Topological Regression Model

TDA Final Project

Kaheon Kim, Leo Jung

Persistence Diagram

Topological Descriptor

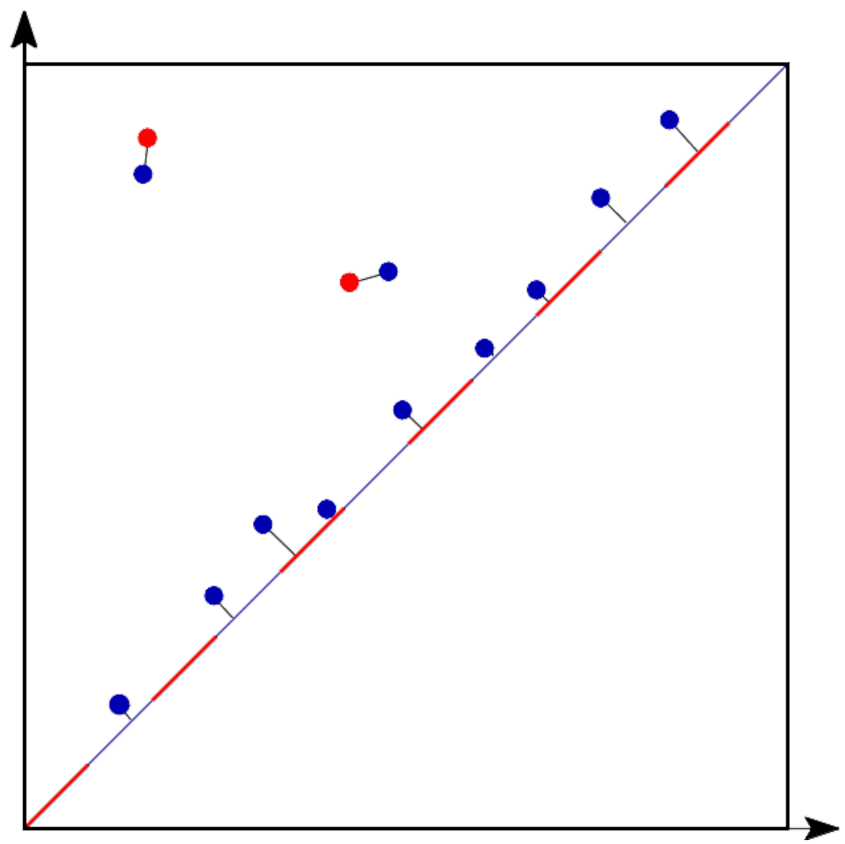
- Representation of topological features such as loops, voids and connected components
- Visualizes b (birth) and d (death) of a topological feature
- Useful in realizing the significant components in vast data.
- Interpretation of Persistence Diagram in Applications

Distance for Persistence Diagrams

Bottleneck distance

The bottleneck distance between PD_1 and PD_2 is defined by

$$d_B(PD_1, PD_2) = \inf_{\gamma} \sup_{x \in PD_1 \cup \Delta} \|x - \gamma(x)\|_\infty$$



