ILAGOR, JOCELYN, M.A. Visualizing Arctic Ice Data with Optimal Transport. (2023)
Directed by Dr. Dr. Thomas Weighill. 24 pp.

Climate change is a growing concern that causes sea level risings, heatwaves, and loss of habitats, and impacts the ecology. The climate change in the Arctic is specifically important as the Arctic helps reflect a significant amount of sunlight. This thesis applies Optimal Transport (OT) and Topological Data Analysis (TDA) to analyse the arctic ice data collected by NASA in 1999-2009. OT and TDA are fields in mathematics that consist of computational methods that study the shape and distribution of data. Our method is based on Wasserstein distance, a geometry-aware distance between distributions. Our method enables visualization tools such as time series plots and low-dimensional embeddings. These visualizations in combination with persistent homology reveal important insights from the data. In particular, we were able to identify missing data in the dataset, and detect and compare the freezing and melting times across years. Our most striking finding is a fundamental asymmetry between the freezing and melting processes. These preliminary findings demonstrate the potential of OT and TDA to reveal structure in climate change data, and more generally to satellite image data.

VISUALIZING ARCTIC ICE DATA WITH OPTIMAL TRANSPORT

by

Jocelyn Ilagor

A Thesis Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Greensboro
2023

Approved by

_____
Committee Chair

Dedico esto a mi familia, por todos sus sacrificios

APPROVAL PAGE

This thesis written by Jocelyn Ilagor has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____
Dr. Thomas Weighill

Committee Members _____
Dr. Yu-Min Chung

_____
Dr. Michael Hull

_____
Dr. Clifford Smyth

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor, Dr. Thomas Weighill. I wouldn't have been able to get through all of this without his support and patience. I want to thank my committee, Dr. Chung, Dr. Hull, and Dr. Smyth, for their time and assistance. As well as the many people I have met and befriended at UNCG these past few years. The support and guidance I have received at UNCG have made me into the person I am today and I cannot express how grateful I am.

Table of Contents

List of Figures

Chapter 1: Introduction

Optimal Transport (OT) can be described as the study of transporting one distribution to another and associating a cost to that transport. As with any other transportation problem, the minimum cost is the most ideal. OT has been increasingly used in computational applications such as color or texture processing, shape manipulation, and machine learning. In this thesis, OT will enable a computational approach to visualizing the Arctic ice change.

Climate change is a growing concern that impacts everyone. Declining sea ice not only reduces the reflection of sunlight but it can also contribute to rising sea levels, and affect ecological and geomorphological processes [18]. The ecological world is not the only one that is impacted by these effects. Its important for the Arctic and Antarctic ice to reflect heat back with less heat being reflected this could lead to intense heatwaves. Sea level rising has been of particular concern as it causes coastal floods and storm surges. Specifically looking at the Arctic, the changes in ice and water levels will lead to changes in the structure and function of ecosystems [14].

The motivation behind the present project was to make use of methods from optimal transport to study the ice change happening in the Arctic. Aside from OT, other methods like Topological Data Analysis (TDA) and Multi-Dimensional Scaling (MDS) were used. TDA is a method that focuses on using topology to analyze datasets.

The basic method consists of computing the distance between the arctic ice using Wasserstein distance (optimal transport distance). This distance takes into account not only the amount of ice but the distribution of the ice in space. This distance enables new analysis techniques such as low dimensional embedding, time series analysis, and persistent homology. These techniques extract insights that might have remained hidden if we only analyzed the images using the human eye.

We briefly outline the various sections of the thesis. In Chapter 2, we discuss the necessary background needed to understand the theory used to compute our results, focusing on optimal transport and persistent homology. The following chapter goes over the methods used to analyze the dataset and the methods used to interpret the results. The data is also discussed in this chapter as well as the errors in the dataset. In Chapter 4, we present the results obtained and discuss some findings from these results. In Chapter 5, we conclude and mention some possible future directions.

1.1 Related Work

Optimal transport. Optimal transport methods have been applied in many fields including statistics, economics, computer graphics and fluid mechanics. Examples of applications of OT to image analysis include image segmentation, watermarking, and color transfer. OT has been used in image segmentation by using various transport based cost functions and relying on primal-dual algorithm to solve the convex optimization problem in the images [16]. Strong security purposes are achieved through watermarking but it is easy for distortion to happen. OT has been used to minimize the global distortion in the images [7]. OT has also been applied to the problem of color transfer between images. OT was able to remove the issues of additional postprocessing [15].

Topological data analysis. TDA consists of methods which quantify the shape of data. TDA methods have been applied to a wide range of fields like neural science, image classification, medical imagining, and gerrymandering. Persistence diagrams have been transformed into Gaussian persistence curves and applied to texture datasets [1]. TDA has been applied to gerrymandering by producing a persistence diagram from election data [2]. Persistent homology has also been used to detect geometric voids in data sets specifically COVID-19 data [4]. More specific to this research TDA has been applied to Arctic ponds formed on the surface of the Arctic sea ice. TDA was applied to melt pond evolution, obtaining information about the geometric structure of the melt ponds [20].

Arctic Ice Analysis. Analysis on the Arctic ice to evaluate the season cycle of the ice has previously been done. Time series of the ice extent also show that there is a positive trend and correlates negatively with high-latitude temperature fluctuations [19]. More recent studies have been conducted to see the impact the changing sea ice will have on biology and human activity in the Arctic. The changes in the sea ice is making it harder for certain species whose habitats might no longer be there as well as native communities that live in the Arctic [9]. This once again sheds light on the importance of this research.

Chapter 2: Background

In the following sections, we recall material provided in the text Computational Optimal Transport written by Gabriel Peyré and Marco Cuturi [13]. Let $[[n]]$ denote the set $\{1, 2, \ldots, n\}$ and $\mathbf{1}_n$ be a vector of ones of length n.

2.1 Optimal Transport Theory

We work with discrete measures throughout since these will be most important for the computational experiments later on, but all of this theory extends to arbitrary measures.

Definition 2.1 (Discrete Measures)**.** A *discrete measure* with weights $\mathbf{a} \in \mathbb{R}^n$ and locations $x_1, ..., x_n \in X$ is defined as $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i}$, where δ_{x_i} is the unit mass concentrated at location x_i.

2.1.1 Monge-Problem

The Monge-Problem seeks a map that matches each point x_i to a single point y_j and must push the mass of α toward the mass of β where α and β are discrete measures [13].

