7 exibe os tempos aproximados para processar os cálculos da topologia em alguns otólitos da amostra. Uma redução drástica do tempo de processamento de horas para minutos foi obtida sobre os dados reduzidos via amostragem probabilística. Observe que, para otólitos maiores o tempo aumenta, chegando a mais de 4 dias para a Pop dos otólitos ACs.

Tabela 7 – Tempo de processamento do cálculo da topologia de alguns otólitos. Comparação entre grafos BM de Strat em 5% e Pop com $\varepsilon = 100$. Esses tempos são referentes a topologia calculada após o procedimento de limpeza, por exemplo, o tempo de processamento para o cálculo da topologia do otólito AS2 sujo é de aproximadamente 8 horas.

Amostra	AS1		AS2		AS3		TO3		AC8	
	Pop	Strat	Pop	Strat	Pop	Strat	Pop	Strat	Pop	Strat
tempo*	1,5 h	5 m	6,5 h	15 m	5 h	10 m	6 h	15 m	4,5 d	5 h

* d: dias; h: horas; m: minutos

[1] O isomorfismo é a principal noção de congruência em topologia

5.6 Homologia Persistente como ferramenta para classificação de otólitos

Assim como as discussões anteriores deixarem incertezas quanto ao uso da Simp para reduzir dados de imagens de μCT a fim de estudar densidade otolítica. Permanece em aberto possibilidades desta técnica de amostragem ser usada para reduzir dados a fim de estudos de outra natureza. Para avaliar a validade desse método em outras aplicações, nesta seção ele será aplicado na redução da resolução das imagens afim de esboçar um novo classificador e/ou descritor de otólitos com base em sua estrutura 3D, e com um tamanho de amostra inferior àqueles 5% usados na aplicação do BM nos estudos de HU.

Para atingir esse objetivo é utilizada outra técnica da TDA, a Homologia Persistente (HP). Essa técnica é capaz de extrair características topológicas de um espaço topológico de qualquer dimensão, nesta tese, imagens tridimensionais de otólitos. Como dito na metodologia, o algoritmo usado para computar o homologia foi o `giotto-tda` (`gtda`).

Para tal, extraiu-se 5 mil voxels por Simp de cada otólito, pois o algoritmo `gtda` trabalha com nuvens de pontos de mesmo tamanho, alé disso, os dados de todos os otólitos são processados de uma vez simultaneamente pelo algoritmo. A escolha por esse tamanho de amostra está apoiada no fator custo computacional, pois cálculos de HP são mais caros computacionalmente que àqueles que geram grafos. Assim, a definição da resolução foi baseada na quantidade de voxels que o algoritmo conseguiria processar com a capacidade computacional disponível para desenvolvimento desta pesquisa. Os testes empíricos apontaram aproximadamente 100 mil voxels, com isso, extraiu-se amostras de tamanho 5 mil voxels de cada um dos 21 otólitos, totalizando 105 mil voxels a serem processados.

Sobre os 105 mil voxels, a capacidade computacional foi reduzida para aproximadamente 1 hora de processamento, a partir de um tempo inicialmente desconhecido sobre um primeiro teste usando apenas uma imagem de 5% de resolução do menor otólito, o AS1. Pode-se dizer que se o algoritmo não consegui concluir a tarefa para apenas uma imagem em 5% de resolução ainda está distante um estudo de classificação de otólitos com imagens 3D em a alta resolução. Será demonstrado que a redução proposta em 5 mil voxels foi suficiente para se alcançar resultados significativos, exibindo uma classificação e/ou discriminação acurada dos otólitos.

O ganho computacional alcançado sobre esse estudo permite ainda inserir características topológicas adicionais a fim de melhorar a acurácia da classificação, levando a resultados que apontam HP como uma metodologia atrativa e eficiente na classificação/discriminação de uma amostra diversificada de otólitos, sugerindo que dois ou mais

estoques pesqueiros possam ser estudados em uma só análise.

A hipótese acima, pode ser dada como possível porque o baixo custo computacional, aliado a uma amostra diversificada, permitiu realizar de modo rápido diversas interações do algoritmo, que exibiram vários valores de classificação permitindo discutir sobre tais valores a discriminação e classificação em relação a divisão amostral por espécie.

Como visto na seção de fundamentos teóricos, a HP mensura as características topológicas por cada dimensão de homologia, número de componentes conectados β_0, números de furos unidimensionais ou furos circulares β_1 e números de vazios ou furos bidimensionais β_2, e essas características são definidas pelos *números de Betti* (Definição 2.3.13), e as associa a um diagrama de persistência (Definições 2.3.15 e 4.2.3).

A partir de cada diagrama de persistência dos 21 otólitos (ver Apêndice F - Figura 48(a–u)), o algoritmo calcula as entropias (Definição 4.2.12) associadas àquelas características topológicas em cada uma das dimensões de homologia (β_0, β_1 e β_2), produzindo três valores de entropia, que podem ser traduzidas para coordenadas x, y e z, respectivamente. Considerando esses valores como coordenadas, um scatter plot 3D produz a chamada matriz de característica, a qual permite avaliar qualitativamente a formação de clusters, correspondentes a grupos de formas. As entropias, ainda podem ser normalizadas (Definição 4.2.13) afim de se ter outra perspectiva visual da matriz de característica, fornecendo uma possibilidade adicional para a interpretação e distinção dos grupos de formas (clusters). Serão adotadas as notações Ex, Ex e Ez para entropias não normalizadas, e Ex_n, Ex_n, Ez_n para entropias normalizadas.

Através das entropias, cada otólito como nuvem de 5 mil voxels fica representado apenas por um ponto tridimensional (Ex, Ex, Ez) ou (Ex_n, Ex_n, Ez_n) em uma matriz de característica, de onde é observar clusters que podem sugerir a classificação por grupos de formas dos otólitos. A Tabela 8 apresenta os resultados das entropias normalizadas e não normalizadas para os dados dos 21 otólitos.

Usando box plots pode-se perceber a variabilidade das entropias, por cada dimensão de homologia. Na Figura 33 tem-se o box plot das entropias não normalizadas (Ex, Ey e Ez) à direita e das normalizadas (Ex_n, Ey_n e Ez_n) à esquerda. O box plot da entropia Ex referente a β_0 indica pouca variabilidade entre seus valores, de fato seus valores de entropia se encontram no intervalo [12.193, 12.207], e os extremos desse intervalo são referentes respectivamente aos otólitos TO5 e AC1 (Tabela 8). As entropias normalizadas estão mais distribuídas, pois a mediana delas está mais próxima do centro das caixas dos box plots (Figura 33–direita).

Tabela 8 – Entropias de persistência não normalizadas e normalizadas para todos os otólitos. Essas entropias, em três coordenadas, representam cada otólito por um único ponto no espaço tridimensional, permitindo verificar diferenciações em suas formas apenas pela observação de clusters formados por esses pontos em um scatter plot 3D, chamado matriz de característica. Pontos próximos referem-se a otólitos de formas similares, pontos distantes a otólitos com formas dissimilares.

	Entropias					
	não normalizadas			normalizadas		
Otólito	Ex	Ey	Ez	Ex_n	Ex_n	Ez_n
OO1	12.201721	10.732908	8.534182	0.753373	0.794286	0.809354
OO2	12.204502	10.737227	8.612502	0.755272	0.795194	0.816041
AS1	12.195558	10.445681	8.012833	0.792155	0.835466	0.897542
AS2	12.202679	10.560703	8.196792	0.771025	0.812242	0.851831
AS3	12.203064	10.564955	8.240434	0.777501	0.821667	0.866354
TA	12.196943	10.545582	8.126741	0.777207	0.818078	0.847034
TO1	12.197761	10.524083	8.297888	0.780306	0.820563	0.873101
TO2	12.200754	10.588859	8.264123	0.770812	0.811872	0.859192
TO3	12.197844	10.539677	8.181661	0.771445	0.810495	0.845916
TO4	12.197562	10.575422	8.233137	0.765721	0.805858	0.841141
TO5	12.193242	10.587996	8.237134	0.769861	0.810010	0.848126
AC1	12.206948	10.749627	8.540693	0.747384	0.786743	0.794213
AC2	12.206865	10.726868	8.647546	0.738109	0.770645	0.791077
AC3	12.203569	10.733415	8.534432	0.736316	0.771396	0.782204
AC4	12.202761	10.735724	8.645976	0.734648	0.769922	0.785183
AC5	12.203657	10.771420	8.630898	0.735524	0.771670	0.782476
AC6	12.198096	10.753127	8.587861	0.739301	0.776547	0.785901
AC7	12.203291	10.739002	8.635719	0.735152	0.771065	0.783849
AC8	12.202515	10.775026	8.666321	0.732600	0.768581	0.770083
HP1	12.203016	10.748326	8.577368	0.735897	0.770496	0.774904
HP2	12.204526	10.754533	8.632251	0.731230	0.765301	0.776506

Os dados das entropias aplicados em uma modelo de *machine learning* (ML) treinam um classificador a fim de produzir o grau de separação quantitativa dos clusters observados na matriz de característica. O modelo aplicado é o *Random Forest* (Definição 4.2.14) acoplado a um *OOB score* (Definição 4.2.15) que fornece o valor quantitativo do grau de separação dos objetos avaliadas.

A escolha do *Random Forest* como modelo de ML nesse problema está apoiada na sua robustez em trabalhar com conjuntos de dados pequenos, em nosso caso 21 otólitos. Este modelo é menos propenso ao *overfitting*, além de apresentar estabilidade e resistência a pequenas variações, assim, em conjuntos de dados pequenos, onde pequenas mudanças podem ter um impacto significativo, o *Random Forest* pode fornecer resultados consistentes.

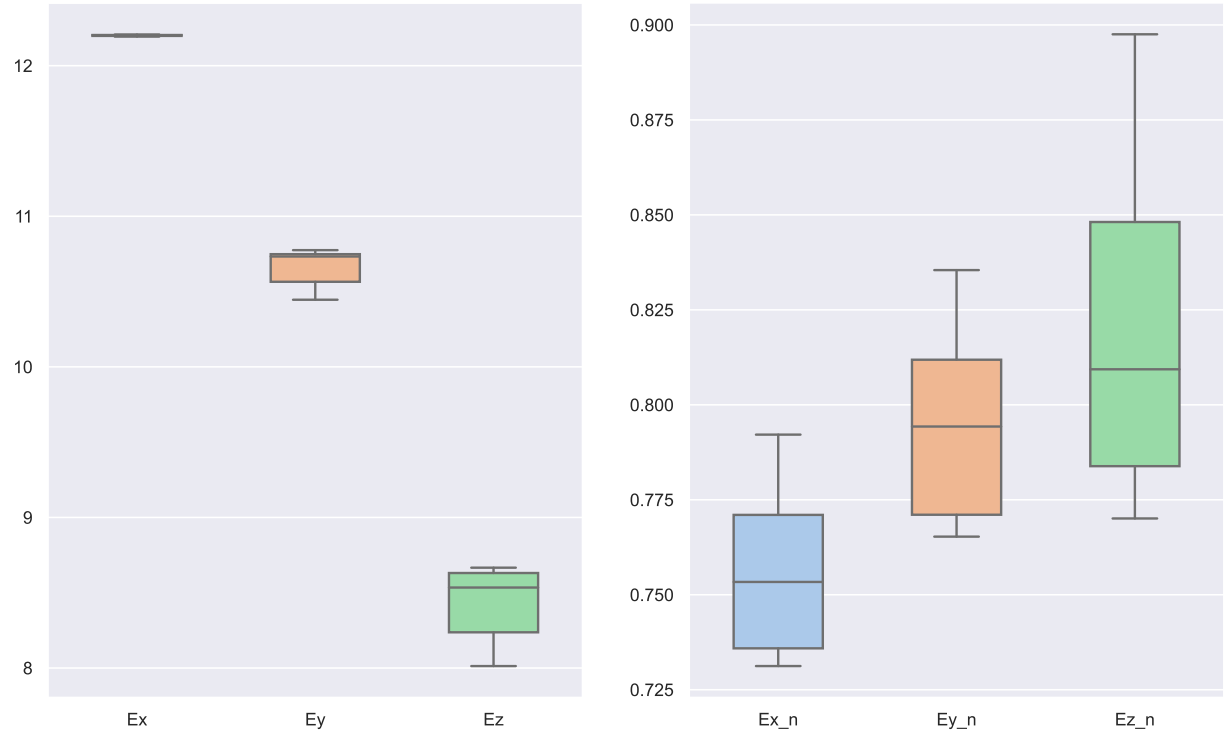


Figura 33 – À esquerda, entropias não normalizadas Ex, Ey e Ez, referentes as divisões de homologia β_0, β_1 e β_2 respectivamente. A amplitude da entropia Ex é de apenas 0.014, o que explica a pouca variabilidade em seus valores mostrada em seu box plot. À direita, as entropias normalizadas Ex_n, Ey_n e Ez_n são referentes as divisões de homologia β_0, β_1 e β_2 respectivamente. Após a normalização os dados das entropias diminuíram a assimetria, uma vez que suas medianas estão mais centralizadas nas caixas dos gráficos. No geral a maior variabilidade acontece para a entropia referente a β_2, porém não é observado *outliers* em nenhum dos casos, permitindo deduzir que nenhum otólitos teve uma classificação extrema relacionada à sua topologia.

A capacidade do *Random Forest* de construir várias árvores de decisão e combinar seus resultados por meio de votação ou média pode ser benéfica. Além disso, o *Random Forest* também é conhecido por lidar melhor no manuseio de características irrelevantes ou ruidosas que possam aparecer em conjuntos de dados pequenos, apresentar boa capacidade de generalização para diferentes conjuntos de dados o que é crucial para que um modelo se adapte e realize bem sua tarefa em conjuntos de dados pequenos, e por fim a possibilidade de usá-lo em situações em que capacidade computacional e tempo disponível para ajustar modelos são limitados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; BREIMAN, 2001).

Em suma, uma matriz de característica é usada para avaliação qualitativa dos grupos de formas, porque possibilita uma valorosa visualização dos dados das entropias. Um modelo de *Random Forest* com uso do *OOB score* é usado para mensurar quantitativamente o grau de dissimilaridade dos grupos de formas para os dois grupos de dados de entropias

(normalizada e não normalizada), produzindo uma classificação quantitativa que pode ser discutida sobre as duas formas das matrizes de características. Agora será discutido os resultados dessas classificações, qualitativa e quantitativa nas duas subseções seguintes.

5.6.1 Classificação qualitativa

Na classificação dos otólitos a matriz de característica foi desempenhadas nos dois casos dos dados das entropias, não normalizada (Figura 34A) e normalizada (Figura 34B), a fim de ampliar as possibilidades de visualização e discussão sobre essa classificação. Na Figura 34A, não normalizada, é evidente a observação que os otólitos se organizaram em dois clusters, um agrupando os otólitos das espécies *Acanthocybium solandri*, *Thunnus albacares* e *Thunnus obesus*, pontos nas cores vermelho, verde e roxo respectivamente (cluster 1), e o outro grupo agrupando os otólitos das espécies *Opisthonema oglinum*, *Acanthurus coeruleus* e *Haemulon plumierii*, pontos nas cores azul, laranja e celeste respectivamente (cluster 2).

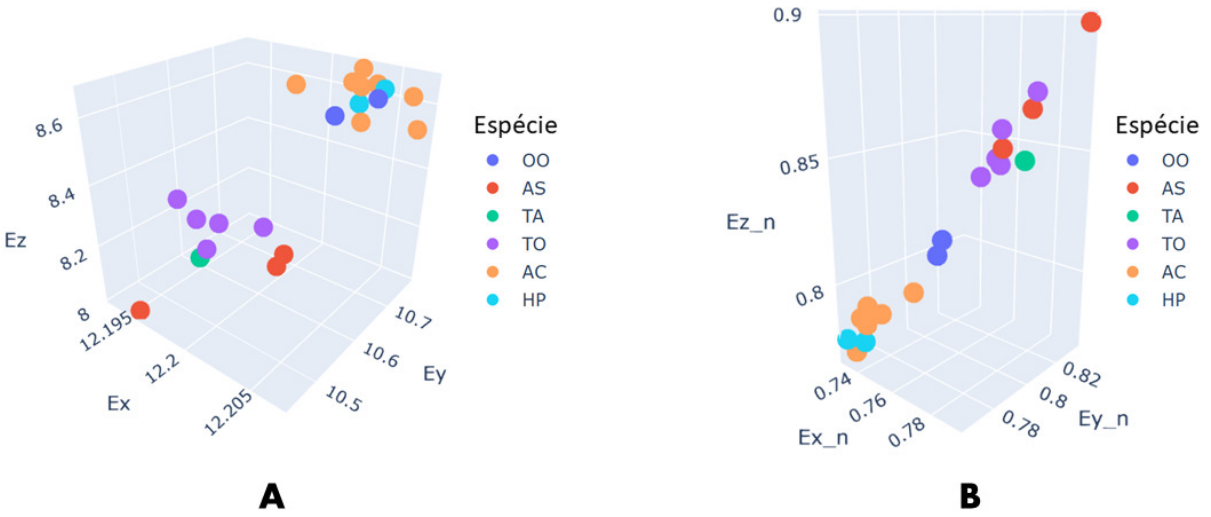


Figura 34 – (A) Matriz de característica não normalizada. É possível observar a formação de dois clusters e/ou grupos distintos, um cluster 1 formado pelos otólitos das espécies *Acanthocybium solandri*, *Thunnus albacares* e *Thunnus obesus*, e um cluster 2 formado pelos otólitos das espécies *Opisthonema oglinum*, *Acanthurus coeruleus* e *Haemulon plumierii*. No cluster 1, apenas um otólito apresentou forma mais dissimilar dos demais, o AS1, ponto vermelho mais baixo na matriz. (B) Matriz de característica normalizada. A normalização deixaram a matriz de característica normalizada com entropias mais distribuídas, como já previsto pela análise dos box plots exibidos na Figura 33–direita. Os dois otólitos de *Opisthonema oglinum* se posicionaram mais ao extremo do cluster 2, indicando que os otólitos dessa espécie são os de forma mais dissimilar dentre àqueles classificados no cluster 2.