Motivation

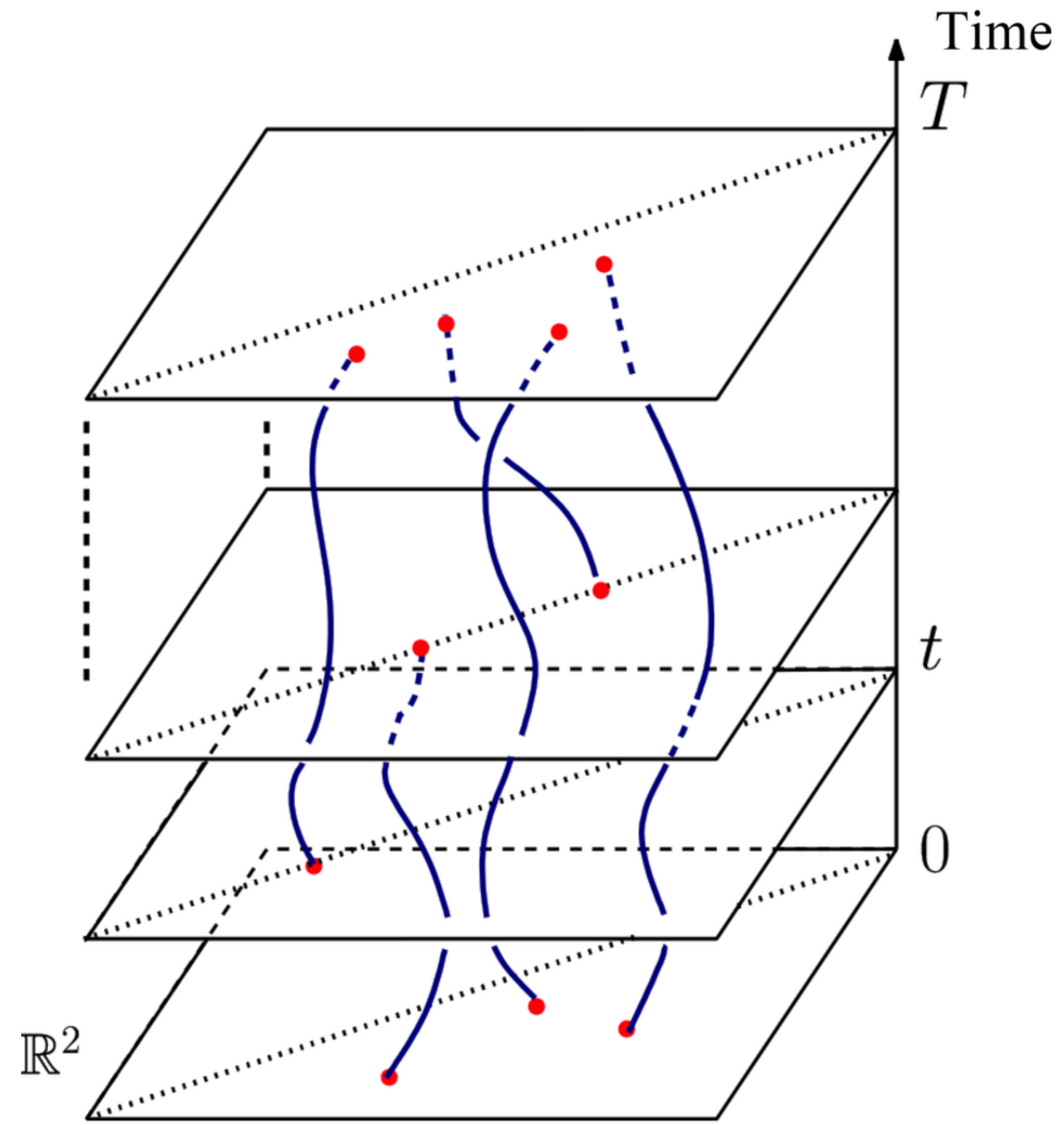
Vine and Vineyard

[Steiner et al, 2006]

- Vine : The trajectory of a single persistent feature as the parameter t changes.
- Vineyard : the collection of all the vines for the entire family $\{f_t\}$

Similar to time-series model

→ Build the regression model for PDs!



Motivation

Regression between probability distributions

Persistence diagram and probability distribution lies in non-Euclidean space.

Several methodologies for regression between probability distributions

- Wasserstein Regression [Chen et al, 2021]