Definition 2.2 (Monge problem)**.** Let $T : \{x_1, ..., x_m\} \longrightarrow \{y_1, ..., y_n\}$ be a mapping. We write $T_{\#}\alpha = \beta$ if $\mathbf{b}_j = \sum_{i:T(x_i)=y_i} \mathbf{a}_i$ for all $j \in [[n]]$. Then for a ground cost function c the Monge-Problem is

$$\min_T \{\sum_i c(x_i, T(x_i)) : T_{\#}\alpha = \beta\}.$$

A mapping T realizing this minimum is called a *Monge map*.

Note that a Monge map may not exist between discrete measures in general.

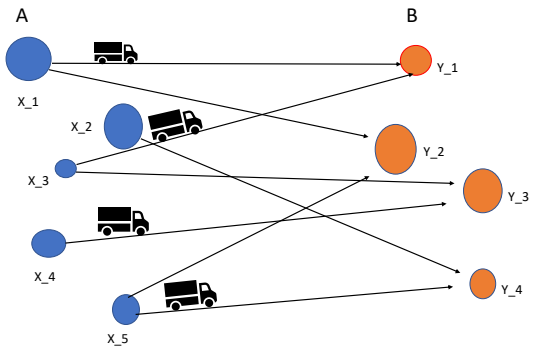


Figure 2.1. Kantorovich's OT Problem through Mines and Factories.

2.1.2 Kantorovich Relaxation

The Kantorovich Relaxation takes the idea of a Monge Problem and relaxes the rules around it. Now, to solve the optimal problem one no longer needs to only assign one location to another. Instead, Kantorovich aims to move any point across several locations which allows for what is known as mass splitting [13]. In order to achieve this a coupling matrix, $\mathbf{P}$, is needed where $\mathbf{P}_{i,j}$ describes the amount of mass flowing from x_i toward y_i.

Definition 2.3 (Set of Coupling Measures)**.** For two discrete measures with weights $\mathbf{a}$ and $\mathbf{b}$ respectively, we define the set of *couplings* as:

$$\mathbf{U}(\mathbf{a}, \mathbf{b}) = \{\mathbf{P} \in \mathbb{R}_+^{n \times m} : \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T\mathbf{1}_n = \mathbf{b}\}$$

We can also note that a coupling $\mathbf{P}$ is in $\mathbf{U}(\mathbf{a}, \mathbf{b})$ if and only if $\mathbf{P}^T$ is in $\mathbf{U}(\mathbf{b}, \mathbf{a})$ making Kantorovich's relation formation always symmetric [13]. We can now move on to Kantorovich's optimal transport problem.

$$L_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle = \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j} \tag{2.1}$$

The easiest way to understand this OT problem is the famous mines and factories example [13]. Looking at Figure 2.1, there are 5 warehouses and 4 factories where each warehouse contains material needed to run the factories. Each warehouse is indexed with an integer i and contains x_i units of the material. The materials need to be moved to the factories with a prescribed quantity y_j needed at factory j. To move the materials we need a to use a transportation company that will charge $\mathbf{C}_{i,j}$ to move a single unit of from location i to j. We assume the price is linear in this example,

the cost of shipping x units of the material from one location to another is $x \cdot C_{i,j}$. Using Equation 2.1 to solve this we obtain a transportation plan $\mathbf{P}^*$ that measures the amounts of goods $\mathbf{P}_{i.j}$ being transported from a warehouse i to a factory j. Now, the total amount needed to be pay to execute the plan is $\langle \mathbf{P}^*, \mathbf{C} \rangle$.

2.1.3 Wasserstein Distance

Wassertein distances are metrics between probability distributions that measure the minimal effort required to reconfigure the mass of one distribution in order to recover the other distribution [12]. To consider the Wasserstein distance we need to assume that the locations x_i and y_i for the discrete measures are in some metric space M with distance d. We then let the fixed cost matrix equal $\mathbf{D}^p = (\mathbf{D}^p_{i,j})_{i,j} \in \mathbb{R}$ where $\mathbf{D}_{i,j} = d(x_i, y_j)$. Then we define the p-Wasserstein distance for $p \geq 1$ associated with the distance matrix $\mathbf{D}$ as

$$W_p(\mathbf{a}, \mathbf{b}) = L_{\mathbf{D}^p}(\mathbf{a}, \mathbf{b})^{\frac{1}{p}} \tag{2.2}$$

where $L_{\mathbf{D}^p}$ represents the optimal solution as in Equation (2.1).

It can be shown that W_p is symmetric, positive, $W_p(\alpha, \beta) = 0$ if and only if $\alpha = \beta$, and it satisfies the triangle inequality [13]. The triangle inequality is as follows

$$\forall \alpha, \beta, \gamma,\ W_p(\alpha, \gamma) \leq W_p(\alpha, \beta) + W_p(\beta, \gamma) \tag{2.3}$$

2.2 Persistent homology

To further understand the results some knowledge on persistence homology is needed. For this, we draw on standard material provided in the text Introduction to Applied Algebraic Topology written by Tom Needham [11].

Definition 2.4 (Convex Set)**.** A subset S is a *convex set* if for any point $x, y \in S$ each point $(1-t)x + ty, t \in [0, 1]$ along the interpolation of x and y is contained in S.

The *convex hull* of S is the smallest convex subset of $\mathbb{R}^k$ which contains S we denote this by $\mathsf{cvx}(S)$. [11].

Definition 2.5 (Simplex)**.** A *simplex* associated to S is the set $\sigma(S) = \mathsf{cvx}(S)$ where each point x_i is called a vertex of $\sigma(S)$ and a pair of points is called an edge of $\sigma(S)$. The face of $\sigma(S)$ is $\sigma(T)$ where $T \subset S$.

A 0-simplex is a point. A 1-simplex is an edge and a 2-simplex is a filled triangle. As for a 3-simplex this is seen as a solid tetrahedron. We can note that a 0-simplex has one face. While a 1-simplex has the faces of the edge AB, A, and B. Now we can define what a simplicial complex is.

Definition 2.6. A *simplicial complex* is a collection of simplices $\mathbf{X}$ in some $\mathbb{R}^n$ that satisfies the following:

1. Given any simplex $\sigma \in \mathbf{X}$ all the faces σ are contained in $\mathbf{X}$.

2. For any simplex $\sigma, \tau \in \mathbf{X}$ the intersection $\sigma \cap \tau$ is either empty or is also a simplex and a face of both σ and τ

We almost have everything we need to understand what persistent homology is. The last thing we need is something called Vietoris-Rips complexes or VR complexes.

Definition 2.7 (VR-Complex)**.** Let $(\mathbf{X}, d)$ be a finite metric space and r a distance parameter. Then the *Vietoris-Rips complex* at parameter r is $VR(\mathbf{X}, r) = (V_r, \Sigma_r)$. The simplex set Σ_r is defined as

$$\Sigma_r = \{\sigma \subset \mathbf{X} \mid d(x, y) \leq r \forall\ x, y \in \sigma\}$$

and $V_r = \mathbf{X}$.