Os otólitos agrupados no cluster 1 são da mesma família, *Scombridae*, sendo que *Thunnus albacares* e *Thunnus obesus* são do mesmo gênero, logo esses otólitos estarem em um mesmo cluster indicando similaridade em suas formas é um resultado esperado e adequado. O agrupamento dos otólitos das espécies *Opisthonema oglinum*, *Acanthurus coeruleus* e *Haemulon plumierii* pelo cluster 2 é menos disperso que os otólitos agrupados pelo o cluster 1, sugerindo que aqueles otólitos possuem maior grau de similaridade entre si, mesmo considerando sua maior variação de idade (Tabela 2).

Para a matriz normalizada, Figura 34B, a separação dos dois grupos de formas é mais sutil que na classificação dada pela matriz não normalizada, Figura 34A. Isso porque os dois otólitos de *Opisthonema oglinum* ficaram mais a extremidade de seu cluster, o que poderia sugerir uma classificação em mais de 2 clusters. Esse resultado sugere que os otólitos de *Opisthonema oglinum* são os mais dissimilares dentre os das espécies agrupadas no cluster 2. Curiosamente a classificação parece colocar os otólitos da amostra em torno de uma reta, e melhor ajustada aos pontos na matriz normalizada, Figura 34B, o que abre hipóteses para investigações de relacionamento dessas características topológicas (entropias) com variáveis do peixe. Essa discussão será retomada na próxima Seção 5.7.

A partir das imagens tomográficas, percebe-se que otólitos agrupados no cluster 2 possuem semelhanças perceptíveis em suas formas (ver Figura 19B e 19C, respectivamente *Acanthurus coeruleus* e *Haemulon plumierii*). Porém, mesmo com essas semelhanças entre suas formas, é interessante questionar por que esses otólitos foram classificados com significativa similaridade de formas entre si (observe na matriz normalizada, Figura 34B, que os otólitos de *Haemulon plumierii* persistem em meio aos de *Acanthurus coeruleus*, ao contrário dos otólitos de *Opisthonema oglinum* que apresentaram algum distanciamento), uma vez que esses peixes não possuem relações taxonômicas, assim, em teoria seria esperado que esses otólitos apresentassem maior dissimilaridade em sua morfologia.

Em uma investigação por espécie, dentre àquelas discriminadas no cluster 2, o *Haemulon plumierii* (grunhido-de-coração), pertencente a família *Haemulidae*, é um peixe encontrado em recifes de coral e ambientes costeiros do Atlântico Ocidental, incluindo o Golfo do México e o Caribe, em direção ao sul até o Brasil, incluindo as Antilhas (SMITH, 1997), vivendo em profundidade média de 3 - 40 m (CERVIGÓN, 1966). O *Acanthurus coeruleus* (cirurgião-patela-azul), família *Acanthuridae*, é também encontrado no mesmo ambiente de recifes de coral das águas tropicais do Atlântico Ocidental, na extensão que vai de Nova York, EUA e Bermudas até o Golfo do México incluindo o Caribe e Brasil além da Ilha da Ascensão (MARSHELL; MUMBY, 2015), vivendo em profundidades entre 2 - 50 m (QUÉRO et al., 1990). O *Opisthonema oglinum* (manjuba), pertence à *Clupeidae*, são peixes encontrados em grandes cardumes nas águas costeiras e oceânicas do Atlântico

Ocidental, que vão do Golfo do Maine (EUA), Bermudas, ao longo do Golfo do México, Caribe e Índias Ocidentais ao sul até Santa Catarina, Brasil. Também pode ser encontrado no Uruguai (NION et al., 2002) e Argentina (MENNI; RINGUELET; ARÁMBURU, 1984). Peixes dessa espécie podem ser encontrados em profundidades entre 1 - 50 m (JR, 2013).

Essas informações rotulam os indivíduos por habitat, e indicam um possível compartilhamento de ambiente entre peixes das espécies agrupadas no cluster 2, porém das três classificadas no cluster 2, a espécie *Opisthonema oglinum* é que está em um habitat de menor intersecção das outras duas espécies, isso pode indicar o fato dessa espécie apresentar forma mais dissimilar dentro de sua classificação no cluster 2. Partindo das informações levantadas na teoria, desses resultados e da premissa que o ambiente induz diferenças morfométricas nos otólitos (CADRIN, 2000; GAULDIE; CRAMPTON, 2002), esse achado deixa uma hipótese a ser investigada na ciência do otólito: a possibilidade de peixes sem relação taxonômica mas que compartilham habitats possuírem similaridade de formas entre otólitos.

Avaliando a classificação em um modo menos geral, a matriz normalizada (Figura 34B), ajuda a perceber que os otólitos da espécie *Acanthurus coeruleus*, que é a mais abundante na amostra, teve a formação de um subcluster com seis dos oito otólitos, referente aos otólitos AC2 — AC7 com idades entre 4 a 8 anos (Tabela 2) . Os outros dois ficaram orbitando tal subcluster nos extremos, o AC1 (1 ano) que ficou abaixo do subcluster e o AC8 (15 anos) que ficou acima do subcluster. Isso sugere que modificações de forma significativas nessa espécie ocorrem apenas em grandes intervalos de idade, e que em idades entre 4 a 8 anos modificações de forma podem ser pouco perceptíveis.

Os dois otólitos de *Haemulon plumierii*, com idades de 8 e 14 anos, tiveram poucas modificação em suas formas apesar desta diferença em 6 anos. Isso sugere que assim como os otólitos da espécie *Acanthurus coeruleus*, que apresentaram mudanças de formas significativas para intervalos de idade de 1 — 4 e de 8 — 15 anos, eles também sofrem poucas modificações de forma mesmo com intervalos relativamente longos de idade. Esse resultado ajuda a apoiar a hipótese levantada anteriormente de que eles podem ser semelhantes devido ao compartilhamento de habitat. Quanto aos dois otólitos de *Opisthonema oglinum*, eles tiveram poucas diferenças em suas formas, mas esse é um resultado esperado tendo em vista que eles possuem praticamente a mesma idade de 1 ano (Tabela 2).

No cluster 1, família *Scombridae*, é onde mais aparece diferenças entre as formas dos otólitos com relação a menores intervalos de idade. A exemplo, na espécie *Acanthocybium solandri* o otólito AS1, de 1.47 anos, foi o mais distante dentre os três de sua espécie, sugerindo que entre 1.47 e 2.89 anos alguma mudança de forma já pode ser observada.

5.6.2 Classificação quantitativa

Nessa classificação a separação dos grupos de diferentes formas é mensurada por um número que representa o grau de dissimilaridade para os grupos de forma na matriz de característica. Tal valor é calculado com uso do *OOB score* (Definição 4.2.15) associado ao modelo de classificação *Random Forest* (Definição 4.2.14).

Neste estudo, observou-se que, devido a amostra ser composta por várias espécies e com diferentes quantidades de otólitos por espécie, o *OOB score* não teve ocorrência única, ao contrário dos exemplos de teste do algoritmo encontrados na documentação do pacote `giotto-tda`, onde lá três grupos de formas são intencionalmente bem definidos. Diante dessa observação, na expectativa de identificar padrões nos valores do *OOB score*, calculados sobre os dados da imagens dos otólitos, quatro cenários com 10 iterações do algoritmo em cada, possibilitaram observar a existência de uma repetição nos valores do *OOB score* calculados neste estudo.

Os quatro cenários, foram montados de acordo com as variáveis que alimentam o *Random Forest*. Dois cenários foram construídos alimentado o modelo com as entropias normalizadas (Definição 4.2.12) e não normalizadas (Definição 4.2.13). Após isso, além das entropias normalizadas e não normalizadas, as métricas topológicas (*Wasserstein*, *Bottleneck*, *landscape*, *persistente image*, *Heat*, *Silhueta* e *Betti* – Definições de 4.2.6 a 4.2.11) foram incorporadas a alimentação do modelo, a fim de melhorar a acurácia da classificação, definindo os outros dois cenários.

Para os 10 processamentos do algoritmo realizados em cada cenário, o registrado das frequências dos valores do *OOB score* foi feito do seguinte modo: ao lado de cada primeira ocorrência de um *score*, dentro de um cenário, foi registrado sua frequência absoluta (F_a), e entre parênteses sua frequência absoluta global (F_g), esta última é a soma de todas as F_as observadas. Os valores únicos de *OOB score* estão destacados em negrito em suas primeiras ocorrências.

Ao observar os valores em negrito, pode-se perceber a ocorrência global de apenas sete valores de *score* para os 40 processamentos do modelo. Uma primeira observação levantada sobre esses valores de *score* é que eles foram iguais a alguma medida de separação (fração amostral) para a amostra dos 21 otólitos. Isso permitiu interessantes *insights* sobre a interpretação dos *score*s e a entender o grau da separação quantitativa das classes dos otólitos. Também serviu como uma comprovação da precisão dos valores de *score*s fornecidos pelo *Random Forest*, uma vez que tais valores foram iguais, em todas as casas decimais, aos valores decimais das suas respectivas frações amostrais. Os cenários montados, as frequências de ocorrência e a fração amostral de cada valor do *score* estão organizadas na Tabela 9.

Tabela 9 – Resultados da classificação quantitativa dos grupos de formas dos otólitos. Quatro cenários foram montados com base na normalização das entropias e na adição de métricas topológicas. 10 iterações do algoritmo em cada cenário foram processadas e os valores de *OOB score*(*score*) registrados. Em cada cenário, as frequências absolutas de cada cenário (F_a) e as frequências absolutas globais (F_g), considerando todos os cenários, foram registradas a cada primeira ocorrência de um *score*. Os resultados revelam a repetição de apenas sete valores de classificação no geral (negrito). Observou-se que cada *score* pode ser escrito na forma de fração e que essas frações correspondem a alguma divisão de separação ou classificação da amostra de estudo. A exemplo, o primeiro valor 0.42857142857142855, que classifica a amostra com aproximadamente 43% de dissimilaridade, é o mesmo valor que 9 dividido por 21, indicando que 9 dos 21 otólitos é dissimilar dos demais da amostra, isso precisamente separa os dois grupos de otólitos revelados pela matriz de característica na classificação qualitativa (9 da família *scombridae*, cluster 1, e os demais 12 do cluster 2).

Sem adição de métricas topológicas						
	Cenário I Entropias não normalizadas			Cenário II Entropias Normalizadas		
	OOB score	F_a (F_g)	Fração amostral	*OOB score*	F_a (F_g)	Fração amostral
1	**0.42857142857142855**	1 (2)	9/21	0.5714285714285714	3 (10)	12/21
2	**0.47619047619047616**	6 (6)	10/21	0.5714285714285714		
3	0.47619047619047616			**0.6666666666666666**	4 (6)	14/21
4	**0.5714285714285714**	2 (10)	12/21	0.5238095238095238	1 (8)	11/21
5	0.47619047619047616			0.6666666666666666		
6	0.5714285714285714			0.42857142857142855	1 (2)	9/21
7	0.47619047619047616			0.6666666666666666		
8	**0.5238095238095238**	1 (8)	11/21	0.5714285714285714		
9	0.47619047619047616			**0.6190476190476191**	1 (5)	13/21
10	0.47619047619047616			0.6666666666666666		
Com adição de métricas topológicas						
	Cenário III Entropias não normalizadas			Cenário IV Entropias Normalizadas		
	OOB score	F_a (F_g)	Fração amostral	*OOB score*	F_a (F_g)	Fração amostral
1	0.5714285714285714	4 (10)	12/21	**0.7142857142857143**	3 (3)	15/21
2	0.5238095238095238	6 (8)	11/21	0.6190476190476191	4 (5)	13/21
3	0.5714285714285714			0.6190476190476191		
4	0.5714285714285714			0.6190476190476191		
5	0.5238095238095238			0.6666666666666666	2 (6)	14/21
6	0.5238095238095238			0.5714285714285714	1 (10)	12/21
7	0.5238095238095238			0.6666666666666666		
8	0.5238095238095238			0.7142857142857143		
9	0.5714285714285714			0.7142857142857143		
10	0.5238095238095238			0.6190476190476191		

Agora será discutido como estes sete valores de *OOB score* fornecem alguma medida de separação/classificação da amostra dos otólitos em relação as espécies, em ordem dos valores de maior frequência absoluta global observada.

Sem dúvidas o *score* mais interessante observado é 0.5714285714285714 pois ele possui a maior frequência global e está presente em todos os quatro cenários. Olhando para a fração amostral 12/21 a qual ele corresponde, vemos que esse valor separa precisamente o cluster 1, onde estão os otólitos da família *Scombridae*, do cluster 2, o qual contém 12 otólitos das espécies *Opisthonema oglinum*, *Acanthurus coeruleus*, e *Haemulon plumierii*. Esse resultado de separação dessas duas classes ainda é fortalecido por mais duas ocorrências do valor 0.42857142857142855 (9/21) (posição 1 do cenário I e 6 do cenário II), complementar do valor de *score* em questão.

O valor 0.5238095238095238 (11/21) de classificação pode está incorporando 2 subclusters como dissimilares, que seria os otólitos de *Opisthonema oglinum*, 2/21, somados com 9/21 referente ao cluster 1 que agrupam a família *Scombridae*. Com isso, o seu complementar 10/21, 0.47619047619047616, (posição 2 do cenário I), consequentemente estaria representando os otólitos das espécies *Acanthurus coeruleus* e *Haemulon plumierii* que completam a amostra. Essa separação pode ser visualizada a partir da *matriz de característica* normalizada, Figura 34B. Essa justificativa pode ser sustentada no fato dos dois otólitos da espécie *Opisthonema oglinum* terem se distanciado de seu cluster 2, por serem os mais dissimilares dentre àqueles classificados em tal cluster, e se aproximado mais do cluster 1 (família *Scombridae*) na matriz de característica normalizada (Figura 34B).

O índice 0.6666666666666666, referente a 66,7% (14/21) de dissimilaridade, sugere que o otólito AS1, por ser o mais afastado dentre os de sua espécie (ponto vermelho mais distante observado nas duas matrizes, Figuras 34A e 34B) e TA por ser de espécie diferente (ponto verde, Figura 34A e 34B) podem ter sido classificados fora de seu cluster 1. Esse resultado sustenta a hipótese da classificação ser dada pelas frações 12/21+1/21+1/21+7/21, uma vez que 7/21 é a fração complementar da classificação em questão e representa o grau de similaridade dos otólitos restantes no cluster 1.

O *score* 0.6190476190476191 (13/21) que capta 61.9% de dissimilaridade (Tabela 9: posições 9 do cenário II e posição 2 do cenário IV) pode está fornecendo um significativo resultado de classificação/separação com aumento de características topológicas, uma vez que seu complementar 8/21 indica apenas 8 otólitos com maior grau de similaridade, o que pode estar correspondendo aos 8 otólitos da espécie *Acanthurus coeruleus*. Isso sugere que a adição de métricas topológicas pode melhor ajustar o modelo a ponto de capturar com precisão a similaridade de otólitos de mesma forma, separando-os por espécie, indicando que um cenário como o IV pode ajudar na classificação de otólitos, perspectivando a discriminação de estoques pesqueiros.

O resultado 0.7142857142857143 referente a fração amostral 15/21, também pode ser melhor interpretado ao olhar para seu complementar 6/21 que diz respeito a similaridade.

Assim este valor 6/21, no caso, estar indicando seis otólitos como similares, a possibilidade mais evidente, seria os seis otólitos da espécie *Acanthurus coeruleus* que estão mais agrupados dentro do cluster 2, como pode ser visto na matriz de característica não normalizada (Figura 34B). Esse resultado é condizente com o fato desses otólitos possuírem pouca variação de idade, de 4 a 8 anos (Tabela 2), sustentando a hipótese levantada no parágrafo anterior que o aumento de características topológicas melhoraram a acurácia da classificação, e aqui, não só por espécie mas também por idade.

Discutindo a classificação quantitativa de modo geral, parece razoável dizer que os 4 cenários estabelecidos, com 10 iterações em cada, são suficientes para se alcançar índices que revelem informações fiéis de classificação de otólitos com base em sua forma 3D usando análise topológica sobre imagens 3D de otólitos de peixes.