  Regression between tangential maps from each distribution

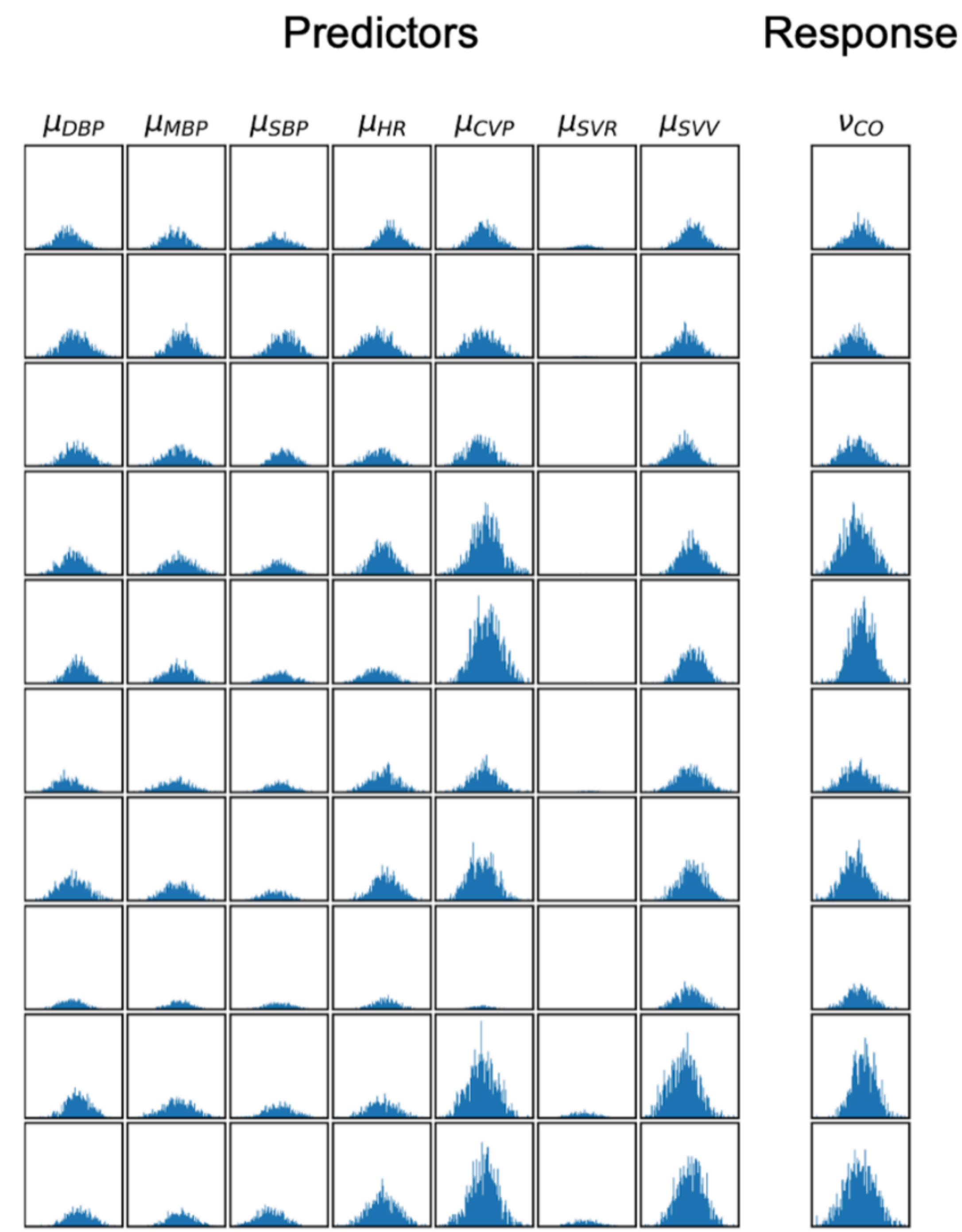
- Distribution-on-Distribution Regression [Ghodrati et al, 2022]

  Regression between distribution by optimal transport map T

Application

- Distributional Data Analysis

  Finance, climate science, medical science.



Distributional predictors and responses from intraoperative anesthesia records

Regression Model for PDs

Topological Regression Model

Goal : To predict response PDs $\{D_i^{(R)}\}_{i=1}^n$ from predictor PDs $\{D_i^{(P)}\}_{i=1}^n$.