So, as long as every pair of points in $\sigma \subset \mathbf{X}$ is at most r apart we can include σ in Σ_r. Note that (V_r, Σ_r) defines an abstract simplicial complex (i.e. a list of vertices and simplices) which has many possible geometric realizations (Definition 2.6). The simplicial homology of this complex depends only on the pair (V_r, Σ_r) and not on the geometric realization.

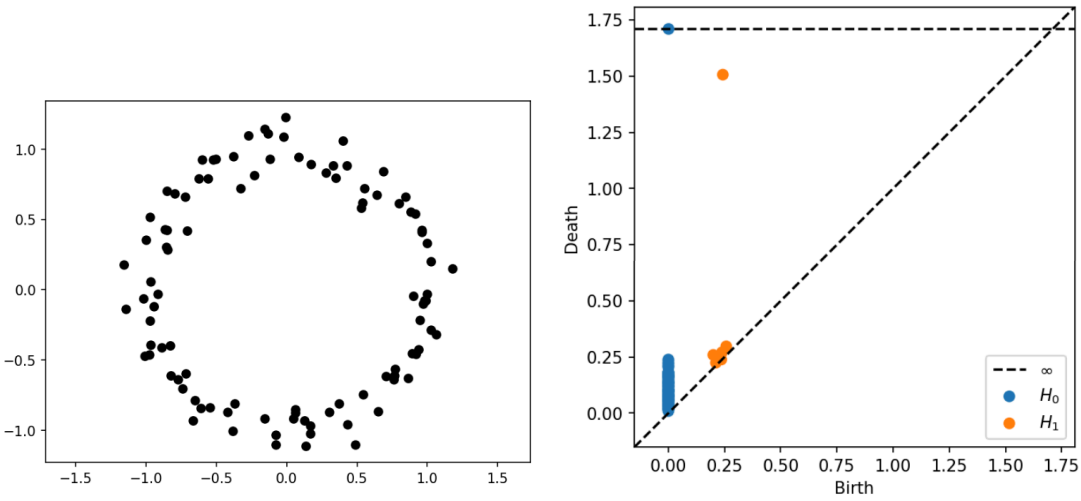
Definition 2.8 (Filtered Simplicial Complexes)**.** A *filtered simplicial complex* is a collection $K = \{K_r\}_{r \geq 0}$ of (finite) simplicial complexes K_r such that $K_r \subseteq K_s$ when $r \leq s$.

The goal of persistent homology is to track the appearance and disappearance of holes and connected components. We will not recall the details of simplicial homology here, see [11] for a description in the F_2 case.

Theorem 2.9 (Fundamental Theorem of Persistent Homology [21])**.** *Let $K_{r_1} \subseteq \cdots K_{r_n}$ be a filtered simplicial complex. and let $i_1*, \ldots, i_n*$ be the induced maps of homology in dimension in k. Then we can choose a basis $\mathcal{B}_j$ of each $H_k(K_{r_j})$ such that for each element $b \in \mathcal{B}_j$, either $i_k * (b) = 0$ or $i_k * (b) \in \mathcal{B}_{j+1}$ and no other $b' \in \mathcal{B}_j$ maps to $i_k * (b)$.*

The Fundamental Theorem of Persistence Homology gives rise to the concept of birth and death. If $b \in \mathcal{B}_j$ is not in the image of $i_{(j-1)*}$, then we say that b is *born* at r_j. For $b \in \mathcal{B}_j$ to *die* at r_m we need the following to hold:

$$i_{j*}(b) \neq 0$$
$$i_{j+1*} \circ i_{j*}(b) \neq 0$$
$$\ldots$$
$$i_{m-1*} \circ \ldots \circ i_{j*}(b) \neq 0$$
$$i_{m*} \circ \ldots \circ i_{j*}(b) = 0$$



(a) A VR Complex

(b) The PD of the VR Complex

Figure 2.2. Detecting a hole through Persistence Diagrams.

If no such m exists then we say that b dies at ∞. A visualization of birth and death is something called persistence diagrams which will be talked about in the following chapter.

Definition 2.10 (Persistence Diagram)**.** The *persistence diagram*(PD) consists of the multi-set $\{(b_i, d_j)\}_{j=1}^{M}$ consisting of all (birth, death) pairs.

From Figure2.2, we can see how persistence diagrams detect holes. We first take note that there are two things being plotted: H_0 and H_1 component. The H_0 represents the connecting components while H_1 represents the holes. Looking at the Figure2.2b we can see that there is one connecting component that "survives". We also see that there is one point in H_1 that is furthest from the diagonal. This means it had a longer life compared to the other points. This is an indication that we have a circle, as we have one component that persists and only one point that is away from the diagonal in H_1. This is proven by Figure2.2a, as that is what we plotted in order to achieve 2.2b.

Chapter 3: Methods

3.1 Data

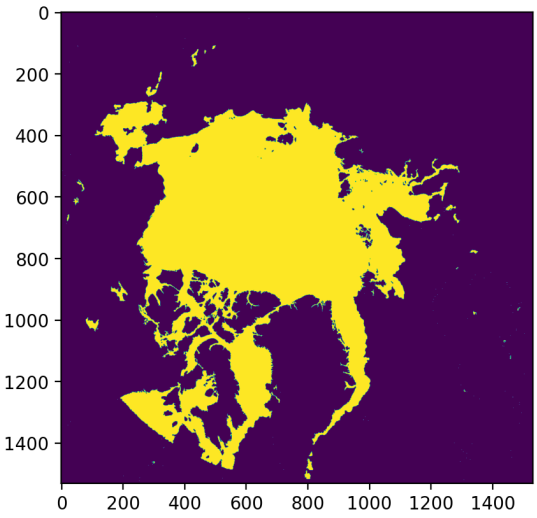
The data used was found through the NASA National Snow and Ice Data Center as *NASA SCP Arctic and Antartic Ice Extent from QuikSCAT, 199-2009, Version 2 (NSIDC-0265)* [5]. The data set provides sea ice extent for the Arctic and Antarctic in Scatterometer Image Reconstruction (SIR) binary image format. The data extends from July 19, 1999 to December 31, 2009. NASA's Quick Scatterometer (QuikScat) satellite obtained 12 individual radar normalized backscatter measurements called *slices*. The summed up measurement of the slices are called *egg* measurements which are stored as image files. An example can be seen in 3.1a. Each day contains a set of both slices and egg images. Only eggs were used in the computations of this thesis. The SIR file used in the implementations contained at least one 512-byte header with information to read the file and projection information to geolocate the data [5].