Com a adição das métricas topológicas junto as entropias não normalizadas, a classificação se restringiu aos dois valores de *score* mais recorrentes, (cenário III, Tabela 9). Esses valores parecem ser os que melhor discriminaram a amostra, pois foi um resultado esperado com base na classificação qualitativa (Figura 34). E, ao adicionar métricas topológicas aos dados das entropia normalizadas (cenário IV, Tabela 9), observou-se uma classificação mais acurada dentro de uma mesma espécie, assim, essa classificação mostrou-se capaz de captar formas dissimilares dentro de uma mesma espécie com indivíduos de idades diferentes. Sugerindo que este último cenário pode ser indicado em investigações com amostras de muitos otólitos de uma mesma espécie e com idades diferentes, o que acontece em estudos de identificação de estoques com base na forma do otólito.

Ainda sobre o cenário IV, os maiores valores para o *score* observados, podem de fato ser atribuídos a alimentação do modelo com mais métricas no sentido de melhorar a acurácia da classificação a nível de separar os otólitos de mais similaridade dentre os demais da amostra. Isso porque na matriz de característica vista na Figura 34B têm-se oito otólitos da espécie *Acanthurus coeruleus* mais semelhantes (mais clusterizados), então era esperado que essa melhora na classificação para os cenários com acréscimos de métricas topológicas capturassem o grau de similaridade daqueles otólitos, o que de fato ocorreu com o valor de dissimilaridade 61,9% já discutido. Para tal considerou-se seu complementar 38,1% de similaridade, que reflete a fração amostral 8/21 daqueles oito otólitos. Esses resultados permitem declarar que o modelo *Random Forest* com o aumento de métricas topológicas classificou com acurácia a amostra de otólitos deste estudo, apesar de pequena.

Viu-se que, a adição de características topológicas pôde fornecer uma classificação precisa pela forma dos otólitos, sugerindo que esta configuração de classificação do cenário IV pode ser útil para classificar otólitos de uma mesma espécie em que podem existir poucas diferenças em suas formas, podendo ser aplicada a indivíduos de mesma idade.

Atribui-se o fato de se ter obtido mais de um valor na classificação ao fato da amostra ter diferentes quantidades de otólitos em diferentes espécies e espécies diferentes com otólitos semelhantes, de uma mesma família, além da diferenciação de idade. No entanto, isso não dificultou a classificação, em contrário, uma amostra diversificada aliada as informações à priori sobre a mesma, ajudaram a verificar a metodologia de Homologia Persistente e validar o modelo de classificação *Random Forest* como métodos interessantes para a classificação de otólitos com base em sua forma 3D.

Uma observação sobre os resultados que ajuda a apoiar a declaração feita no parágrafo anterior, é o de os sete valores únicos de *OOB score* observados (considerando todos os cenários) possuírem média aritmética de suas frações igual a 12/21, que corresponde ao valor de classificação com maior frequência, indicando que em média existe a separação de dois clusters, um com 12 e outro com 9 otólitos, como visto na matriz de característica mostrada na Figura 34A.

De modo geral, os resultados da classificação quantitativa foram de acordo com as informações à priori da amostra, espécies, idades e formas, que tiveram mais força com a percepção de que cada valor de *OOB score* corresponde a alguma fração da amostra. As suposições alcançadas sobre tais resultados, por serem obtidos sobre uma amostra com apenas 21 otólitos e ainda distribuídos em seis espécies, permitem sugerir que a Análise Topológica de Dados pode ser usada como ferramenta na discriminação de otólitos com amostras maiores.

Em outras palavras, a declaração acima diz que, embora com uma amostra pequena e diversificada, os resultados levantados pelas classificações qualitativa e quantitativa abre a possibilidade para aplicar esses métodos em amostras maiores para averiguar diferenças de formas por idade, bem como para verificar se otólitos de uma mesma espécie são de diferentes estoques pesqueiros. Isso porque os resultados mostraram-se consistentes com as informações disponíveis a cerca da amostra, permitindo declarar que os mesmos foram acurados como prometia a teoria do modelo de *Machine learning* aplicado, *Random Forest*.

5.7 Relacionamento entre características topológicas e variáveis do peixe

Nesta seção será explorado os relacionamentos entre as entropias e as variáveis do peixe idade, comprimento e radiodensidade (HU). Para tal, análises de regressão são desempenhadas para as espécies que possuem pelo menos 5 indivíduos, que no caso são *Thunnus obesus*, 5 indivíduos, e *Acanthurus coeruleus*, 8 indivíduos (Tabela 1).

5.7.1 Análise sobre dados dos otólitos da espécie *Thunnus obesus*

A matriz de correlação da Figura 35 ajuda a verificar preliminarmente o grau de associação entre variáveis para dados da espécie *Thunnus obesus*. Observa-se correlações forte $(0,60 \leq |\rho| \leq 0,79)$ ou muito forte $(0,80 \leq |\rho| \leq 1)$ e negativas entre as entropias normalizadas Ex_n, Ey_n e Ez_n e as variáveis, idade (-0,8), comprimento (-0,9) e HU (-0,7), respectivamente. Uma correlação moderada $(0,40 \leq |\rho| \leq 0,59)$ é observada da associação entre a entropia não normalizada Ey coma as variáveis idade, comprimento e HU, 0,4 para cada relação mencionada.

Para qualquer entropia, normalizada ou não, a correlação com as variáveis, idade, comprimento ou HU é a mesma, diferindo apenas se a relação é crescente ou decrescente. Por exemplo, a correlação entre a entropia Ex_n e as variáveis idade, comprimento e HU é -0,80, -0,80 e 0,80 respectivamente. E entre a entropia Ey e essas variáveis a correlação é 0,4, 0,4 e -0,4, respectivamente.

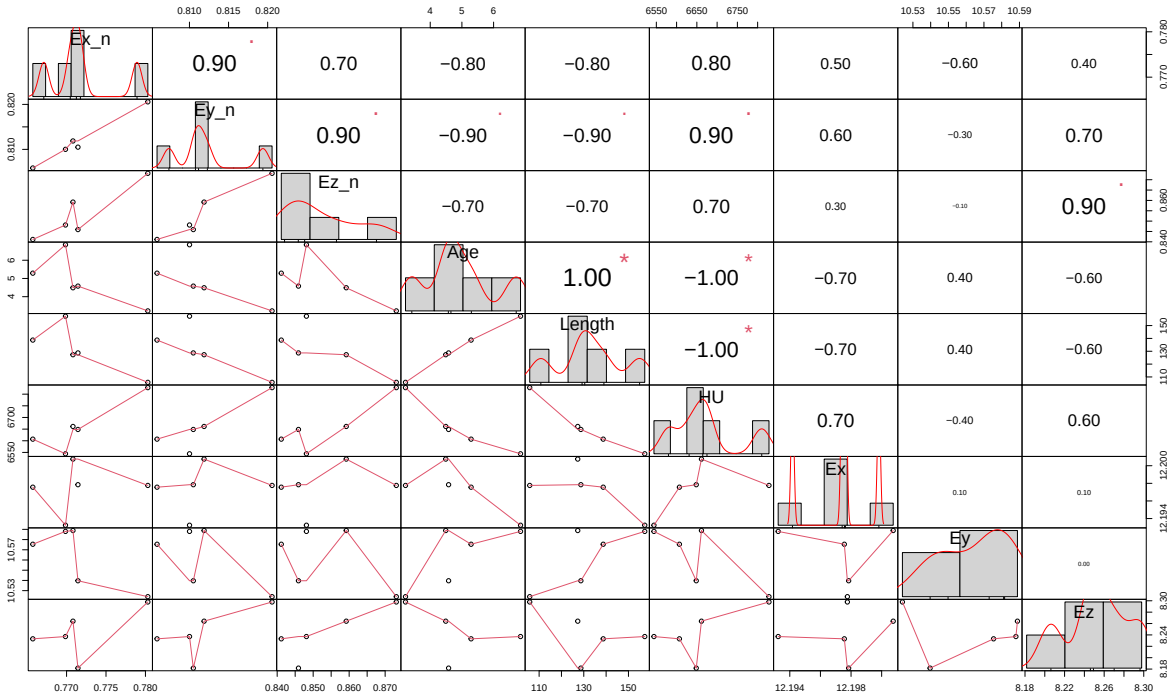


Figura 35 – Matriz de dispersão para dados dos otólitos da espécie *Thunnus obesus*. Na diagonal principal, estão as entropias normalizadas Ex_n, Ey_n, Ez_n, variáveis do peixe idade (Age), comprimento (Length) e radiodensidade em unidades hounsfield (HU) e por fim as entropias não normalizadas Ex, Ey e Ez. A entronpia Ey_n parace ser a que melhor se relacionaou com as variáveis do peixe, idade, comprimento e HU, -0.9, -0.9 e 0.9, respectivamente. O (·) ao lado de alguns valores significa relação estatisticamente não significativa. O (*) significa relação significativa ao nível de 0,05%.

A Tabela 10 contém os resultados da análise de regressão feita em todos os pareamentos das entropias com as variáveis idade, comprimento e HU. Nela há as informações do tipo de relacionamento das entropias com as variáveis, se linear ou exponencial (exp), o $p - valor$, o coeficiente de correlação de Spearman (ρ), o modelo, o R^2-ajustado (R^2) e o Critério de Informação de Akaike (AIC).

Tabela 10 – Informações da análise de regressão das entropias normalizadas e não normalizadas com variáveis do peixe para a espécie *Thunnus obesus*. Relacionamento das entropias normalizadas Ex_n, Ey_n e Ez_n e não normalizadas Ex, Ey e Ez com as variáveis idade, comprimento e radiodensidade (HU), avaliado por dois tipos de modelagem, linear e exponencial linearizada (exp). ρ – coeficiente de correlação de Spearman. R^2 – R^2 ajustado. AIC – Critério de Informação de Akaike. Nesta espécie, a única relação significativa pela análise de regressão foi entre a entropia Ey_n com a variável HU, com ambos os modelos explicando 70% dos dados de HU, o modelo exp foi preferível por apresentar menor AIC.

Entropias normalizadas						
entropia	Relacionamento com a idade					
	tipo	p-valor	ρ	modelo	R^2	AIC
Ex_n	linear	0.185	-0.80	Y = 140 - 175X	0.32	18.49
	exp	0.115	-0.80	log(Y) = 33 - 40X	0.49	1.34
Ey_n	linear	0.187	-0.90	Y = 145 - 172X	0.32	18.52
	exp	0.116	-0.90	log(Y) = 34 - 40X	0.49	1.36
Ez_n	linear	0.181	-0.70	Y = 68 - 73X	0.33	18.42
	exp	0.120	-0.70	log(Y) = 16 - 17X	0.48	1.42
entropia	Relacionamento com o comprimento					
	tipo	p-valor	ρ	modelo	R^2	AIC
Ex_n	linear	0.129	-0.80	Y = 2211 - 2694X	0.45	43.92
	exp	0.099	-0.80	log(Y) = 22 - 22X	0.53	-5.53
Ey_n	linear	0.130	-0.90	Y = 2288 - 2656X	0.45	43.92
	exp	0.099	-0.90	log(Y) = 22 - 21X	0.53	-5.52
Ez_n	linear	0.132	-0.70	Y = 1087 - 1119X	0.45	43.97
	exp	0.105	-0.70	log(Y) = 13 - 9X	0.52	-5.35
entropia	Relacionamento com a radiodensidade (HU)					
	tipo	p-valor	ρ	modelo	R^2	AIC
Ex_n	linear	0.051	0.80	Y= - 6721 + 17338X	0.69	58.34
	exp	0.052	0.80	log(Y) = 7 + 3X	0.69	-29.67
Ey_n	linear	0.049	0.90	Y = - 7214 + 17089X	0.70	58.20

Tabela 10 – Continuação

	exp	0.050	0.90	log(Y) = 7 + 2.6X	0.69	-29.80
Ez_n	linear	0.059	0.70	Y = 584 + 7115X	0.65	58.84
	exp	0.061	0.70	log(Y) = 8 + X	0.66	-29.20
Entropias não normalizadas						
entropia	Relacionamento com a idade					
	tipo	p-valor	ρ	modelo	R^2	AIC
Ex	linear	0.177	-0.70	Y = 4297 - 352X	0.34	18.36
	exp	0.257	-0.70	log(Y) = 783 - 64X	0.19	3.64
Ey	linear	0.155	0.40	Y = - 344 + 33X	0.39	17.99
	exp	0.128	0.40	log(Y) = - 74 + 7X	0.45	1.66
Ez	linear	0.477	-0.60	Y = 113 - 13X	-0.09	20.91
	exp	0.392	-0.60	log(Y) = 27 - 3X	0.00	4.7
entropia	Relacionamento com o comprimento					
	tipo	p-valor	ρ	modelo	R^2	AIC
Ex	linear	0.237	-0.70	Y = 55327 - 4525X	0.23	45.64
	exp	0.285	-0.70	log(Y) = 398 - 32X	0.14	-2.51
Ey	linear	0.134	0.40	Y = - 4963 + 482X	0.44	44.02
	exp	0.123	0.40	log(Y) = - 35 + 3X	0.45	-4.87
Ez	linear	0.410	-0.60	Y = 1870 - 211X	-0.02	47.05
	exp	0.369	-0.60	log(Y) = 19 - 2X	0.03	-1.85
entropia	Relacionamento com a radiodensidade (HU)					
	tipo	p-valor	ρ	modelo	R^2	AIC
Ex	linear	0.424	0.70	Y = - 218633 + 18470X	-0.04	64.42
	exp	0.418	0.70	log(Y) = - 25 + 3X	-0.03	-23.70
Ey	linear	0.117	-0.40	Y = 36145 - 2792X	0.48	60.92
	exp	0.118	-0.40	log(Y) = 13 - 0.5X	0.48	-27.16
Ez	linear	0.268	0.60	Y = - 5848 + 1517X	0.17	63.28
	exp	0.271	0.60	log(Y) = 7 + 0.2X	0.17	–24.78

Diante das informações da Tabela 10, a escolha por um melhor modelo deve-se basear no maior ρ (maior correlação entre as variáveis), maior R^2 (o que mais explica os dados) e menor AIC (maior qualidade quanto ao ajuste do modelo aos dados). Essas informações conduzem aos modelos linear $Y = -7314 + 17089X$ e ao exp $log(Y) = 7 + 2,6X$ obtidos a partir do relacionamento entre Ey_n e HU, o qual a análise de regressão teve significância estatística com p-valor $\leq 0,05$.

Dentre todos os modelos apresentados na Tabela 10, esses foram os únicos em que a análise de regressão desempenhada captou significância estatística ($p - valor \leq 0,05$), p–valor = 0.049 para o linear e p–valor = 0,050 para o exp. Explicando cerca de 70% (R^2) dos dados em ambos os dois casos e, apesar da forte correlação entre as variáveis, $\rho = 0,9$, o modelo exp mostra-se preferível por ter apresentado AIC mais baixo, -29,78, em relação ao modelo linear, 58,22, sugerindo que a HU pode ser explicada pela entropia Ey_n através de uma relação exponencial.

Na Figura 36 tem-se a análise gráfica da regressão para entropia Ey_n versus HU, caso linear. A esquerda é possível observar como fica a reta de regressão ajustada aos dados, e a direita os gráficos decorrentes da análise dos resíduos. Quanto a análise dos resíduos (quatro gráficos a direita), espera-se observar, no gráfico `Residuos vs Fitted`, pontos em torno de $x = 0$ desprovidos de padrões específicos, como requisito para que o ajuste capte a relação entre Ey_n e HU. Apesar de não parecerem seguir padrão específico, todos os pontos estão acima da reta $x = 0$, o que não permite argumentar sobre a hipótese deste gráfico, deixando alguma incerteza se o ajuste captou a relação entre as variáveis.

Para que os resultados dos testes de significância e intervalos de confiança associados aos parâmetros do modelo sejam válidos, os resíduos devem ser normalmente distribuídos, isso se confirma se o gráfico `Q-Q Residuals` apresentar pontos em torno de sua reta. Além dessa hipótese ser confirmada pelo gráfico, o teste de Shapiro com estatística do teste $W = 0,86592$ e p–valor = 0,2503 > 0,05 confirma tal hipótese quantitativamente.

O gráfico `Scale-Location` diz respeito a hipótese de homocedásticidade (variância constante dos resíduos), que é confirmada caso seja observado os pontos nesse gráfico sem um padrão de comportamento. Como os dados são poucos, um teste de Bartlett confirma tal hipótese ao apresentar Bartlett's K-squared = 1,6457, df = 1, p–valor = 0,1995 > 0,05.

O gráfico `Residuals vs Laverage` apresenta os cortes da distância de Cook, a qual avalia a influência de pontos individuais sobre as estimativas dos coeficientes do modelo. Essa análise diz que é conveniente avaliar pontos acima da linha em 0,5 e pontos acima da linha em 1 devem ser analisados com cautela, sobre a hipótese de serem *outliers*. Nesse caso, o ponto 1, acima da linha em 1, refere-se ao otólito TO1, o qual sua HU é a única com média na casa dos 6.800 HUs, enquanto os demais estão entre 6.500 HUs e 6.670 HUs (ver Tabela 2), logo tal observação para o otólito TO1 não deve ser considerada como um *outlier* e sim como uma observação de média relativamente distante das demais. O ponto 4, acima de 0,5, refere-se ao otólito TO4. Esses resultados levantados para esses dois otólitos fazem sentido, uma vez que eles são os dois pontos mais distantes na reta de regressão, justificando o motivo de terem sido chamados à atenção no gráfico da distância de Cook.