We define **topological regression model**

$$D_i^{(R)} \ = \ (T_{\varepsilon_i} \circ T^\star)_\# \, D_i^{(P)}, \qquad \mathbb{E}\Big[T_{\varepsilon_i}(z) \,\Big|\, z\Big] = z,$$

where

- $T^\star : \overline{\mathscr{U}} \to \overline{\mathscr{U}}$ is a model parameter map
- Pushforward $T_\# D = \Big\{T(b_i, d_i)\Big\}_{i \in I}$ where $D = \Big\{(b_i, d_i)\Big\}_{i \in I}$.
- $\mathscr{U} = \{(b, d) : b < d\}$

Regression Model for PDs

Estimation of parameter T

Finding the map T that minimizes average squared-L2 distance from optimal matchings from $PD^{(P)}$ to $PD^{(R)}$, i.e.

$$\hat{T} \in \arg\min_{T \in \mathcal{T}} L(T) := \frac{1}{n} \sum_{i=1}^{n} \sum_{x \in D_i^{(P)}} \left\| T(x) - \gamma_i(x) \right\|_2^2,$$

- $\mathcal{T} := \left\{ T : \overline{\mathcal{U}} \to \overline{\mathcal{U}} \;\middle|\; T(\Delta) = \Delta,\, d'(b,d) - b'(b,d) \geq 0 \right\},$
- γ_i is an optimal matching from $D_i^{(P)}$ to $D_i^{(R)}$, i.e.

$$\gamma_i \in \arg\min_{\gamma \in \Gamma(D_i^{(P)}, D_i^{(R)})} \sum_{x \in D_i^{(P)}} \|x - \gamma(x)\|_\infty,$$

Regression Model for PDs

Estimation of parameter T

T can be parametrized into $T_\theta : \mathbb{R}^2 \to \mathbb{R}^2$, so that

$$\hat{\theta} \ \in \ \arg\min_{\theta \in \Theta} \tilde{L}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \sum_{x \in D_i^{(P)}} \left\| T_\theta(x) - \gamma_i(x) \right\|_2^2$$

- T_θ can be linear model, polynomial, and neural network, etc.
- If T_θ is affine (i.e. $T_\theta(x) = Ax + b$, where $\theta = \left\{A \in \mathbb{R}^{2\times 2}, b \in \mathbb{R}^2\right\}$), the loss function $\tilde{L}(T_\theta)$ is strictly convex, guaranteeing a unique solution.

Numerical Studies

Preliminary Analysis

- 2 climate datasets from NOAA (National Centers For Environmental Information)

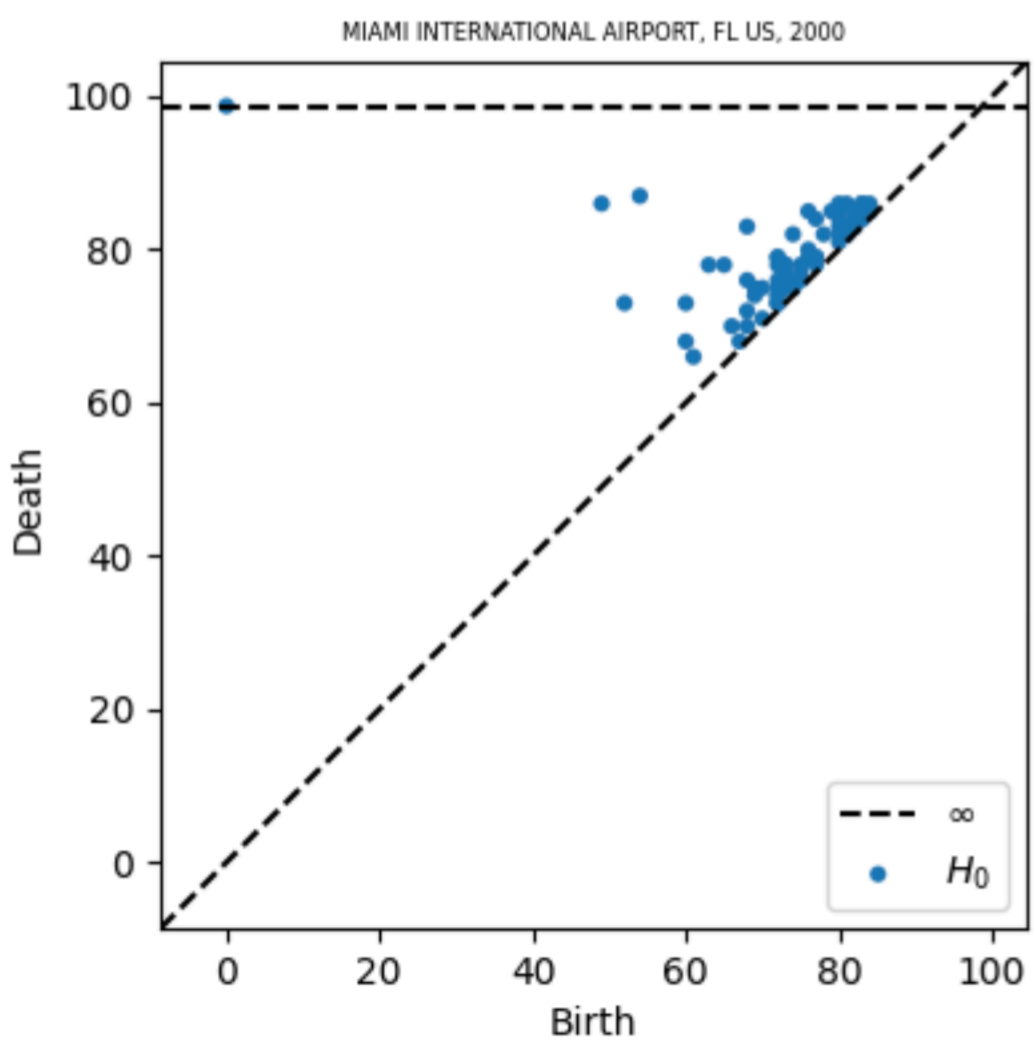
- Records of daily temperature in a single year at 10 different stations (airports) in 2000 and 2020.