3.1.1 Errors in Data Set

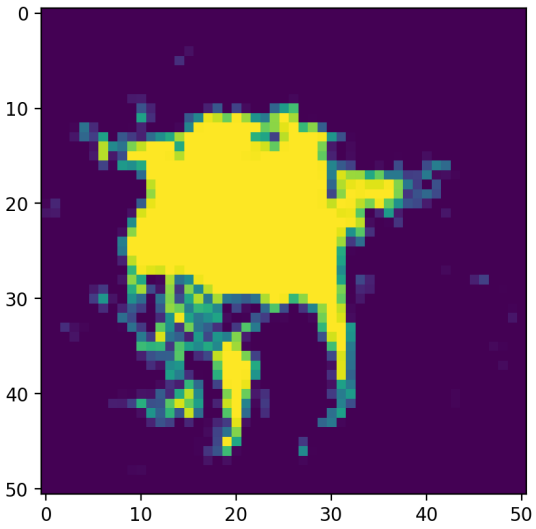
It was noted that high winds can make the ocean appear like ice, while surfaces melt conditions can make sea ice appear like ocean. As well as rapid ice edge motion can lead to misclassification errors.

The QuickSCAT suffered a power anomaloy on November 18, 1999 making the egg images blank for that day. Other years also contained blank eggs or just missing dates.The SeaWinds which is a scatterometer that flies on the QuickSCAT collects data more than 99% of the time. However, there are data gaps due to spacecraft anomalies [6]. Not considering this every year except for leap years contained all 365 eggs. The only exception being 2008 that had blank egg files. This was fixed later on by removing the blank images and replacing them with the previous egg (see Figure 4.4).

(a) Raw Egg



(b) Binary Image



(c) Low Resolution Image

Figure 3.1. The visual representation of the image processing done to obtain the low resolution images.

3.2 Computing distances between images

3.2.1 Image to Discrete Distributions

The goal was to convert an egg into a discrete distribution in the plane. Each image had a size of 1530 by 1530. This would later prove to be an issue in regards to computational time, but first we focused on finding ice pixels. We isolated all the pixels that represented the ice in the image. To do this the image had to be turned into a binary array of 1 for ice or 0 for no ice as seen in Figure 3.1b.

Given that the images had a size of 1530 by 1530 computation of Wasserstein distances later on in the process was taking too long so the resolution of the images had to be changed. In order to do this the images had to be partitioned. The images were partitioned by taking a 30 by 30 block and summing the values for each block we can see the outcome in Figure 3.1c. Each block was represented by discrete distribution:

- locations were the coordinates of the central pixel in each block,
- weights were equal to the sum over the 30 by 30 block divided by the sum of all such values

These same weights were used as the weights used to compute Wasserstein distances later on. We unravel the 51 by 51 array of weights to a vector of length 2601 for this purpose. We now have the low resolution ice images as well as the associated weight vectors all we need is the cost matrix.

3.2.2 Cost matrices

We consider two different geometric cost matrices.

Pixel distance (G_{pixel})

The G_{pixel} cost matrix used the Euclidean distance between pixels.

$$d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3.1}$$

Geographic distance (G_{geo})

This cost matrix used the Geographic distance (Haversine formula) between locations. The Haversine formula determines the distance between two points on a sphere given their longitudes and latitudes. The Haversine formula allows the Haversine of θ to be computed from the latitude φ and longitude λ of two points. Note that θ is the angle between two points on a sphere is given by:

$$\theta = \frac{d}{r} \tag{3.2}$$

where d is the spherical distance between two points and r is the radius of the sphere. The Haversine formula of two points is[17]:

$$hav(\theta) = \frac{\sin^2(\varphi_2 - \varphi_1)}{2} + \cos(\varphi_1)\cos(\varphi_2)\frac{\sin^2(\lambda_2 - \lambda_1)}{2} \tag{3.3}$$

where φ_1, φ_2 are the latitude of point 1 and latitude of point 2, and λ_1, λ_2 are the longitude of point 1 and longitude of point 2. Finally, the distance is obtain by applying arcsin:

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos\varphi_1 \cos\varphi_2 \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{3.4}$$

3.2.3 Optimal transport distance

Recall the optimization problem from the previous section. Given a cost matrix G (either G_{pixel} or G_{geo}) we get:

$$L_{\mathbf{G}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{G}, \mathbf{P} \rangle = \sum_{i,j} \mathbf{G}_{i,j} \mathbf{P}_{i,j} \tag{3.5}$$

where $\mathbf{a}$ and $\mathbf{b}$ are the vectors of weights associated to two egg images.

Its important to note that since the discrete distributions have a fixed set of possible locations we can precompute G_{geo} and G_{pixel} just once.

3.3 Visualizing with MDS

After obtaining the pair-wise distance for all years Multi-Dimensional Scaling (MDS) was used to analyze the results. The aim of MDS is to find the representation in as few dimensions as possible, while still reproducing the dissimilarities [8].

3.4 Finding holes with TDA

Topological Data Analysis (TDA) was also used to interpret the results. It was seen through MDS that certain years had "holes". To visuually see this persistence homology was used in order to obtain the persistence diagrams.

3.5 Implementation

The main Python libraries used were `ot`, and `numpy`. The majority of the functions used to transform the eggs into arrays where provided by the user guide [5], as well as how to convert pixels to latitude and longitude points.

3.5.1 Cost Matrix G

The functions used to compute the cost matrix were loadsir, distance, geodistance, pix2latlon_custom. The loadsir function and the pix2latlon_custom was provided by the user guide while the latter had to be customized to work with the newer version of python. The loadsir function was used to load the sir image and head array into python. The head array is the scaled header information block that contains scale factors to convert floating-point data to integers [5]. The pix2latlon_custom function was used to convert pixels to latitude/longitude points. The distance function was the adaptation of the Haversine function formula. The geodistance function took two ice pixel arrays as well the header then used pix2latlon_custom to convert the pixels to latitude and longitude points. Then distance was applied to compute the haversine distance between those points. The geodistance function was used to obtain the cost matrix G_{geo}.

3.5.2 Wasserstein Distance

Wasserstein distances were computed by using the function `ot.emd2` from the Python OT [3]. Given the year, the files were loaded, the icepixels were obtained and turned into lower resolution images. The weights were then normalized. The function then took the sample weights as well as the cost matrix and returned the optimal transport loss. To be able to compute the distances for every year the longleaf high perfomance cluster at UNC Chapel Hill was used.