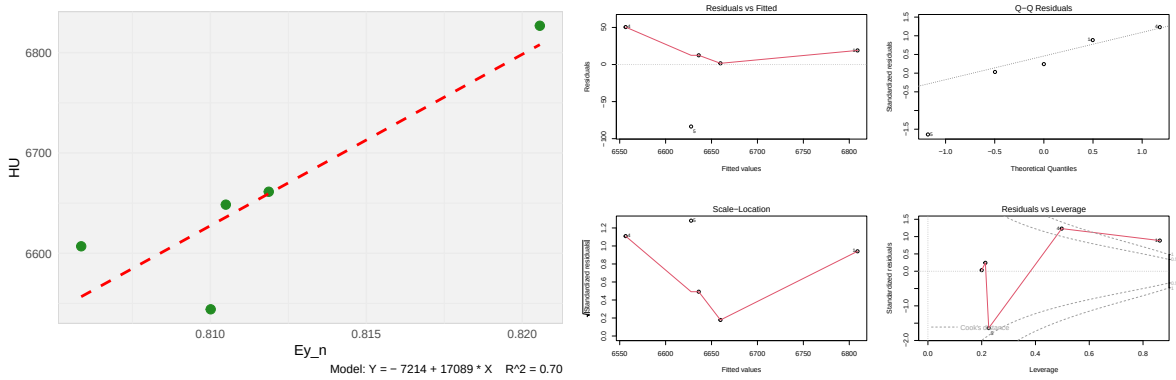


Figura 36 – Análise gráfica da regressão (caso linear) em dados dos otólitos da espécie *Thunnus obesus*. A esquerda o gráfico com a reta de regressão ajustada aos dados. A direita gráficos da análise dos resíduos: o `Residuos vs Fitted` deixa incertezas quanto ao ajuste de E_y e radiodensidade HU, ao serem observados pontos apenas acima da reta $x = 0$. O gráfico `Q-Q Residuals` indica normalidade dos resíduos devido aos pontos se concentrarem em torno da reta do gráfico. A hipótese de homocedasticidade dos resíduos é verificada ao observar que os pontos do gráfico `Scale-Location` não apresentam padrão de comportamento. A análise gráfica dos resíduos é concluída com a observação dos pontos influentes através da distância de Cook a partir do gráfico `Residuals vs Laverage`, a qual, os dois pontos acima das linhas em 0,5 e em 1 não são *outliers* mas apenas são as maiores HUs dos cinco otólitos da espécie *Thunnus obesus*.

A Figura 37 exibe o resultado gráfico da análise de regressão do relacionamento da entropia Ey_n vs HU (caso exp). No geral observa-se uma modelo com capacidade de explicação de HU relativamente próxima ao modelo linear visto anteriormente (Figura 36). O detalhe é que o AIC foi menor para este caso exp, porém ambos os p–valores estão próximos a 0,05, apesar disso não ser um inconveniente para o resultado, um conjunto de dados maior ajudaria a comprovar essa associação, o que pode resultar em uma maior explicação (R^2-ajustado) da variável resposta para além dos 70% neste caso exp. A análise dos resíduos para o caso exponencial pode herdar a discussão já vista no caso linear (ver Figuras 36 e 37), uma vez que elas também tiveram resultados similares.

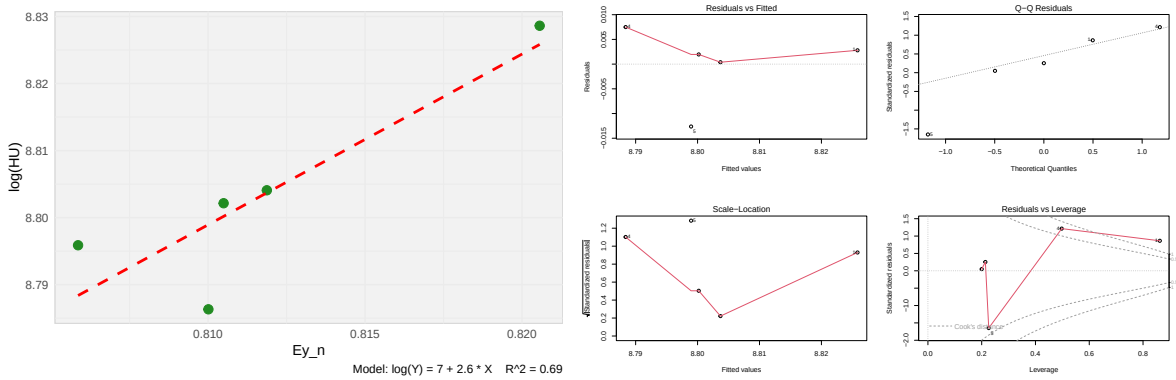


Figura 37 – Análise gráfica da regressão (caso exp) em dados dos otólitos da espécie *Thunnus obesus*. A esquerda gráfico com a reta de regressão ajustada aos dados de log(HU). A direita os gráficos da análise dos resíduos: o `Residuos vs Fitted` revela que Ey_n se ajusta aos dados de radiodensidade HU ao serem observados pontos abaixo e acima da reta $x = 0$. o gráfico `Q-Q Residuals` indica a normalidade dos resíduos devido aos pontos se concentrarem em torno da reta do gráfico. A hipótese de homocedasticidade dos resíduos é verificada ao observar que os pontos do gráfico `Scale-Location` não apresentam padrão de comportamento. Por fim, com a observação dos pontos influentes através da distância de Cook a partir do gráfico `Residuals vs Laverage` a análise é concluída. Observa-se uma análise de resíduos similar ao caso linear, indicando que os dois modelos podem explicar os dados de HU.

Em termos gerais sobre as relações avaliadas na Tabela 10, é provável que outras relações significativas possam aparecer sobre uma quantidade maior de dados, uma vez que fortes correlações são observadas. Essa análise ainda sugere que equações exponenciais devem melhor modelar tais relações avaliadas.

5.7.2 Análise sobre dados dos otólitos da espécie *Acanthurus coeruleus*

A matriz de dispersão e correlação na Figura 38 traz o pareamento entre as entropias e as variáveis do peixe idade, comprimento e HU para a espécie *Acanthurus coeruleus*. Para as entropias normalizadas correlações forte $(0,6 \leq |\rho| \leq 0,79)$ e muito forte $(0,8 \leq |\rho| \leq 1)$ são observadas a partir do relacionamento de Ex_n com o comprimento (-0,81) e com HU (0,81), de Ey_n com o comprimento (-0,64) e HU (0,67), e de Ez_n com a idade (-0,61) e com o comprimento (-0,64).

Para as entropias não normalizadas, é observada uma correlação forte do relacionamento de Ex com a idade (-0,71) e com HU (0,76), e de Ez com HU (-0,69). Correlação muito forte foi observada apenas entre Ex e o comprimento (-0,83). A entropia Ey teve correlação fraca ou moderada $(0,2 \leq |\rho| \leq 0,59)$ em relação as 3 variáveis avaliadas.

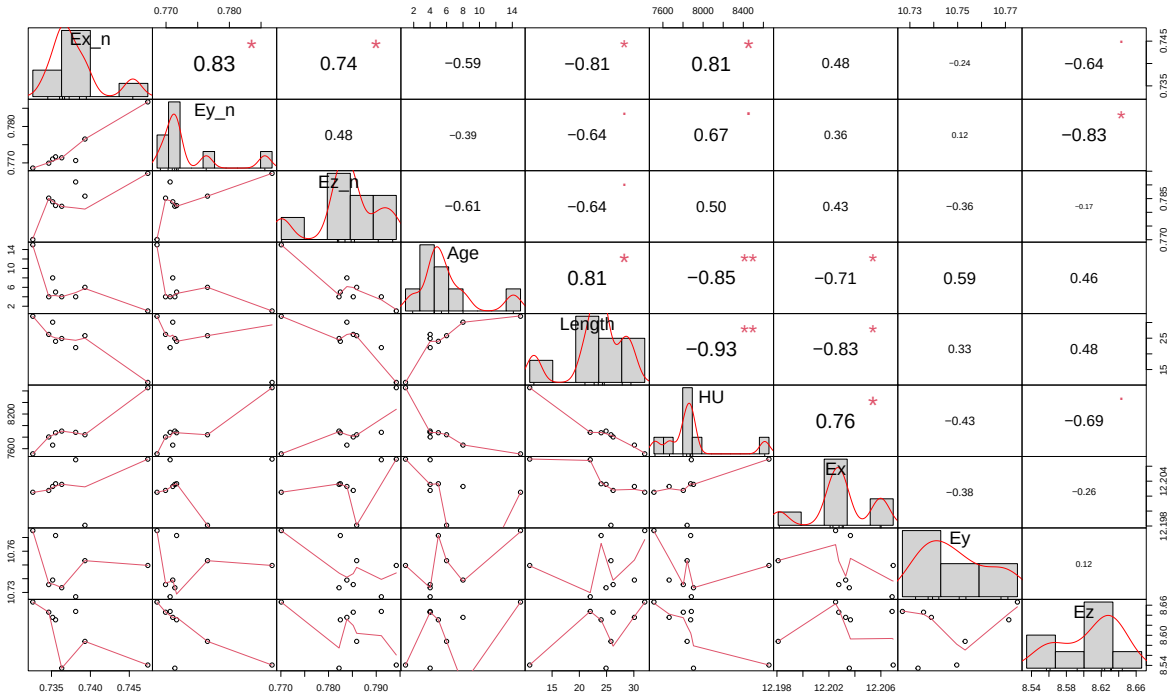


Figura 38 – Matriz de dispersão e correlação para dados dos otólitos da espécie *Acanthurus coeruleus*. Na diagonal principal, estão as entropias normalizadas Ex_n, Ey_n, Ez_n, variáveis do peixe idade (Age), comprimento (Length) e radiodensidade (HU) e por fim as entropias não normalizadas Ex, Ey e Ez. Em relação a idade, são observados relacionamentos fortes ($0,6 \leq |\rho| \leq 0,79$) com Ez_n (-0,61) e com Ex (-0,71). Em relação ao comprimento (Length) observa-se relações fortes ($0,6 \leq |\rho| \leq 0,79$) e muito fortes ($0,8 \leq |\rho| \leq 1$) com Ex_n (-0,81), Ey_n (-0,64), Ez_n (-0,64) e Ex (-0,83). Dos relacionamentos com HU, a relação é forte com E_y (0,67), Ez (-0,69) e Ex (0,76), e muito forte com Ex_n (0,81). (·) correlação estatísticamente não significativa. (*) correlação significativa ao nível de 0,05%. (**) correlação significativa ao nível de 0,01%.

A Tabela 11 mostra o resultado da análise de regressão entre as entropias e as variáveis dos peixes da espécie *Acanthurus coeruleus*. Para essa espécie, considerando as entropias normalizadas, mesmo com apenas 7 observações, todos os modelos apresentam significância estatística (p-valor ≤ 0.05), com exceção de apenas 2 casos de relacionamento, ambos lineares, que foram entre: Ex_n vs idade e Ey_n vs idade. Ainda sobre a variável idade, houve casos que explicaram pelo menos 70% das observações: Ex_n vs idade (exp) e Ex_n e Ez_n vs idade. Como esperado, os modelos mais significativos na análise são os que mais explicam a variável resposta, aqui, R^2 em pelo menos 80%. Casos: Ex_n vs comprimento e Ex_n e Ey_n vs HU, tanto linear quanto exp.

Do relacionamento das entropias não normalizadas com HU, é muito provável

estimar essa variável por qualquer uma das entropias, uma vez que possuem correlações de moderada ($0,40 \leq |\rho| \leq 0,59$) a muito forte ($0,8 \leq |\rho| \leq 1$) e explicam pelo menos 50% dos dados desta variável HU. Nesse cado as relações exponenciais são as preferíveis porque possuem menores AICs.

Ainda em relação às entropias não normalizadas, observa-se apenas um caso com significância estatística, decorrente do relacionamento de Ez com HU. Apesar de possuir correlação moderada (-0,68) essa entropia só é capaz de explicar 42% da variável HU. Isso induz, que esse relacionamento pode ter ajuste melhorado para uma amostra maior de otólitos dessa espécie.

Tabela 11 – Informações da análise de regressão das entropias normalizadas e não normalizadas com variáveis dos peixes da espécie *Acanthurus coeruleus*. Relacionamento das entropias normalizadas Ex_n, Ey_n e Ez_n e não normalizadas Ex, Ey e Ez com as variáveis idade, comprimento e radiodensidade (HU), avaliado por dois tipos de modelagem, linear e exponencial linearizada (exp). ρ – coeficiente de correlação de Spearman. R^2 – R^2 ajustado. AIC – Critério de Informação de Akaike.

Entropias normalizadas						
entropia	Relacionamento com a idade					
	tipo	p-valor	ρ	modelo	R^2	AIC
Ex_n	linear	0.068	-0.59	Y = 464 - 621X	0.36	45.74
	exp	0.005	-0.59	log(Y) = 111 - 148X	0.72	12.06
Ey_n	linear	0.150	-0.39	Y = 313 - 397X	0.20	47.56
	exp	0.017	-0.39	log(Y) = 83 - 105X	0.58	15.28
Ez_n	linear	0.003	-0.61	Y = 418 - 525X	0.76	37.62
	exp	0.008	-0.61	log(Y) = 73 - 91X	0.67	13.36
entropia	Relacionamento com o comprimento					
	tipo	p-valor	ρ	modelo	R^2	AIC
Ex_n	linear	0.001	-0.81	Y = 985 - 1303X	0.84	41.38
	exp	0.000	-0.81	log(Y) = 54 - 69X	0.87	-7.69
Ey_n	linear	0.006	-0.64	Y = 742 - 927X	0.69	46.57
	exp	0.002	-0.64	log(Y) = 43 - 51X	0.77	-3.31
Ez_n	linear	0.015	-0.64	Y = 588 - 719X	0.59	48.80
	exp	0.032	-0.64	log(Y) = 30 - 35X	0.48	3.43
entropia	Relacionamento com a radiodensidade (HU)					
	tipo	p-valor	ρ	modelo	R^2	AIC
Ex_n	linear	0.000	0.81	Y = - 43926 + 70273X	0.88	102.45

Tabela 11 – Continuação

	exp	0.000	0.81	log(Y) = 3 + 9X	0.87	-41.38
Ey_n	linear	0.001	0.67	Y = - 32615 + 52381X	0.81	105.96
	exp	0.001	0.67	log(Y) = 4 + 6X	0.80	-37.65
Ez_n	linear	0.027	0.50	Y = - 20355 + 36012X	0.51	113.69
	exp	0.024	0.50	log(Y) = 5.5 + 4.5X	0.53	-30.61
Entropias não normalizadas						
entropia	Relacionamento com a idade					
	tipo	p-valor	ρ	modelo	R^2	AIC
Ex	linear	0.334	-0.71	Y = 7265 - 595X	0.01	49.20
	exp	0.152	-0.71	log(Y) = 1887 - 154X	0.19	20.53
Ey	linear	0.151	0.59	Y = - 1410 + 132X	0.20	47.58
	exp	0.380	0.59	log(Y) = - 167 + 16X	-0.01	22.37
Ez	linear	0.117	0.46	Y = - 420 + 50X	0.25	47.01
	exp	0.071	0.46	log(Y) = - 85 + 10X	0.35	18.75
entropia	Relacionamento com o comprimento					
	tipo	p-valor	ρ	modelo	R^2	AIC
Ex	linear	0.129	-0.83	Y = 16346 - 1337X	0.23	53.93
	exp	0.125	-0.83	log(Y) = 863 - 70X	0.24	6.64
Ey	linear	0.650	0.33	Y = - 715 + 69X	-0.12	56.95
	exp	0.773	0.33	log(Y) = - 21 + 2X	-0.15	9.9
Ez	linear	0.099	0.48	Y = - 650 + 78 X	0.27	53.32
	exp	0.105	0.48	log(Y) = - 31 + 4X	0.27	6.25
entropia	Relacionamento com a radiodensidade (HU)					
	tipo	p-valor	ρ	modelo	R^2	AIC
Ex	linear	0.192	0.76	Y = - 753033 + 62353X	0.14	118.27
	exp	0.195	0.76	log(Y) = - 84 + 8X	0.14	-25.72
Ey	linear	0.712	-0.43	Y = 39749 - 2964X	-0.14	120.53
	exp	0.692	-0.43	log(Y) = 13.18 - 0.4X	-0.13	-23.52
Ez	linear	0.051	-0.69	Y = 48153 - 4675X	0.41	115.26
	exp	0.049	-0.69	log(Y) = 13.96 - 0.6X	0.42	-28.86

Dar-se-á preferência a discussão da análise gráfica para o relacionamento linear e exponencial de Ex_n vs idade. A justificativa dessa escolha, é que a variável idade é de muito interesse de ser estimada em estudos de otólitos. Por outro lado, esse relacionamento possui correlação apenas moderada ($0.4 \leq |\rho| \leq 0.59$), útil para justificar a validade do modelo para poucos dados, além de ter apresentado um caso não significativo (linear, com

p-valor = 0.068) e outro significativo (exponencial, com p-valor = 0.005), sendo que eles explicam 36 e 72% (R^2-ajustado) dos dados de idade, respectivamente. Além disso o AIC indica o exponencial como modelo que melhor se ajusta aos dados de idade.