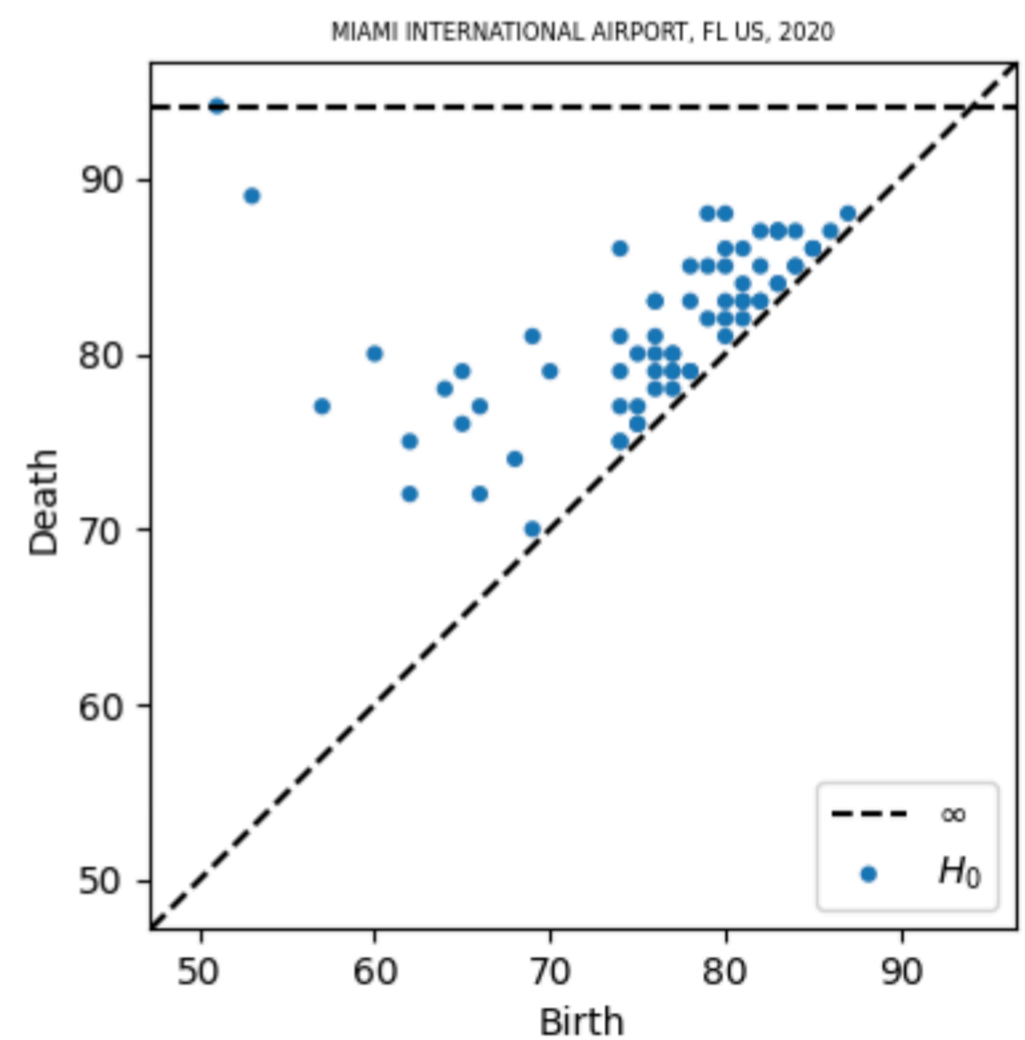
- Perform regression analysis with $T_\theta = Ax + b$ between the persistence diagram at each station using the measured temperature at their respective years.