Chapter 4: Results

We have now successfully turned each image into a discrete distribution, and can compute pairwise Wasserstein distances. We can now make sense of our results through visual representations.

4.1 Time Series

In Figure 4.1, we can see the ice change for the whole year over a seven day interval. More precisely, we compute, for each day i, the ice change from day i to day $i + 6$ using our Wasserstein distance. To smooth out the graphs further, we replace each value with the seven day average. Thus Figure 4.1 represents the rolling seven day average of the per-week ice change for each year. The results are identical even though a different G matrix was used.

We can see that there are two peaks around mid April and late September. The peaks around April being the one where the ice is melting and the peaks around late September indicating freezing. It is also easy to see the change in ice through the seasons. There is also variability between the years and the peak sizes. We can see that 2008 has the highest peak between all the years. The contrast in peak size can be seen when comparing 2008 with 2002. Not only is the peak smaller during the freezing period but the freezing period for 2002 came earlier compared to 2008. Another difference can be seen when the ice is melting. For 2003, the peak hits its highest around early April while in 2008 the peak does not happen until mid to late April.

4.2 MDS Plots

We were able to find what years were missing dates through MDS. When the starting point did not end in the same area as the ending point, this was a clear indicator of missing data. This visualisation of the data with MDS also helped see how some years had a "hole" or an opening as the points made their way back to the starting point as

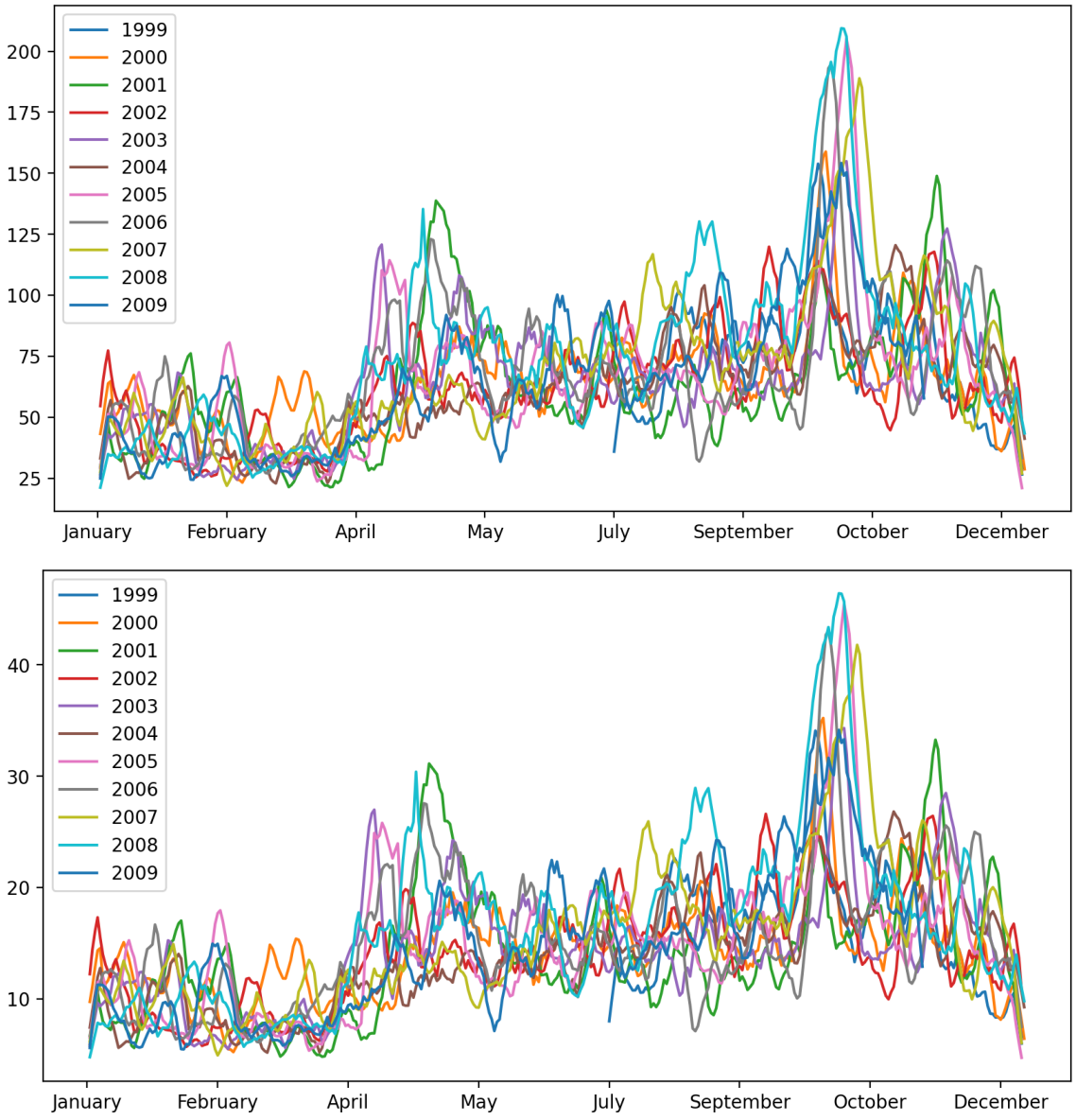


Figure 4.1. Time Series of G_{geo} (above) and G_{pixel} (below) showing the smoothed per week ice change throughout the year. Different choices of G yielded almost identical graphs.

seen in Figures 4.2. Both G_{geo} and G_{pixel} yielded almost identical results hence, only the MDS plots for G_{geo} are shown.

While the MDS plots for 2006 show the points' trajectory make its way back to the start we can see the opposite for 1999 in Figure 4.3. The MDS plot shows that the starting point did not end in the same area as the ending point. This indicated that the year was missing days which is confirmed by the user guide. The most interesting thing can be seen in Figures 4.2. We took note of a hole or cycle structure in all plots, which provided us with a different insight.

The hole corresponds to a different melting and freezing trajectory. There is a lack of symmetry that shows that the freezing of the ice does not look like the melting. Not every year had a clear indicator of having a "hole" (for example 2003) so, persistent homology was used to confirm the "hole" even when it was not clear on the MDS embedding.

4.3 Persistence Diagrams

The initial analysis conducted revealed there was an issue for 2008. We can see in Figure 4.4a that there is something wrong with the persistence diagram for 2008. The difference is seen in the H_0 components which are the blue dots in Figure 4.4. The outlier is obvious in the corrupted image as there is an H_0 component that lived longer than the other H_0 components. This resulted in finding out that 2008 had blank eggs. Removing the outliers gave us Figure4.4b.