O resultado da análise gráfica da regressão para o caso Ex_n vs idade (Age) linear, para os otólitos da espécie *Acanthurus coeruleus*, pode ser visto na Figura 39. Na figura, à esquerda, está o ajuste do modelo linear aos dados e, à direita, a análise dos resíduos.

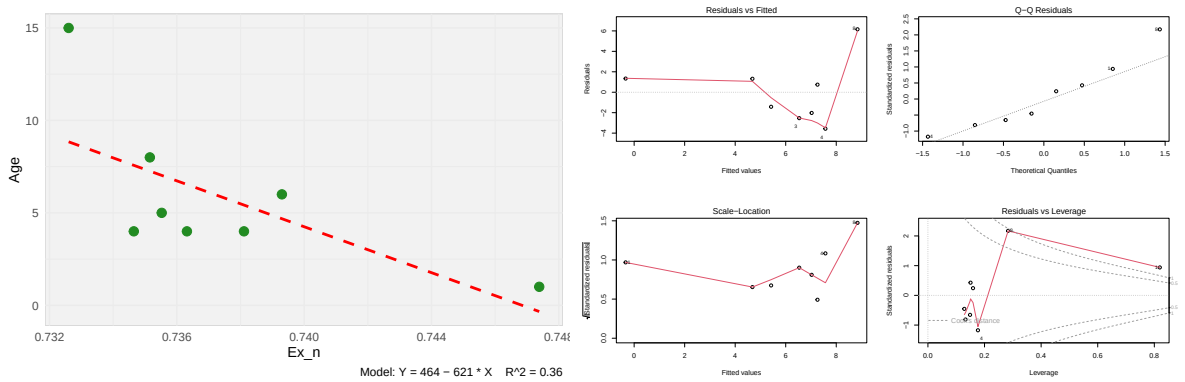


Figura 39 – Análise gráfica da regressão entre as variáveis Ex_n vs Age (caso linear) em dados dos otólitos da espécie *Acanthurus coeruleus*. À esquerda, gráfico com a reta de regressão ajustada aos dados. À direita, gráficos da análise dos resíduos: o `Residuos vs Fitted` revela que Ex_n capta algum ajuste aos dados de idade (Age) ao serem observados pontos abaixo e acima da reta $x = 0$. O gráfico `Q-Q Residuals` indica a normalidade dos resíduos devido aos pontos se concentrarem em torno da reta diagonal deste gráfico. A hipótese de homocedasticidade dos resíduos é verificada ao observar que os pontos do gráfico `Scale-Location` não apresentam comportamento específico. A partir do gráfico `Residuals vs Laverage` é possível observar os pontos influentes através da distância de Cook. Nesse gráfico, dois pontos são chamados à atenção sobre a possibilidade de serem observações extremas. O ponto 1, que está fora das distâncias de Cook (linhas tracejadas), e 8, que está próximo à linha mais distante. Sabe-se, dos dados, que esses pontos são referentes aos otólitos AC1, de 1 ano, e o otólito AC8, de 15 anos, respectivamente, dentro da espécie avaliada (Tabela 2), ou seja, são apenas as observações mais distantes de idade dentre os indivíduos, e não *outiliers*.

A Figura 40 mostra o resultado gráfico da análise de regressão para o caso exponencial (exp) entre as veriáveis Ex_n vs idade (Age). É evidente o melhor ajuste do modelo exp (Figura 40, à esquerda) em relação ao ajuste do modelo linear (Figura 39, à esquerda). Quanto à análise dos resíduos, a diferença evidente se dar no gráfico das distâncias de Cook (Figura 40, à direita), onde não revela a possibilidade de pontos extremos, o que fortalece a justificativa deste caso exp apresentar um melhor ajuste.

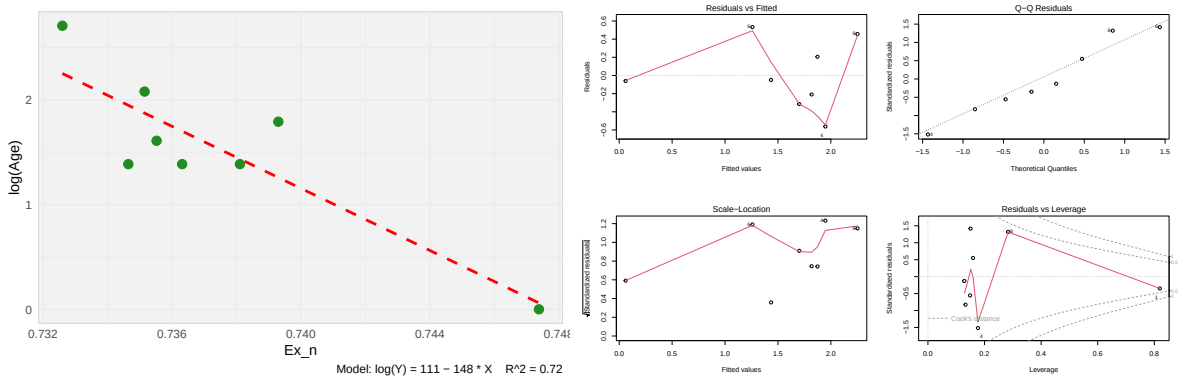


Figura 40 – Análise gráfica da regressão entre as variáveis Ex_n vs Age (caso exponencial - exp) em dados dos otólitos da espécie *Acanthurus coeruleus*. À esquerda, gráfico com a reta de regressão ajustada aos dados, onde é possível perceber um ajuste melhor que no caso linear, isso ainda pode ser fortalecido pelo R^2, que praticamente dobrou a taxa de explicação da variável resposta ao experimentar o exponencial linearizado. À direita, gráficos da análise dos resíduos: `Residuos vs Fitted` revela que Ex_n capta algum ajuste aos dados de idade (Age) ao serem observados pontos abaixo e acima da reta $x = 0$; o gráfico `Q-Q Residuals` indica a normalidade dos resíduos devido aos pontos se concentrarem em torno da reta diagonal deste gráfico. A hipótese de homocedasticidade dos resíduos é verificada ao observar que os pontos do gráfico `Scale-Location` não apresentam comportamento específico. No gráfico `Residuals vs Laverage`, referente à distância de Cook é possível observar que nenhum dos pontos foram chamados à atenção sob a possibilidade de serem considerados pontos influentes ou observações extremas, em contrário ao ajuste no caso linear, exprimindo uma real possibilidade da existência de determinação de idade pela entropia Ex_n, nessecitando apenas de uma amostra maior para comprovar tal hipótese.

6 Conclusão

Usando amostragem probabilística é possível reduzir 95% dos dados de imagens tomográficas 3D e estudar o fenômeno da variação de densidade óssea de otólitos, usando Análise Topológica de Dados (TDA), como se estivesse trabalhando nos dados brutos. Tal redução nos dados proporciona significativo ganho computacional na aplicação das técnicas Mapper e Homologia Persistente da TDA. Essa possibilidade de trabalhar com apenas 5% de resolução das imagens para estudos de densidade, revela que os parâmetros configurados no dispositivo de microtomografia computadorizada (μCT) na investigação de Vasconcelos-Filho et al. (2019) de fato conduziram a uma fina resolução para obtenção de imagens tomográficas.

Somente análises estatísticas, sobre imagens de μCT, podem levar um pesquisador de otólitos a utilizar amostras menores a 5% sob o risco de perder informações relevantes sobre a densidade otolítica. Na definição de um tamanho adequado de resolução das imagens, um procedimento de Validação Topológica de Amostras, mostra-se interessante, porque existem otólitos de estruturas irregulares (*Acanthocybium solandri*, *Thunnus albacares* e *Thunnus obesus*) e, dependendo do tamanho e formato do otólito regiões de interesse podem não ser incluídas na análise. Tal procedimento mostra-se útil então para definir uma resolução de imagem mínima capaz de tornar todos os otólitos de uma amostra comparáveis ao estudar a densidade óssea de otólitos.

O algoritmo Ball Mapper (BM) configurado com o parâmetro $\varepsilon = 100$ demonstra fornecer um resumo topologicamente fiel de imagens tomográficas de otólitos de peixes, que podem ser interpretados de forma simples em relação aos aspectos combinados de forma e densidade. Além disso, a técnica BM provou ser eficiente para identificar sujeiras e anomalias em otólitos, bem como suporte para eliminá-las, servindo também, como técnica de pré-processamento e segmentação para imagens 3D de μCT.

A técnica de Homologia Persistente, mesmo diante de uma amostra pequena de otólitos, demonstra abrir possibilidades para realizar estudos de separação de otólitos, uma vez que revelou qualitativamente as separações de classes dos otólitos presentes na amostra estudada nesta tese, ao separar os peixes da família *scombridae* dos demais. Tal separação pôde ser confirmada por uma classificação quantitativa usando o modelo de *Machine Learning*, *Random Forest*, alimentado com características topológicas extraídas a partir de diagramas de persistência obtidos dos otólitos.

Ao aumentar a quantidade de características topológicas que alimentaram o modelo, a classificação quantitativa revelou valores que sugerem separações dos otólitos pela sua forma em diferentes níveis, exibindo valores de classificação condizentes com os indivíduos da amostra, com resultados que parecem classificar os otólitos por espécie e por idade, indicando proeminência dessa técnica em estudos de classificação pela forma do otólito, o que pode ser de grande utilidade para estudos de identificação e discriminação de estoques pesqueiros.

Análises de regressão revelaram a existência de correlações de características topológicas (entropias de persistência), extraídas dos diagramas de persistência obtidos dos otólitos, com variáveis do peixe, idade, comprimento e densidade, expressando a possibilidade de estimar essas variáveis para as espécies *Thunnus obesus* e *Acanthurus coeruleus* tendo as entropias de persistência como variáveis explicativas. Tal análise demonstrou ainda que caso exista relações entre essas variáveis e as entropias, é provável que elas sejam exponenciais. Apesar de evidências dos ajustes e das explicações das variáveis do peixe pelas entropias serem consistentes, uma amostra maior seria útil para avaliar tais hipóteses.

Em geral, usando TDA, foram apresentadas evidências qualitativas (cor nos grafos e visualização da separação de classes por matriz de característica) e quantitativas (análise estatística não paramétrica, invariantes topológicos e *OOB score*) de que é possível estudar a densidade otolítica e classificar otólitos pela forma a baixo custo computacional a partir de imagens de alta resolução obtidas por tomografia computadorizada. Além disso, com características topológicas (entropias de persistência) parece possível estimar idade, comprimento e radiodensidade (HU) usando uma análise de regressão linear simples.

As revelações desta tese abrem possibilidades para aplicação de amostragem probabilística em outras áreas que possuam bancos de dados de alta dimensão, visto que este estudo apresenta uma alternativa a redução da dimensionalidade, aos estudos envolvendo TDA, ao demonstrar ser possível, obter ganhos computacionais extras ao reduzir a dimensionalidade dos dados antes da aplicação de técnicas da TDA. Possibilita ainda estimar idade de peixes e a realizar classificação de estoques pesqueiros com base na forma do otólito para várias espécies simultaneamente.

Referências

ADAMS, D. C.; ROHLF, F. J.; SLICE, D. E. Geometric morphometrics: Ten years of progress following the 'revolution'. **Italian Journal of Zoology**, v. 71, n. 1, p. 5–16, Jan 2004. ISSN 1125-0003, 1748-5851.

ADAMS, H. et al. Persistence images: A stable vector representation of persistent homology. **Journal of Machine Learning Research**, v. 18, 2017.

AHLFORS, L. V. **Complex analysis: an introduction to the theory of analytic functions of one complex variable**. 3d ed. ed. New York: McGraw-Hill, 1979. (International series in pure and applied mathematics). ISBN 978-0-07-000657-7.

AMÉZQUITA, E. J. et al. The shape of things to come: Topological data analysis and biology, from molecules to organisms. **Developmental Dynamics**, Wiley Online Library, v. 249, n. 7, p. 816–833, 2020.

ANASTASI, A. **Psychological Testing**. 4th. ed. [S.l.]: Macmillan, 1976.

ANDREWS, D. F.; HERZBERG, A. M. **Data: a collection of problems from many fields for the student and research worker**. New York: Springer-Verlag, 1985.

ASHWORTH, E. **Exploration of the relationship between somatic and otolith growth, and development of a proportionality-based back-calculation approach based on traditional growth equations**. Tese (Doutorado), 03 2016.

BAAS, N. A. et al. **Topological Data Analysis: The Abel Symposium 2018**. Cham: Springer International Publishing, 2020. v. 15. (Abel Symposia, v. 15). ISBN 978-3-030-43407-6. Disponível em: <http://link.springer.com/10.1007/978-3-030-43408-3>.

BARDARSON, H. et al. To glue or not to glue? reassembling broken otoliths for population discrimination. **Journal of Fish Biology**, v. 84, n. 5, p. 1626–1633, maio 2014. ISSN 00221112.

BERGER, R.; CASELLA, G. **Statistical Inference**. 2. ed. Florence, AL: Duxbury Press, 2001.

BOLLE, L. J. et al. Growth changes in plaice, cod, haddock and saithe in the north sea: a comparison of (post-) medieval and present-day growth rates based on otolith measurements. **Journal of sea research**, Elsevier, v. 51, n. 3-4, p. 313–328, 2004.

BONIS, T. et al. Persistence-based pooling for shape pose recognition. In: SPRINGER. **Computational Topology in Image Context: 6th International Workshop, CTIC 2016, Marseille, France, June 15-17, 2016, Proceedings 6**. [S.l.], 2016. p. 19–29.

BRADBURY, I. et al. Resolving natal tags using otolith geochemistry in an estuarine fish, rainbow smelt osmerus mordax. **Marine Ecology Progress Series**, v. 433, p. 195–204, Jul 2011. ISSN 0171-8630, 1616-1599.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

BUBENIK, P. et al. Statistical topological data analysis using persistence landscapes. **J. Mach. Learn. Res.**, v. 16, n. 1, p. 77–102, 2015.

BUZUG, T. M. **Computed Tomography**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. 311–342 p. ISBN 978-3-540-74658-4. Disponível em: <https://doi.org/10.1007/978-3-540-74658-4_16>.

CADRIN, S. X. Advances in morphometric identification of fishery stocks. **Reviews in Fish biology and Fisheries**, Springer, v. 10, p. 91–112, 2000.

CADRIN, S. X.; FRIEDLAND, K. D. The utility of image processing techniques for morphometric analysis and stock identification. **Fisheries Research**, Elsevier, v. 43, n. 1-3, p. 129–139, 1999.

CAMPANA, S. Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. **Journal of fish biology**, Wiley Online Library, v. 59, n. 2, p. 197–242, 2001.

CAMPANA, S. E. Chemistry and composition of fish otoliths: pathways, mechanisms and applications. **Marine ecology progress series**, v. 188, p. 263–297, 1999.

CAMPANA, S. E. Otolith science entering the 21st century. **Marine and freshwater research**, CSIRO Publishing, v. 56, n. 5, p. 485–495, 2005.

CAMPANA, S. E.; THORROLD, S. R. Otoliths, increments, and elements: keys to a comprehensive understanding of fish populations? **Canadian Journal of Fisheries and Aquatic Sciences**, NRC Research Press Ottawa, Canada, v. 58, n. 1, p. 30–38, 2001.

CARLSSON, G. Topology and data. **Bulletin of the American Mathematical Society**, v. 46, n. 2, p. 255–308, 2009.

CARLSSON, G. et al. Topological data analysis and machine learning theory. In: **Proc. BIRS Workshop**. [S.l.: s.n.], 2012. p. 1–11.

CARLSSON, G. et al. Persistence barcodes for shapes. In: **Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing**. New York, NY, USA: Association for Computing Machinery, 2004. (SGP '04), p. 124–135. ISBN 3905673134. Disponível em: <https://doi.org/10.1145/1057432.1057449>.

CARRIÈRE, M.; OUDOT, S. Y.; OVSJANIKOV, M. Stable topological signatures for points on 3d shapes. In: WILEY ONLINE LIBRARY. **Computer graphics forum**. [S.l.], 2015. v. 34, n. 5, p. 1–12.

CERVIGÓN, F. Los peces marinos de venezuela. **Fundación Científica Los Roques**, 1966.

CHAZAL, F. et al. Gromov-hausdorff stable signatures for shapes using persistence. In: WILEY ONLINE LIBRARY. **Computer Graphics Forum**. [S.l.], 2009. v. 28, n. 5, p. 1393–1403.

CHAZAL, F. et al. Stochastic convergence of persistence landscapes and silhouettes. In: **Proceedings of the thirtieth annual symposium on Computational geometry**. Kyoto Japan: ACM, 2014. p. 474–483. ISBN 978-1-4503-2594-3. Disponível em: <https://dl.acm.org/doi/10.1145/2582112.2582128>.

CHAZAL, F.; MICHEL, B. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. **Frontiers in Artificial Intelligence**, v. 4, p. 667963, Sep 2021. ISSN 2624-8212.