Numerical Studies

Preliminary Analysis

- Persistence Diagram (Miami)

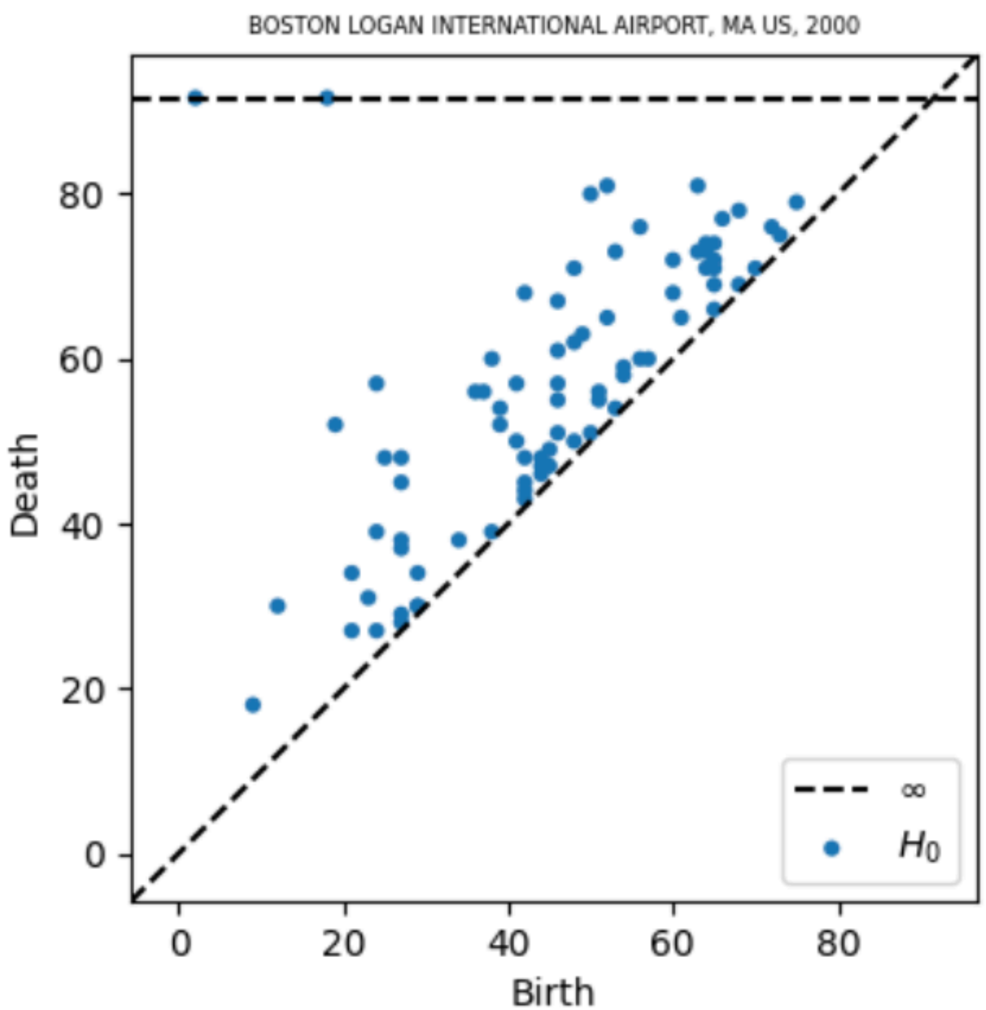