Aside from detecting an abnormality the persistence diagram also confirmed the circle found in the MDS plots. We can see that if a component has a short life (dies shortly after it is born) it will be close to the diagonal [10]. With this in mind we can see in Figures 4.5 that there is a point in H_1 that has a long lifetime which means it is far from the diagonal. Having this point indicates we have a hole, in other words, the rough topology of the data is that of a circle. Once again, the circle represents that there is a different freezing and melting pattern, as if it was symmetrical there would be no hole. This can be seen in Figure 4.6. We can also see that unlike the other years 2009 does not have a hole this is due to 2009 having missing data.

This finding demonstrates the usefulness of persistent homology for finding structure in highly non-linear, high dimensional spaces. Persistence diagrams do not rely on the planar embedding created by MDS, but use the original distances which shows that the freezing and melting time is not symmetrical, and this is not just an artifact of the embedding.

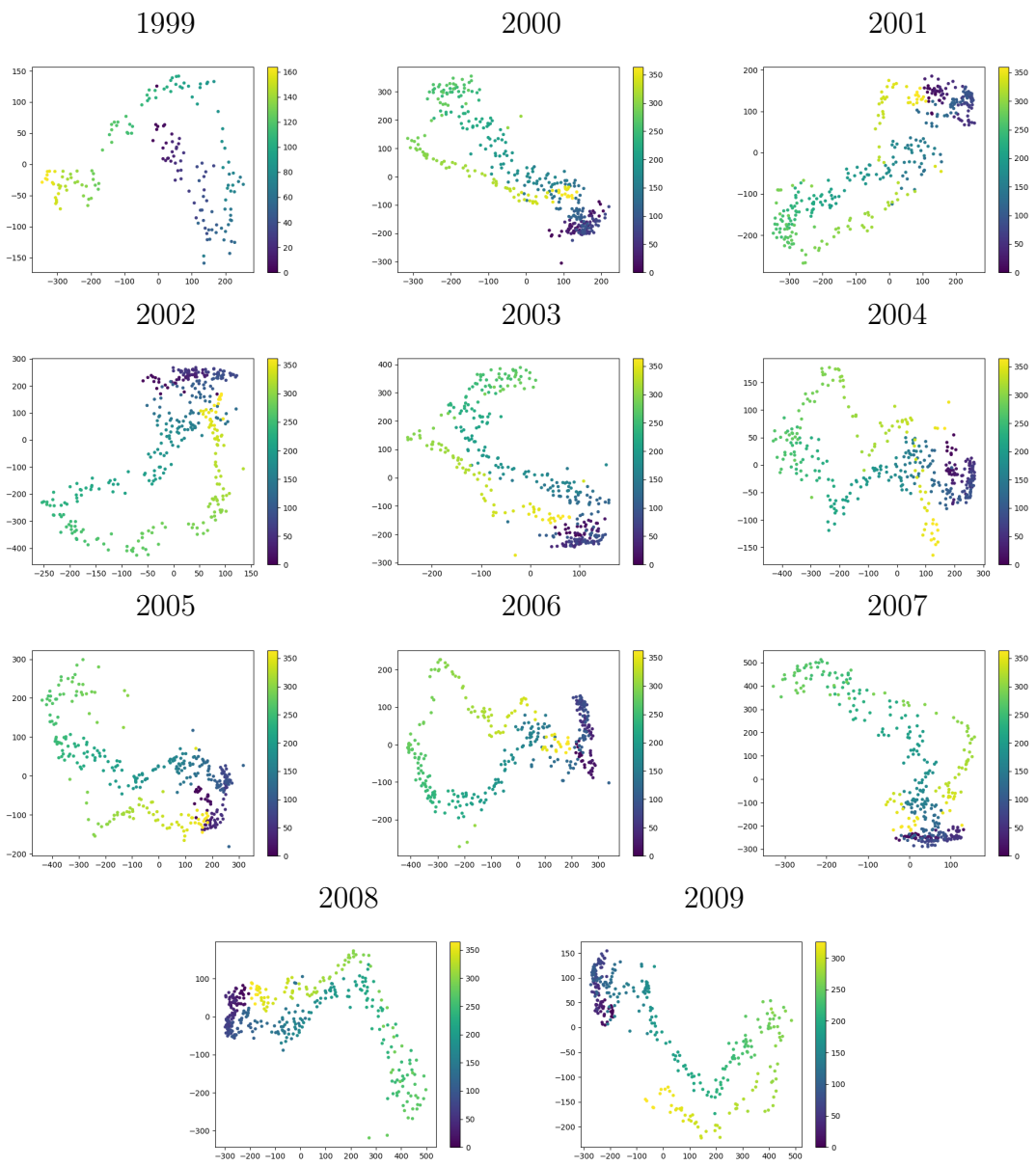


Figure 4.2. MDS plots for 1999-2009 using optimal transport distance. The color indicates the day of the year.

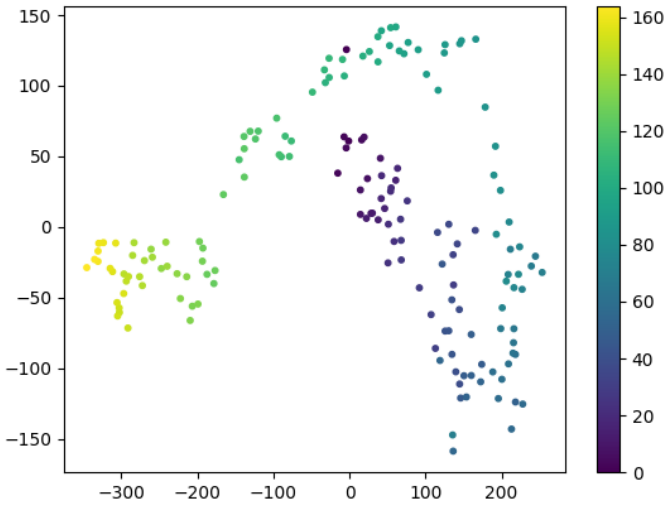
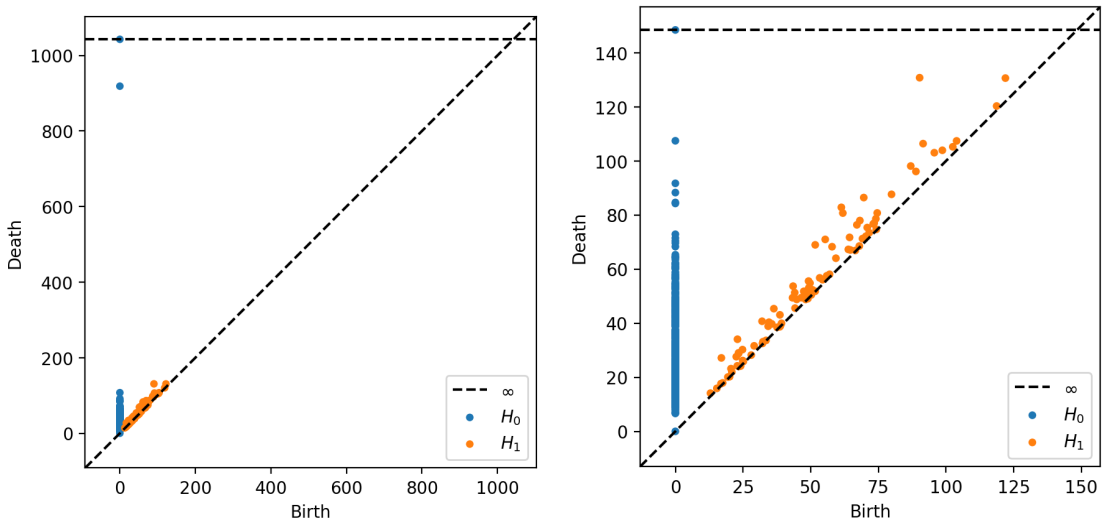