CHEN, Y.; VOLIĆ, I. Topological data analysis model for the spread of the coronavirus. **PLOS ONE**, v. 16, n. 8, p. e0255584, Aug 2021. ISSN 1932-6203.

CHITWOOD, D. H. et al. Topological mapper for 3d volumetric images. In: SPRINGER. **International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing**. [S.l.], 2019. p. 84–95.

COCHRAN, W. G. **Sampling techniques**. 3d ed. ed. New York: Wiley, 1977. (Wiley series in probability and mathematical statistics). ISBN 978-0-471-16240-7.

CONDAL, F.; GUIDA, G. Fractal analysis of otoliths contour. Unpublished, 2020. Disponível em: <http://rgdoi.net/10.13140/RG.2.2.21969.56161>.

de Almeida, P. R. C. et al. The use of the shape and chemistry of fish otoliths as a subpopulational discrimination tool for eugerres brasilianus in lagoon systems in the southwest atlantic ocean. **Fisheries Research**, v. 267, p. 106795, 2023. ISSN 0165-7836. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0165783623001881>.

DELFINADO, C. J. A.; EDELSBRUNNER, H. An incremental algorithm for betti numbers of simplicial complexes on the 3-sphere. **Computer Aided Geometric Design**, Elsevier, v. 12, n. 7, p. 771–784, 1995.

DEO, S. **Algebraic Topology**. Singapore: Springer Singapore, 2018. v. 27. (Texts and Readings in Mathematics, v. 27). ISBN 978-981-10-8734-9. Disponível em: <http://link.springer.com/10.1007/978-981-10-8734-9>.

DEY, T. K.; GUHA, S. Computing homology groups of simplicial complexes in r 3. **Journal of the ACM (JACM)**, ACM New York, NY, USA, v. 45, n. 2, p. 266–287, 1998.

DUARTE-NETO, P.; HIGA, F. M.; LESSA, R. P. Age and growth estimation of bigeye tuna, thunnus obesus (teleostei: Scombridae) in the southwestern atlantic. **Neotropical Ichthyology**, SciELO Brasil, v. 10, p. 148–158, 2012.

DUARTE-NETO, P. et al. The use of sagittal otoliths in discriminating stocks of common dolphinfish (coryphaena hippurus) off northeastern brazil using multishape descriptors. **ICES Journal of Marine Science**, v. 65, n. 7, p. 1144–1152, Oct 2008. ISSN 1095-9289, 1054-3139.

DUARTE-NETO, P. et al. Multifractal properties of a closed contour: A peek beyond the shape analysis. **PLoS ONE**, v. 9, n. 12, p. e115262, Dec 2014. ISSN 1932-6203.

DULVY, N. K. et al. Climate change and deepening of the north sea fish assemblage: a biotic indicator of warming seas. **Journal of Applied Ecology**, Wiley Online Library, v. 45, n. 4, p. 1029–1039, 2008.

DŁOTKO, P. Ball mapper: a shape summary for topological data analysis. arXiv, n. arXiv:1901.07410, Jan 2019. ArXiv:1901.07410 [math]. Disponível em: <http://arxiv.org/abs/1901.07410>.

DŁOTKO, P.; QIU, W.; RUDKIN, S. Financial ratios and stock returns reappraised through a topological data analysis lens. arXiv, n. arXiv:1911.10297, Nov 2019. ArXiv:1911.10297 [q-fin]. Disponível em: <http://arxiv.org/abs/1911.10297>.

DŁOTKO, P.; QIU, W.; RUDKIN, S. Topological data analysis ball mapper for finance. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2206.03622>.

DŁOTKO, P.; RUDKIN, S. Visualising the evolution of english covid-19 cases with topological data analysis ball mapper. arXiv, n. arXiv:2004.03282, Apr 2020. ArXiv:2004.03282 [physics, q-bio]. Disponível em: <http://arxiv.org/abs/2004.03282>.

EDELSBRUNNER; LETSCHER; ZOMORODIAN. Topological persistence and simplification. **Discrete & Computational Geometry**, Springer, v. 28, p. 511–533, 2002.

EDELSBRUNNER, H.; HARER, J. et al. Persistent homology-a survey. **Contemporary mathematics**, Providence, RI: American Mathematical Society, v. 453, n. 26, p. 257–282, 2008.

EDELSBRUNNER, H.; HARER, J. L. **Computational topology: an introduction**. [S.l.]: American Mathematical Society, 2022.

ELSDON, T. S. et al. Otolith chemistry to describe movements and life-history parameters of fishes: hypotheses, assumptions, limitations and inferences. **Oceanography and marine biology: an annual review**, v. 46, n. 1, p. 297–330, 2008.

FAROOQ, O. et al. The generalized euler characteristics of the graphs split at vertices. **Entropy**, v. 24, n. 3, p. 387, Mar 2022. ISSN 1099-4300.

FASY, B. T. et al. Confidence sets for persistence diagrams. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 42, n. 6, p. 2301 – 2339, 2014. Disponível em: <https://doi.org/10.1214/14-AOS1252>.

FISHER, M.; HUNTER, E. Digital imaging techniques in otolith data capture, analysis and interpretation. **Marine Ecology Progress Series**, v. 598, p. 213–231, 2018.

GALLARDO-CABELLO, M. et al. Analysis of the otoliths sagitta, asteristus and lapillus of yellowfin mojarra gerres cinereus (perciformes: Gerreidae) in the coast of colima and jalisco, mexico. **Open Journal of Ocean and Coastal Sciences**, v. 2, p. 18–33, 01 2015.

GAULDIE, R.; CRAMPTON, J. An eco-morphological explanation of individual variability in the shape of the fish otolith: comparison of the otolith of hoplostethus atlanticus with other species by depth. **Journal of Fish Biology**, Wiley Online Library, v. 60, n. 5, p. 1204–1221, 2002.

GAULDIE, R.; NELSON, D. Otolith growth in fishes. **Comparative Biochemistry and Physiology Part A: Physiology**, v. 97, n. 2, p. 119–135, 1990. ISSN 0300-9629. Disponível em: <https://www.sciencedirect.com/science/article/pii/030096299090159P>.

GEBREMEDHIN, S. et al. Scientific methods to understand fish population dynamics and support sustainable fisheries management. **Water**, MDPI, v. 13, n. 4, p. 574, 2021.

GEFFEN, A. Otolith ring deposition in relation to growth rate in herring (clupea harengus) and turbot (scophthalmus maximus) larvae. **Marine Biology**, Springer, v. 71, p. 317–326, 1982.

GHRIST, R. Barcodes: The persistent topology of data. **Bulletin of the American Mathematical Society**, v. 45, n. 01, p. 61–76, Oct 2007. ISSN 0273-0979.

GHRIST, R. **Elementary applied topology: edition 1.0**. s. l.: Createspace, 2014. ISBN 978-1-5028-8085-7.

GONZALEZ, R. C.; WOODS, R. E. **Digital image processing**. Fourth edition. New York, NY: Pearson, 2018. ISBN 978-0-13-335672-4.

GREEN, B. S. et al. Introduction to otoliths and fisheries in the tropics. In: _____. **Tropical Fish Otoliths: Information for Assessment, Management and Ecology**. Dordrecht: Springer Netherlands, 2009. (Reviews: Methods and Technologies in Fish Biology and Fisheries, v. 11), p. 1–22. ISBN 978-1-4020-3582-1. Disponível em: <http://link.springer.com/10.1007/978-1-4020-5775-5_1>.

HABERTHÜR, D. et al. Microtomographic investigation of a large corpus of cichlids. **PLOS ONE**, Public Library of Science, v. 18, n. 9, p. 1–11, 09 2023. Disponível em: <https://doi.org/10.1371/journal.pone.0291003>.

HAGBERG, A.; CONWAY, D. Networkx: Network analysis with python. **URL: https://networkx. github. io**, 2020.

HAIMOVICI, M. et al. Otolith atlas for marine fishes of the southwestern atlantic occurring along southern brazil (28° s-34° s). **Marine and Fishery Sciences (MAFIS)**, v. 37, 11 2023.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The elements of statistical learning: data mining, inference, and prediction**. 2nd ed. ed. New York, NY: Springer, 2009. (Springer series in statistics). ISBN 978-0-387-84857-0.

HATCHER, A. **Algebraic topology**. [S.l.]: by Cambridge University Press, 2002.

HE, T. et al. The use of otolith shape to identify stocks of redlip mullet, liza haematocheilus. **Pakistan Journal of Zoology**, v. 52, n. 6, 2020. ISSN 00309923. Disponível em: <http://researcherslinks.com/current-issues/The-Use-of-Otolith-Shape-Liza-haematocheilus/20/1/3352/html>.

HU, Y. et al. Otolith microchemistry reveals life history and habitat use of coilia nasus from the dayang river of china. **Fishes**, v. 7, n. 6, 2022. ISSN 2410-3888. Disponível em: <https://www.mdpi.com/2410-3888/7/6/306>.

HUBER, P. J. Projection pursuit. **The annals of Statistics**, JSTOR, p. 435–475, 1985.

HÜSSY, K. et al. Evaluation of otolith shape as a tool for stock discrimination in marine fishes using baltic sea cod as a case study. **Fisheries Research**, v. 174, p. 210–218, Feb 2016. ISSN 01657836.

JR, W. D. A. Field guide to fishes of the chesapeake bay. **Copeia**, The American Society of Ichthyologists and Herpetologists 810 East 10th . . . , v. 2013, n. 4, p. 784–785, 2013.

KAHN, D. W. **Topology: an introduction to the point-set and algebraic areas**. New York: Dover, 1995. ISBN 978-0-486-68609-7.

KALISH, J. M. Otolith microchemistry: validation of the effects of physiology, age and environment on otolith composition. **Journal of Experimental Marine Biology and Ecology**, v. 132, n. 3, p. 151–178, 1989. ISSN 0022-0981. Disponível em: <https://www.sciencedirect.com/science/article/pii/0022098189901263>.

KALTON, G. **Introduction to survey sampling**. Beverly Hills: Sage Publications, 1983. (Sage university papers series). ISBN 978-0-8039-2126-9.

KAMMEYER, H. **Introduction to algebraic topology**. Cham, Switzerland: Birkhäuser, 2022. (Compact Textbooks in Mathematics). ISBN 978-3-030-98313-0.

KARTHIK, R.; ABHISHEK, S. **Machine Learning Using R: With Time Series and Industry-Based Use Cases in R**. [S.l.: s.n.], 2019. v. 2. 1 p.

KERBER, M.; MOROZOV, D.; NIGMETOV, A. **Geometry helps to compare persistence diagrams**. [S.l.]: ACM New York, NY, USA, 2017.

KISH, L. **Survey sampling**. New York: Wiley, 1995. (A Wiley Interscience Publication). ISBN 978-0-471-48900-9.

KOLMOGOROV, A. Sulla determinazione empirica di una legge didistribuzione. **Giorn Dell'inst Ital Degli Att**, v. 4, p. 89–91, 1933.

ŁAWNICZAK, M. et al. The relationship between the euler characteristic and the spectra of graphs and networks. In: SPRINGER. **13th Chaotic Modeling and Simulation International Conference**. [S.l.], 2021. p. 487–497.

LAZAR, N.; RYU, H. The shape of things: Topological data analysis. **Chance**, Taylor & Francis, v. 34, n. 2, p. 59–64, 2021.

LEE, M.-S. et al. Polyketide synthase plays a conserved role in otolith formation. **Zebrafish**, v. 16, n. 4, p. 363–369, Aug 2019. ISSN 1545-8547, 1557-8542.

LESNICK, M. Studying the shape of data using topology. **The Institute Letter**, p. 10–11, 2013.

LESSA, R. et al. Otolith microstructure analysis with otc validation confirms age overestimation in atlantic thread herring opisthonema oglinum from north-eastern brazil. **Journal of fish biology**, Wiley Online Library, v. 73, n. 7, p. 1690–1700, 2008.

LI, C.; OVSJANIKOV, M.; CHAZAL, F. Persistence-based structural recognition. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2014. p. 1995–2002.

LI, M. et al. Topological data analysis as a morphometric method: Using persistent homology to demarcate a leaf morphospace. **Frontiers in Plant Science**, v. 9, 2018. ISSN 1664-462X. Disponível em: <https://www.frontiersin.org/articles/10.3389/fpls.2018.00553>.

LI, M. et al. Persistent homology and the branching topologies of plants. **American Journal of Botany**, v. 104, n. 3, p. 349–353, mar. 2017. ISSN 0002-9122, 1537-2197.

LI, W. et al. Otolith shape analysis as a tool to identify two pacific saury (cololabis saira) groups from a mixed stock in the high-seas fishing ground. **Journal of Ocean University of China**, Springer, v. 20, p. 402–408, 2021.

LIMA, E. L. **Elementos de topologia geral**. [S.l.]: Ao Livro Técnico, Editora da Universidade de São Paulo, 1970.

LIMA, E. L. **Fundamental groups and covering spaces**. Natick, MA: A K Peters, Ltd, 2003.

LIMA, E. L. **Homologia básica**. 2. ed. Rio de Janeiro, RJ: Associação Instituto Nacional de Matemática Pura e Aplicada, 2021. (Coleção Projeto Euclides). ISBN 978-85-244-0286-9.

LOMBARTE, A.; LLEONART, J. Otolith size changes related with body growth, habitat depth and temperature. **Environmental biology of fishes**, Springer, v. 37, p. 297–306, 1993.

LUM, P. Y. et al. Extracting insights from the shape of complex data using topology. **Scientific Reports**, v. 3, n. 1, p. 1236, Dec 2013. ISSN 2045-2322.

MA, G. Using topological data analysis to process time-series data: A persistent homology way. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2020. v. 1550, n. 3, p. 032082.

MANOGARAN, G.; LOPEZ, D.; CHILAMKURTI, N. In-mapper combiner based mapreduce algorithm for processing of big climate data. **Future Generation Computer Systems**, v. 86, p. 433–445, Sep 2018. ISSN 0167739X.

MARSHELL, A.; MUMBY, P. J. The role of surgeonfish (acanthuridae) in maintaining algal turf biomass on coral reefs. **Journal of Experimental Marine Biology and Ecology**, Elsevier, v. 473, p. 152–160, 2015.

MCBRIDE, R. S.; RICHARDSON, A. K.; MAKI, K. L. Age, growth, and mortality of wahoo, acanthocybium solandri, from the atlantic coast of florida and the bahamas. **Marine and Freshwater Research**, CSIRO Publishing, v. 59, n. 9, p. 799–807, 2008.

MENDOZA, R. R. Otoliths and their applications in fishery science. **Croatian Journal of Fisheries: Ribarstvo**, Agronomski fakultet Zagreb, v. 64, n. 3, p. 89–102, 2006.

MENNI, R. C.; RINGUELET, R. A.; ARÁMBURU, R. H. **Peces marinos de la Argentina y Uruguay**. [S.l.]: Editorial Hemisferio Sur, 1984.

MOORE, B. et al. Feasibility of automating otolith ageing using ct scanning and machine learning. **New Zealand Fisheries Assessment Report**, v. 58, p. 23, 2019.

MORALES-NIN, B. Review of the growth regulation processes of otolith daily increment formation. **Fisheries research**, Elsevier, v. 46, n. 1-3, p. 53–67, 2000.

MUNKRES, J. R. **Topology**. 2. ed. ed. Upper Saddle River, NJ: Prentice Hall, 2000. ISBN 978-0-13-181629-9.

MUNKRES, J. R. **Elements of algebraic topology**. Boca Raton London New York: CRC Press Taylor & Francis, 2018. (The advanced book program). ISBN 978-0-201-62728-2.

MYERS, A.; MUNCH, E.; KHASAWNEH, F. A. Persistent homology of complex networks for dynamic state detection. **Physical Review E**, v. 100, n. 2, p. 022314, Aug 2019. ISSN 2470-0045, 2470-0053.

MYERS, S. C. et al. An efficient protocol and data set for automated otolith image analysis. **Geoscience Data Journal**, v. 7, n. 1, p. 80–88, jun. 2020. ISSN 2049-6060, 2049-6060.

NAVA, E. et al. Digital imaging tool to enhance otolith microstructure for estimating age in days in juvenile and adult fish. **IEEE Journal of Oceanic Engineering**, v. 43, n. 1, p. 48–55, 2018.

NAZIR, A.; KHAN, M. A. Using otoliths for fish stock discrimination: Status and challenges. **Acta Ichthyologica et Piscatoria**, Pensoft Publishers, v. 51, n. 2, p. 199–218, 2021.

NICOLAU, M.; LEVINE, A. J.; CARLSSON, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 108, n. 17, p. 7265–7270, 2011.

NION, H. et al. Peces del uruguay: Lista sistemática y nombres comunes. Montevideo (Uruguay) DINARA/INFOPESCA, 2002.

NJ, M. **The Global Challenge for Government Transparency: The Sustainable Development Goals (SDG) 2030 Agenda**. 2015. Available from: <https://worldtop20.org/global-movement?gclid=Cj0KCQjw1vSZBhDuARIsAKZlijTFVx_wQ4oZoclAbMBqpKQowL2zjjfzQdk2cZUCMvCEfvkz-A-IbXYaAi4VEALw_wcB#>. Last accessed: may/4/2022.