Figure 4.3. Detecting missing dates for 1999 through MDS.



(a) corrupted images included (b) corrupted images removed

Figure 4.4. Vietoris-Rips persistence diagrams for 2008 with and without corrupted images. The presence of the corrupted images (outliers) is clearly detected by the H_0 diagram in the left image.

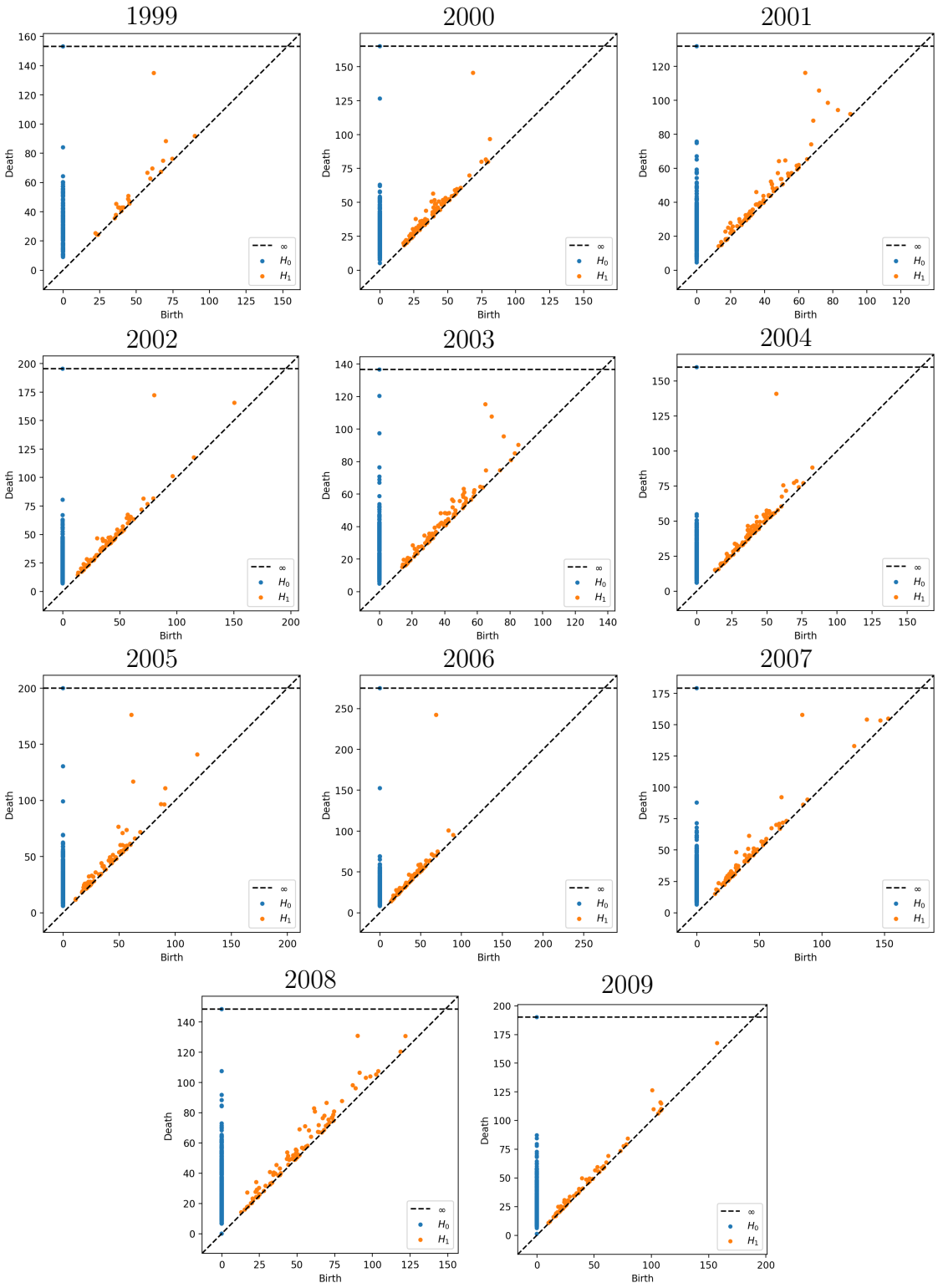


Figure 4.5. Vietoris-Rips persistence diagrams for 1999-2009.