NOAA'S, N. O. M. C. **What is an Otolith?** 2013. Acessado em: 12 de dezembro de 2023. Disponível em: <https://www.youtube.com/watch?v=igywmjc1MdU&list=PL8lAZbDUj3FlLOvefUR96hfcP9kF2tNos&index=1>.

NOLF, D. Studies on fossil otoliths—the state of the art. **Recent developments in fish otolith research**, University of South Carolina Press Columbia, SC, v. 19, p. 513–544, 1995.

OTTER, N. et al. A roadmap for the computation of persistent homology. **EPJ Data Science**, v. 6, n. 1, p. 17, Dec 2017. ISSN 2193-1127. ArXiv:1506.08903 [physics, q-bio].

PANFILI, J. et al. **Manual of fish sclerochronology**. [S.l.]: Ifremer-IRD coedition, 2002. ISBN 978-2-84433-067-3.

PAVLOV, D. A. Otolith morphology and relationships of several fish species of the suborder scorpaenoidei. **Journal of Ichthyology**, v. 61, n. 1, p. 33–47, jan. 2021. ISSN 1555-6425.

PAYAN, P. et al. Chemical composition of saccular endolymph and otolith in fish inner ear: lack of spatial uniformity. **American Journal of Physiology-Regulatory, integrative and comparative physiology**, American Physiological Society Bethesda, MD, v. 277, n. 1, p. R123–R131, 1999.

PAYAN, P. et al. Endolymph chemistry and otolith growth in fish. **Comptes Rendus Palevol**, Elsevier, v. 3, n. 6-7, p. 535–547, 2004.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. **the Journal of machine Learning research**, JMLR. org, v. 12, p. 2825–2830, 2011.

PEREA, J. A. Topological time series analysis. **Notices of the American Mathematical Society**, American Mathematical Society, AMS, v. 66, n. 5, p. 686–694, 2019.

PEREA, J. A.; HARER, J. Sliding windows and persistence: An application of topological methods to signal analysis. **Foundations of Computational Mathematics**, Springer, v. 15, p. 799–838, 2015.

PIERA, J. et al. Otolith shape feature extraction oriented to automatic classification with open distributed data. **Marine and Freshwater Research**, CSIRO Publishing, v. 56, n. 5, p. 805–814, 2005.

POPPER, A. N.; FAY, R. R. Sound detection and processing by fish: critical review and major research questions (part 2 of 2). **Brain, behavior and evolution**, S. Karger AG Basel, Switzerland, v. 41, n. 1, p. 26–38, 1993.

POPPER, A. N.; HOXTER, B. The fine structure of the sacculus and lagena of a teleost fish. **Hearing Research**, Elsevier, v. 5, n. 2-3, p. 245–263, 1981.

POPPER, A. N.; RAMCHARITAR, J.; CAMPANA, S. E. Why otoliths? insights from inner ear physiology and fisheries biology. **Marine and freshwater Research**, CSIRO Publishing, v. 56, n. 5, p. 497–504, 2005.

QIU, W.; RUDKIN, S.; DŁOTKO, P. Refining understanding of corporate failure through a topological data analysis mapping of altman's z-score model. **Expert Systems with Applications**, v. 156, p. 113475, Oct 2020. ISSN 09574174.

QUÉRO, J.-C. et al. Check-list of the fishes of the eastern tropical atlantic: Clofeta. Lisbon (Portugal) Junta Nacional de Investigacao Cientifica e Tecnologica, 1990.

RABADÁN, R.; BLUMBERG, A. J. **Topological Data Analysis for Genomics and Evolution: Topology in Biology**. 1. ed. Cambridge University Press, 2019. ISBN 978-1-316-67166-5. Disponível em: <https://www.cambridge.org/core/product/identifier/9781316671665/type/book>.

REANI, Y.; BOBROWSKI, O. A coupled alpha complex. **ArXiv**, abs/2105.08113, 2021. Disponível em: <https://api.semanticscholar.org/CorpusID:234763222>.

REEB, G. Sur les points singuliers d'une forme de pfaff completement integrable ou d'une fonction numerique [on the singular points of a completely integrable pfaff form or of a numerical function]. **Comptes Rendus Acad. Sciences Paris**, v. 222, p. 847–849, 1946.

REININGHAUS, J. et al. A stable multi-scale kernel for topological machine learning. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2015. p. 4741–4748.

REY, J. et al. Otolith microstructure analysis reveals differentiated growth histories in sympatric black hakes (merluccius polli and merluccius senegalensis). **Fisheries Research**, Elsevier, v. 179, p. 280–290, 2016.

ROBINS, V. Towards computing homology from finite approximations. In: **Topology proceedings**. [S.l.: s.n.], 1999. v. 24, n. 1, p. 503–532.

RUCCO, M. et al. Characterisation of the idiotypic immune network through persistent entropy. In: SPRINGER. **Proceedings of ECCS 2014: European Conference on Complex Systems**. [S.l.], 2016. p. 117–128.

RUZZANTE, D. E. et al. Biocomplexity in a highly migratory pelagic marine fish, atlantic herring. **Proceedings of the Royal Society B: Biological Sciences**, v. 273, n. 1593, p. 1459–1464, jun. 2006. ISSN 0962-8452, 1471-2954.

SAGGAR, M. et al. Towards a new approach to reveal dynamical organization of the brain using topological data analysis. **Nature Communications**, v. 9, n. 1, p. 1399, Dec 2018. ISSN 2041-1723.

SALIMI, N. et al. Fully-automated identification of fish species based on otolith contour: using short-time fourier transform and discriminant analysis (stft-da). **PeerJ**, PeerJ Inc., v. 4, p. e1664, 2016.

SASAGAWA, T.; MUGIYA, Y. Biochemical properties of water-soluble otolith proteins and the immunobiochemical detection of the proteins in serum and various tissues in the tilapia oreochromis niloticus. **Fisheries science**, The Japanese Society of Fisheries Science, v. 62, n. 6, p. 970–976, 1996.

SCHULZ-MIRBACH, T. et al. Enigmatic ear stones: what we know about the functional role and evolution of fish otoliths: The role of fish otoliths in inner ear function. **Biological Reviews**, v. 94, n. 2, p. 457–482, Apr 2019. ISSN 14647931.

SCHWARZHANS, W. Sexual and ontogenetic dimorphism in otoliths of the family ophidiidae. **Cybium**, Société française d'ichtyologie, v. 18, n. 1, p. 71–98, 1994.

SILVA, N. N. d. **Amostragem probabilística: um curso introdutório**. São Paulo: EDUSP, 1998. ISBN 978-85-314-0423-8.

SILVA, V. de; MOROZOV, D.; VEJDEMO-JOHANSSON, M. Topological persistence and simplification. In: IEEE. **Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science**. [S.l.], 2004. p. 454–463.

SINGH, G.; MEMOLI, F.; CARLSSON, G. Topological methods for the analysis of high dimensional data sets and 3d object recognition. **Eurographics Symposium on Point-Based Graphics**, The Eurographics Association, p. 10 pages, 2007. ISSN 1811-7813.

SINGH, N. et al. Topological descriptors of histology images. In: SPRINGER. **Machine Learning in Medical Imaging: 5th International Workshop, MLMI 2014, Held in Conjunction with MICCAI 2014, Boston, MA, USA, September 14, 2014. Proceedings 5**. [S.l.], 2014. p. 231–239.

SMIRNOV, N. V. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. **Bull. Math. Univ. Moscou**, v. 2, n. 2, p. 3–14, 1939.

SMITH, C. **Field guide to tropical marine fishes of the Caribbean, the Gulf of Mexico, Florida, the Bahamas, and Bermuda. National Audubon Society**. [S.l.]: New York: Chanticleer Press Edition, 1997.

SOMKUNWAR, R.; VAZE, V. M. A comparative study of graph isomorphism applications. **International Journal of Computer Applications**, Foundation of Computer Science, v. 162, n. 7, p. 34–37, 2017.

SPEARMAN, C. The proof and measurement of association between two things. Appleton-Century-Crofts, 1961.

SPONAUGLE, S. Otolith microstructure reveals ecological and oceanographic processes important to ecosystem-based management. **Environmental Biology of Fishes**, Springer, v. 89, p. 221–238, 2010.

STÉQUERT, B.; PANFILI, J.; DEAN, J. M. Age and growth of yellowfin tuna, thunnus albacares, from the western indian ocean, based on otolith microstructure. **Oceanographic Literature Review**, v. 12, n. 43, p. 1275, 1996.

STRANSKY, C. Chapter seven - morphometric outlines. In: CADRIN, S. X.; KERR, L. A.; MARIANI, S. (Ed.). **Stock Identification Methods (Second Edition)**. Second edition. San Diego: Academic Press, 2014. p. 129–140. ISBN 978-0-12-397003-9. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780123970039000072>.

SVETOCHEVA, O.; STASENKOVA, N.; FUKS, G. **Guide to the bony fishes otoliths of the white sea**. [S.l.: s.n.], 2007. v. 3. 1–46 p. ISBN 1502-8828.

TAUZIN, G. et al. giotto-tda: A topological data analysis toolkit for machine learning and data exploration. **The Journal of Machine Learning Research**, JMLRORG, v. 22, n. 1, p. 1834–1839, 2021.

TORRES, G. J.; LOMBARTE, A.; MORALES-NIN, B. Sagittal otolith size and shape variability to identify geographical intraspecific differences in three species of the genus merluccius. **Journal of the Marine Biological Association of the United Kingdom**, Cambridge University Press, v. 80, n. 2, p. 333–342, 2000.

TUCKEY, N. P. et al. Automated image analysis as a tool to measure individualised growth and population structure in chinook salmon (oncorhynchus tshawytscha). **Aquaculture, Fish and Fisheries**, Wiley Online Library, v. 2, n. 5, p. 402–413, 2022.

UN. **Transforming our world: 2030 Agenda for sustainable Development**. 2015. Available from: <https://sdgs.un.org/2030agenda>. Last accessed: may/4/2022. Department of Economic and Social Affairs.

VAN, N. Q.; Q.T, V.; V.D, T. Otolith shape utilization for the stock identification of caroun croaker, johnius carouna (cuvier, 1830) (perciformes: Sciaenidae), from the east vietnam sea. **Iranian Journal of Fisheries Sciences**, v. 21, p. 864–879, 08 2022.

VASCONCELOS-FILHO, J. E. et al. Peeling the otolith of fish: Optimal parameterization for micro-ct scanning. **Frontiers in Marine Science**, v. 6, p. 728, Nov 2019. ISSN 2296-7745.

WALLACE, A. H. **An introduction to algebraic topology**. [S.l.]: Courier Corporation, 2011.

WANG, Y.; DICOSIMO, J. **National Observer Program 2016 Fishery Observer Attitudes and Experiences Survey**. US Department of Commerce, National Oceanic and Atmospheric Administration – NOAA, 2019. Disponível em: <https://spo.nmfs.noaa.gov/sites/default/files/TMSPO186.pdf>.

WASSERMAN, L. Topological data analysis. **Annual Review of Statistics and Its Application**, v. 5, n. 1, p. 501–532, Mar 2018. ISSN 2326-8298, 2326-831X.

Wikipedia contributors. **Reeb graph**. 2024. Wikipedia, The Free Encyclopedia. Page name: Reeb graph. Disponível em: <https://en.wikipedia.org/w/index.php?title=Reeb_graph&oldid=1193572595>.

WU, D. et al. Mechanistic basis of otolith formation during teleost inner ear development. **Developmental cell**, Elsevier, v. 20, n. 2, p. 271–278, 2011.

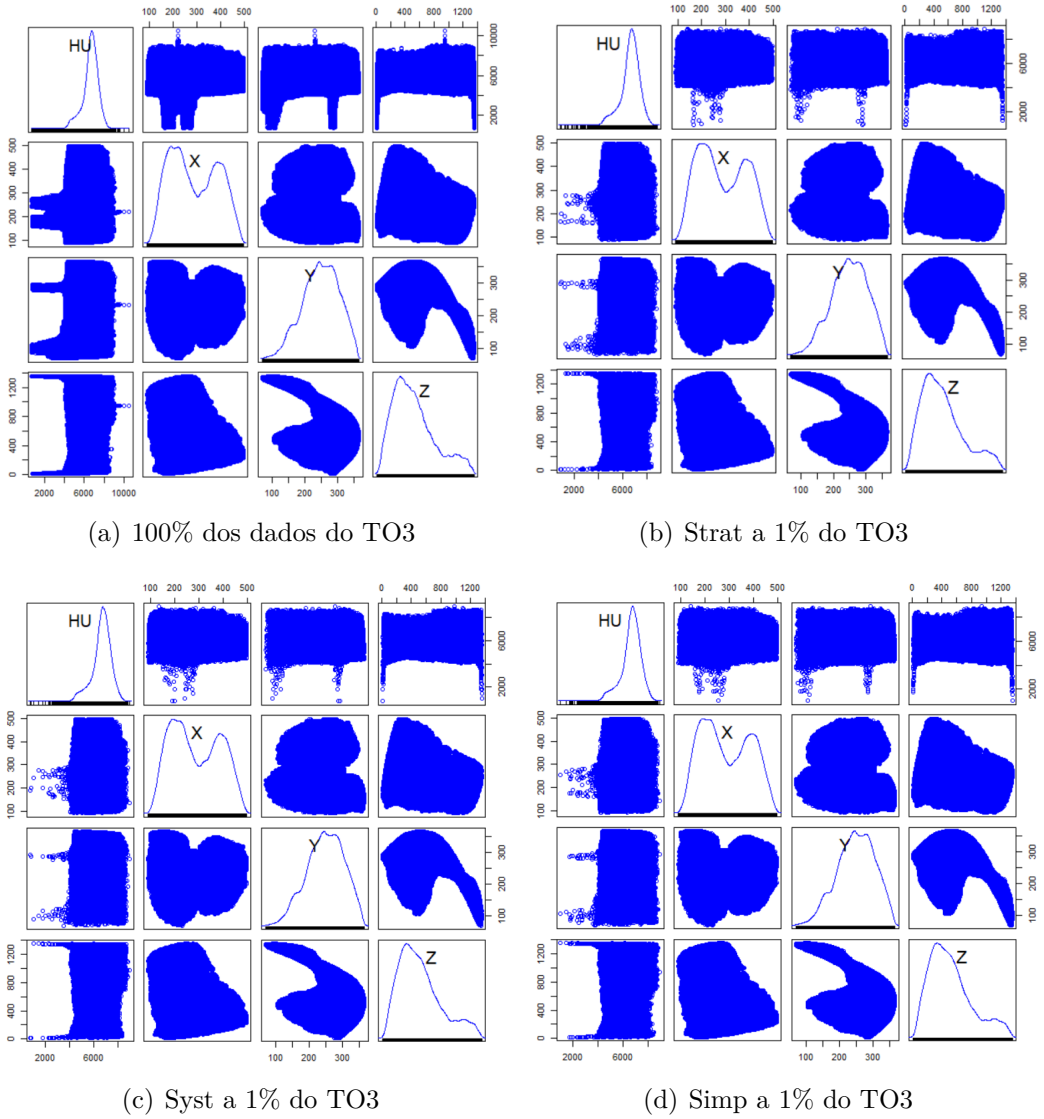
ZOMORODIAN, A.; CARLSSON, G. Computing persistent homology. **Discrete & Computational Geometry**, v. 33, n. 2, p. 249–274, fev. 2005. ISSN 0179-5376, 1432-0444.

APÊNDICE A – Algoritmo usado para a extração dos voxels e valores de HU

```
        library("Momocs")
        library("jpeg")
        library("tiff")
        library("autothresholdr")
        library("ijtiff")

setwd(choose.dir())

  Imagens <- list.files(pattern = ".tif")
  filename <- noquote(basename(getwd()))
  Names <- tools::file_path_sans_ext(basename(Imagens))
  dpi = 300

        Output <- matrix(NA, ncol=4)

for (I in 1:length(Imagens)) {
        Img <- ijtiff::read_tif(Imagens[I])[,,1,1]
        x2 <- as.vector(Img)
        A <- subset(x2, x2 < mean(subset(x2,
        x2 > quantile(x2, 0.99))))
        B <- subset(x2,
        x2 > mean(subset(x2, x2 > quantile(x2, 0.99))))

threshold <- mean(A, na.rm = T)
        + ((mean(B, na.rm = T) - mean(A, na.rm = T))) / 2

        if (threshold < 500 | is.nan(threshold)) {
        next
                }

```

```
31          if (threshold > 500) {
32                  Img2 <- ifelse(Img < threshold, NA, 1)
33                  Img3 <- Img * Img2
34
35  Tab <- cbind(which(!is.na(Img3), arr.ind = T),
36          rep(I, nrow(which(!is.na(Img3), arr.ind = T))),
37                      Img3[which(!is.na(Img3), arr.ind = T)])
38
39                          Output <- rbind(Output, Tab)
40                  }
41
42                  cat("\014")
43                  progress(I / length(Imagens) * 100)
44
45                  if (I == length(Imagens)) {
46                          cat("Done!\n")
47                          beep()
48                  }
49          }
50
51          Output <- Output[-1,]
52
53          colnames(Output) <- c("X", "Y", "Z", "HU")
54
55          Nome <- tools::file_path_sans_ext(basename(getwd()))
56          write.table(Output, paste("", Nome, ".txt", sep=""),
57          quote = F, sep = "\t", row.names = F, col.names = T)
```

APÊNDICE B – Matrizes de Dispersão para Samps a 1% do otólito TO3



(a) 100% dos dados do TO3

(b) Strat a 1% do TO3

(c) Syst a 1% do TO3

(d) Simp a 1% do TO3

Figura 41 – Matrizes de dispersão do TO3 (a) dos dados brutos e (b-d) das amostras a 1% proveniente dos diferentes sistemas de amostragem, Strat, Syst e Simp respectivamente. Independentemente do sistema de amostragem empregado (b-c), são observadas as mesmas distribuições de probabilidade quando comparadas com a matriz de dispersão dos dados brutos (a). Indicando que amostras a 1% podem ser representativas da Pop.