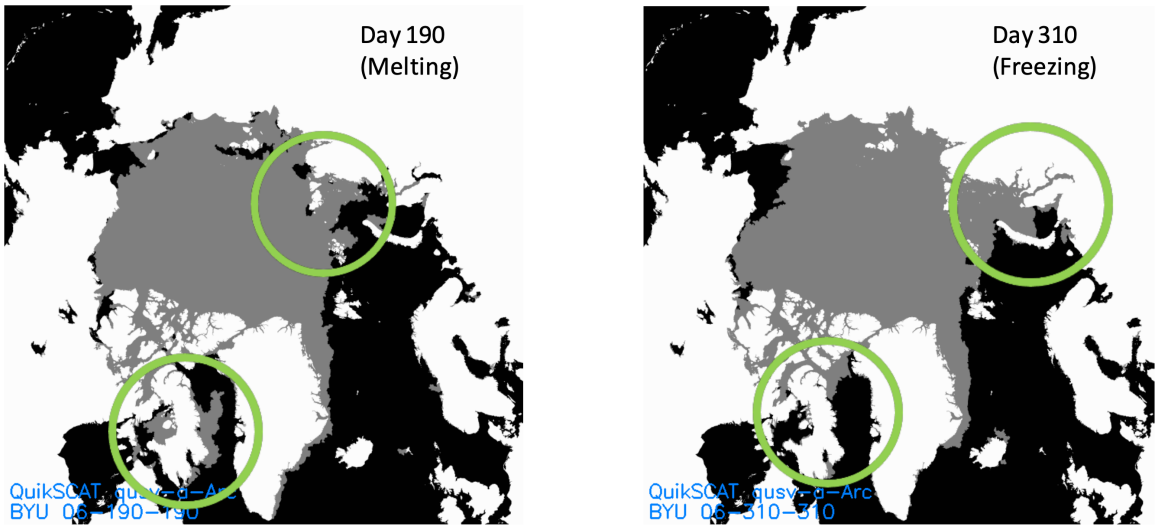


Figure 4.6. Asymmetry in melting and freezing for 2006.

Chapter 5: Conclusion

5.1 Findings

Optimal transport is a growing field that finds the optimal solutions to a variety of problems. The main goal of this thesis was to find ways to visualize the ice change in the Arctic using optimal transport. We were able to find a way to change an image into a discrete distribution, create cost matrices, and find holes using MDS and TDA. Not only were we able to do this but we were able to use visualization tools and persistent homology to make concrete findings:

- The melting and freezing process are different. Instead of having a symmetric process a lot of the years had a "hole" indicating the lack of symmetry in the freezing and melting process.
- The time series also shows the different seasons in which the ice is forming and melting for all the years, namely September and April respectively. While some years started earlier than others they are all relatively close to each other.
- The data for certain years contained corrupted images and/or missing time periods.

5.2 Future work

From our results, we can see that we found the freezing and melting processes were not symmetric. It would be interesting if those years that do not have the same melting and freezing time will remain asymmetric using a different weights for the cost matrix or a different distance measure. The resolution also had to be lowered due to computational constraints but given the opportunity computing the entire process with the raw eggs to see if there us a significant difference in the results. Given that this process was only done with the Arctic eggs the process can be redone with the Antarctic eggs. Specifically looking at the years that were asymmetric would they still remain asymmetric when looking at the Antarctic eggs. There was not enough

time to explore the footprint of climate change in the data, even though it was a big motivation behind this research. Further research is needed to determine what these results mean for climate change. The methods used throughout this research can also be applied to other fixed regions or images such as deforestation, wetland data, soil levels, and habitat growth. OT has demonstrated the potential it has to reveal structure in climate change data.

References

[1] Yu-Min Chung, Michael Hull, and Austin Lawson. Smooth Summaries of Persistence Diagrams and Texture Classification. pages 840–841, 2020.

[2] Duchin, Moon and Needham, Tom and Weighill, Thomas. The (Homological) Persistence of Gerrymandering. *Foundations of Data Science*, 4(4):581–622, 2022.

[3] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[4] Hickok, Abigail and Needell, Deanna and Porter, Mason A. Analysis of Spatial and Spatiotemporal Anomalies Using Persistent Homology: Case Studies with COVID-19 Data. *SIAM Journal on Mathematics of Data Science*, 4(3):1116–1144, 2022.

[5] D. G. Long. NASA SCP Arctic and Antarctic Ice Extent from QuikSCAT, 1999-2009, Version 2, 2013.

[6] David G. Long. private communication, 2023.

[7] Benjamin Mathon, François Cayre, Patrick Bas, and Benoît Macq. Optimal Transport for Secure Spread-Spectrum Watermarking of Still Images. *IEEE Transactions on Image Processing*, 23(4):1694–1705, 2014.

[8] A. Mead. Review of the Development of Multidimensional Scaling Methods. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41(1):27–39, 1992.

[9] Walter N Meier, Greta K Hovelsrud, Bob EH Van Oort, Jeffrey R Key, Kit M Kovacs, Christine Michel, Christian Haas, Mats A Granskog, Sebastian Gerland,

Donald K Perovich, et al. Arctic Sea Ice in Transformation: A Review of Recent Observed Changes and Impacts on Biology and Human Activity. *Reviews of Geophysics*, 52(3):185–217, 2014.

[10] Elizabeth Munch. A User's Guide to Topological Data Analysis. *Journal of Learning Analytics*, 4(2):47–61, 2017.

[11] Tom Needham. Introduction to Applied Algebraic Topology. `https://research.math.osu.edu/tgda/courses/math-4570/LectureNotes`, 2019. Accessed: 2023–3-18.

[12] Victor M. Panaretos and Yoav Zemel. Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, mar 2019.

[13] Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[14] Terry D Prowse, Frederick J Wrona, James D Reist, John J Gibson, John E Hobbie, Lucie MJ Lévesque, and Warwick F Vincent. Climate Change Effects on Hydroecology of Arctic Freshwater Ecosystems. *AMBIO: A Journal of the Human Environment*, 35(7):347–358, 2006.

[15] Julien Rabin, Sira Ferradans, and Nicolas Papadakis. Adaptive Color Transfer with Relaxed Optimal Transport. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4852–4856, 2014.

[16] Julien Rabin and Nicolas Papadakis. Convex Color Image Segmentation with Optimal Transport Distances. In *Scale Space and Variational Methods in Computer Vision: 5th International Conference, SSVM 2015, Lège-Cap Ferret, France, May 31-June 4, 2015, Proceedings 5*, pages 256–269. Springer, 2015.

[17] C. C. Robusto. The Cosine-Haversine Formula. *The American Mathematical Monthly*, 64(1):38–40, 1957.

[18] Edward AG Schuur, Benjamin W Abbott, Roisin Commane, Jessica Ernakovich, Eugenie Euskirchen, Gustaf Hugelius, Guido Grosse, Miriam Jones, Charlie Koven, Victor Leshyk, et al. Permafrost and climate change: carbon cycle feedbacks from the warming arctic. *Annual Review of Environment and Resources*, 47:343–371, 2022.

[19] John E Walsh and Claudia M Johnson. An Analysis of Arctic Sea Ice Fluctuations, 1953–77. *Journal of Physical Oceanography*, 9(3):580–591, 1979.

[20] Wilfred Offord and Michael Coughlan and Ian J. Hewitt and Heather A. Harrington and Gillian Grindstaff. Topological Data Analysis Detects Percolation Thresholds in Arctic Melt-Pond Evolution, 2022.

[21] Afra Zomorodian and Gunnar Carlsson. Computing Persistent Homology. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, pages 347–356, 2004.