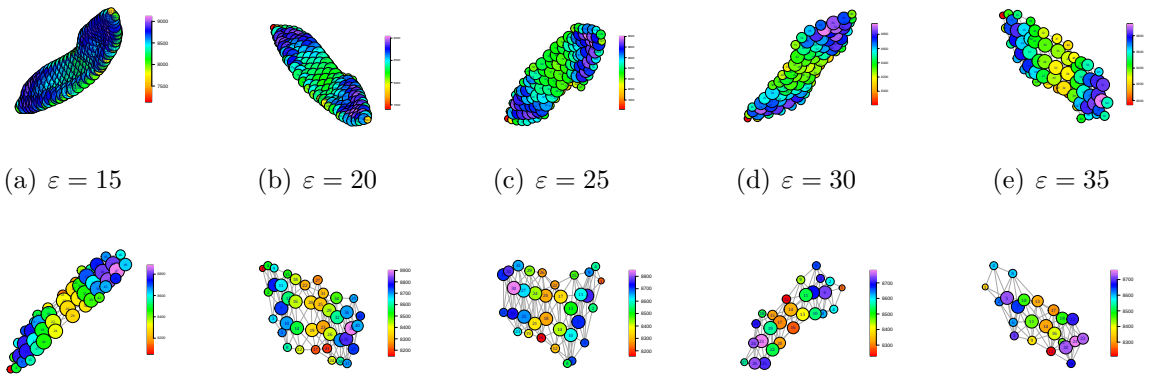
APÊNDICE C – Tabela de frequência dos voxels do otólito TO3

Tabela 12 – Tabela de frequência mostrando a distribuição agrupada de voxels ao longo das fatias Z (imagens 2D) para Pop e amostras a 5% extraídas a partir dos três sistemas de amostragem aplicados sobre os dados da imagem do otólito TO3. As amostragens Strat e Syst mostraram-se equivalentes em proporção à Pop, o que sugere representatividade em distribuição. A Simp apresenta perdas de voxels ao longo das fatias Zs, o que leva a consequente perda dos valores de densidade referentes àqueles voxels não amostrados.

Intervalo de	Pop		Strat		Syst		Simp	
classe dos Zs	voxels	%	voxels	%	voxels	%	voxels	%
0 ⊣ 50	56070	0,24	2801	0,24	2804	0,24	2729	0,24
50 ⊣ 100	396599	1,73	19825	1,73	19830	1,73	19798	1,73
100 ⊣ 150	813022	3,54	40650	3,54	40651	3,54	40649	3,54
150 ⊣ 200	1154246	5,03	57710	5,03	57712	5,03	58001	5,06
200 ⊣ 250	1456300	6,35	72817	6,35	72815	6,35	72532	6,32
250 ⊣ 300	1726947	7,53	86348	7,53	86348	7,53	86785	7,57
300 ⊣ 350	1860257	8,11	93014	8,11	93012	8,11	93154	8,12
350 ⊣ 400	1816758	7,92	90841	7,92	90838	7,92	90798	7,91
400 ⊣ 450	1715558	7,48	85783	7,48	85778	7,48	85765	7,48
450 ⊣ 500	1635852	7,13	81792	7,13	81793	7,13	81850	7,13
500 ⊣ 550	1571408	6,85	78572	6,85	78570	6,85	78418	6,84
550 ⊣ 600	1416025	6,17	70802	6,17	70802	6,17	70791	6,17
600 ⊣ 650	1208151	5,27	60408	5,27	60407	5,27	60409	5,27
650 ⊣ 700	975671	4,25	48783	4,25	48784	4,25	48554	4,23
700 ⊣ 750	815404	3,55	40775	3,55	40770	3,55	40937	3,57
750 ⊣ 800	665208	2,90	33263	2,90	33260	2,90	33231	2,90
800 ⊣ 850	543390	2,37	27172	2,37	27170	2,37	27034	2,36
850 ⊣ 900	444606	1,94	22226	1,94	22230	1,94	22184	1,93
900 ⊣ 950	336331	1,47	16815	1,47	16817	1,47	16825	1,47
950 ⊣ 1000	341672	1,49	17082	1,49	17083	1,49	17270	1,51
1000 ⊣ 1050	330361	1,44	16522	1,44	16518	1,44	16304	1,42
1050 ⊣ 1100	358670	1,56	17936	1,56	17934	1,56	17934	1,56
1100 ⊣ 1150	363709	1,59	18181	1,58	18185	1,59	18173	1,58
1150 ⊣ 1200	323216	1,41	16158	1,41	16161	1,41	16077	1,40
1200 ⊣ 1250	290991	1,27	14547	1,27	14550	1,27	14617	1,27
1250 ⊣ 1300	225633	0,98	11282	0,98	11281	0,98	11271	0,98
1300 ⊣ 1350	100930	0,44	5044	0,44	5047	0,44	5059	0,44
1350 ⊣ 1400	690	0,00	34	0,00	34	0,00	35	0,00
Total	22943675	100.0	1147184	100.0	1147184	100.0	1147184	100.0

APÊNDICE D – VTA das outras espécies

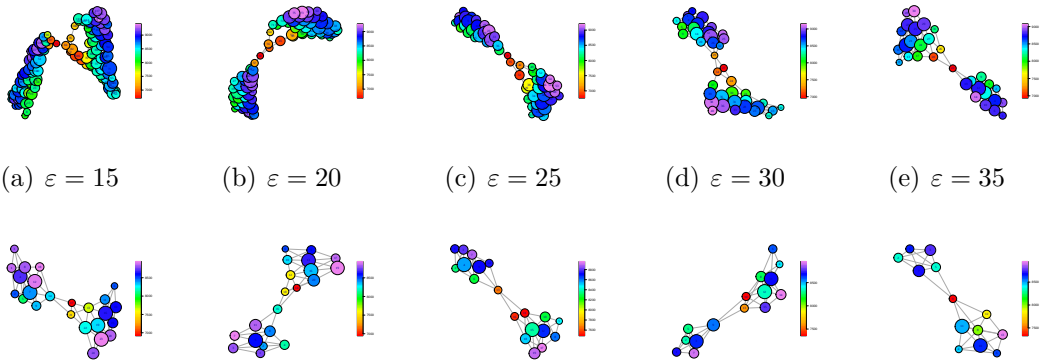
Figura 42 – VTA para Strat a 5% do OO1 - Z=325



(a) $\varepsilon = 15$ (b) $\varepsilon = 20$ (c) $\varepsilon = 25$ (d) $\varepsilon = 30$ (e) $\varepsilon = 35$

(f) $\varepsilon = 40$ (g) $\varepsilon = 45$ (h) $\varepsilon = 50$ (i) $\varepsilon = 55$ (j) $\varepsilon = 60$

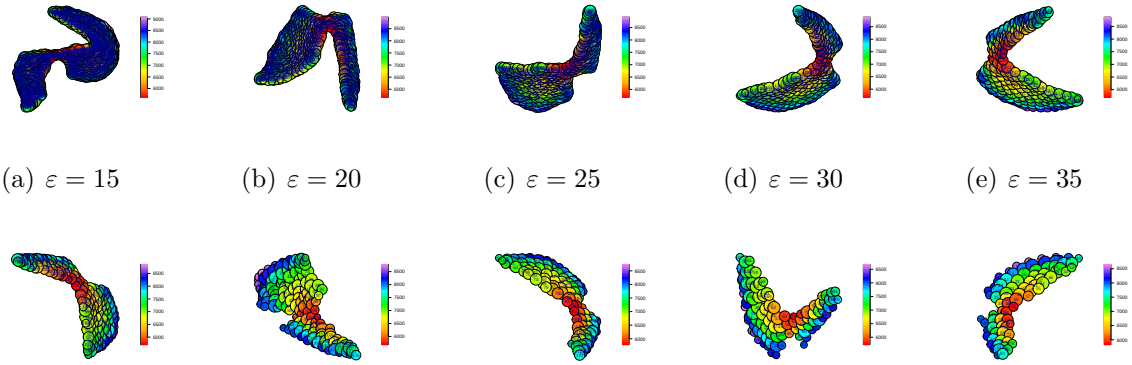
Figura 43 – VTA para Strat a 5% do TA - Z=425



(a) $\varepsilon = 15$ (b) $\varepsilon = 20$ (c) $\varepsilon = 25$ (d) $\varepsilon = 30$ (e) $\varepsilon = 35$

(f) $\varepsilon = 40$ (g) $\varepsilon = 45$ (h) $\varepsilon = 50$ (i) $\varepsilon = 55$ (j) $\varepsilon = 60$

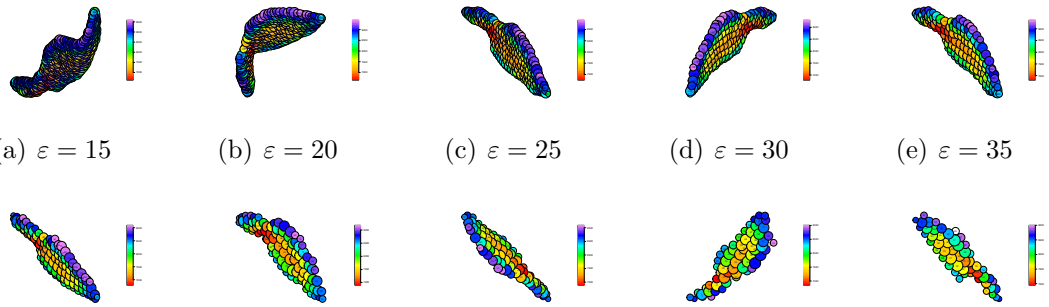
Figura 44 – VTA para Strat a 5% do AC42 - Z=75



(a) $\varepsilon = 15$ (b) $\varepsilon = 20$ (c) $\varepsilon = 25$ (d) $\varepsilon = 30$ (e) $\varepsilon = 35$

(f) $\varepsilon = 40$ (g) $\varepsilon = 45$ (h) $\varepsilon = 50$ (i) $\varepsilon = 55$ (j) $\varepsilon = 60$

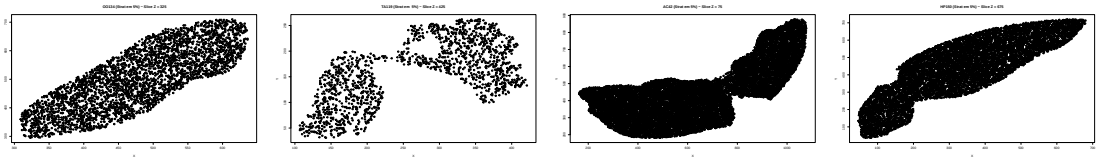
Figura 45 – VTA para Strat a 5% do HP1 - Z=625



(a) $\varepsilon = 15$ (b) $\varepsilon = 20$ (c) $\varepsilon = 25$ (d) $\varepsilon = 30$ (e) $\varepsilon = 35$

(f) $\varepsilon = 40$ (g) $\varepsilon = 45$ (h) $\varepsilon = 50$ (i) $\varepsilon = 55$ (j) $\varepsilon = 60$

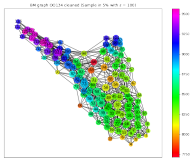
Figura 46 – Slices de referência usados no VTA das demais espécies



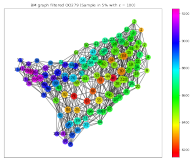
(a) OO1 - Slice Z = 325 (b) TA - Slice Z = 425 (c) AC2 - Slice Z = 75 (d) HP1 - Slice Z = 625

APÊNDICE E – Topologia dos otólitos

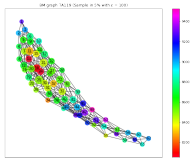
Figura 47 – Topologia representando a densidade para os demais otólitos



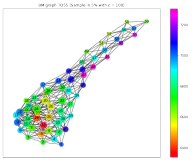
(a) OO1



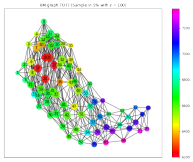
(b) OO2



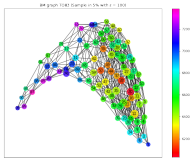
(c) TA



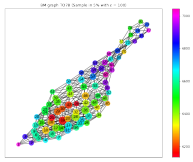
(d) TO1



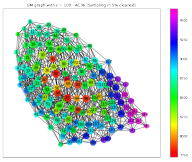
(e) TO2



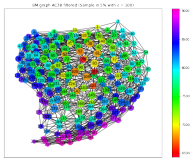
(f) TO4



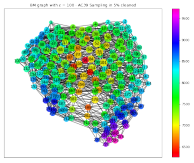
(g) TO5



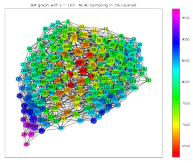
(h) AC1



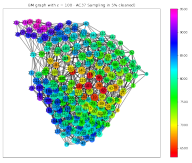
(i) AC2



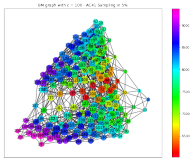
(j) AC3



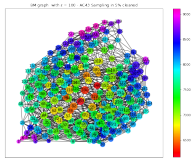
(k) AC4



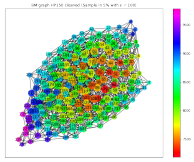
(l) AC5



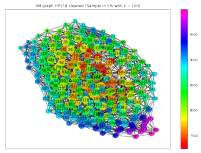
(m) AC6



(n) AC7



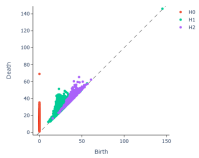
(o) HP1



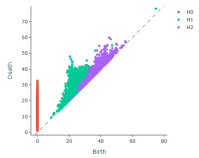
(p) HP2

APÊNDICE F – Homologia dos otólitos

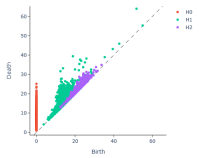
Figura 48 – Diagramas de persistência dos otólitos



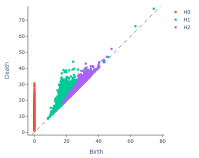
(a) OO1



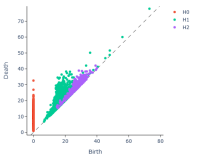
(b) OO2



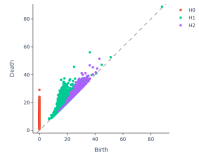
(c) AS1



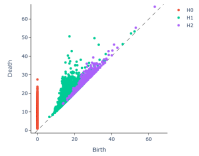
(d) AS2



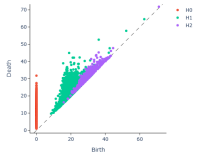
(e) AS3



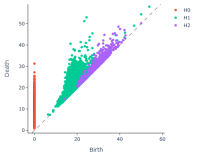
(f) TA



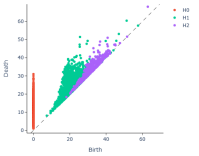
(g) TO1



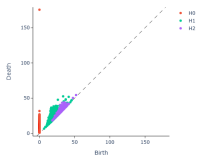
(h) TO2



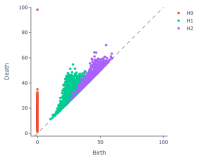
(i) TO3



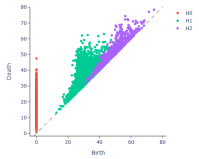
(j) TO4



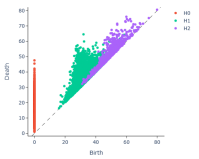
(k) TO5



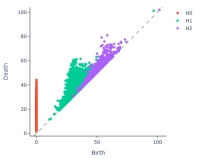
(l) AC1



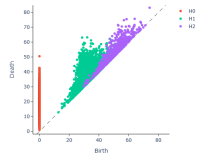
(m) AC2



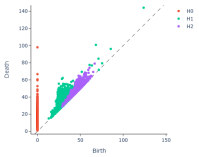
(n) AC3



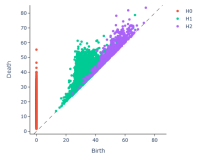
(o) AC4



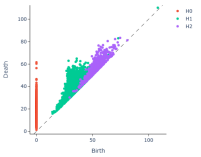
(p) AC5



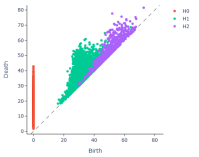
(q) AC6



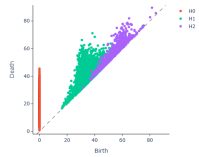
(r) AC7



(s) AC8



(t) HP1



(u) HP2