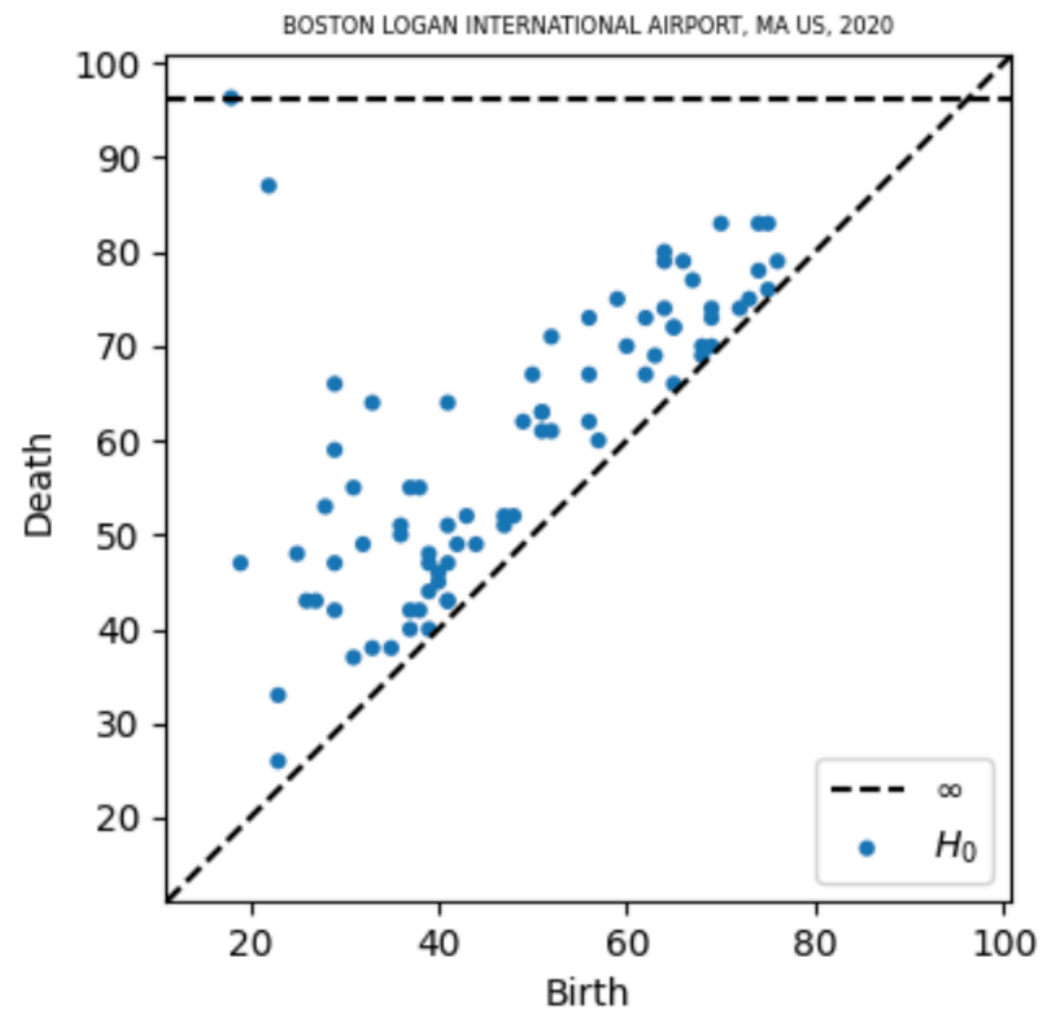


Numerical Studies

Preliminary Analysis

- Persistence Diagram (Boston)

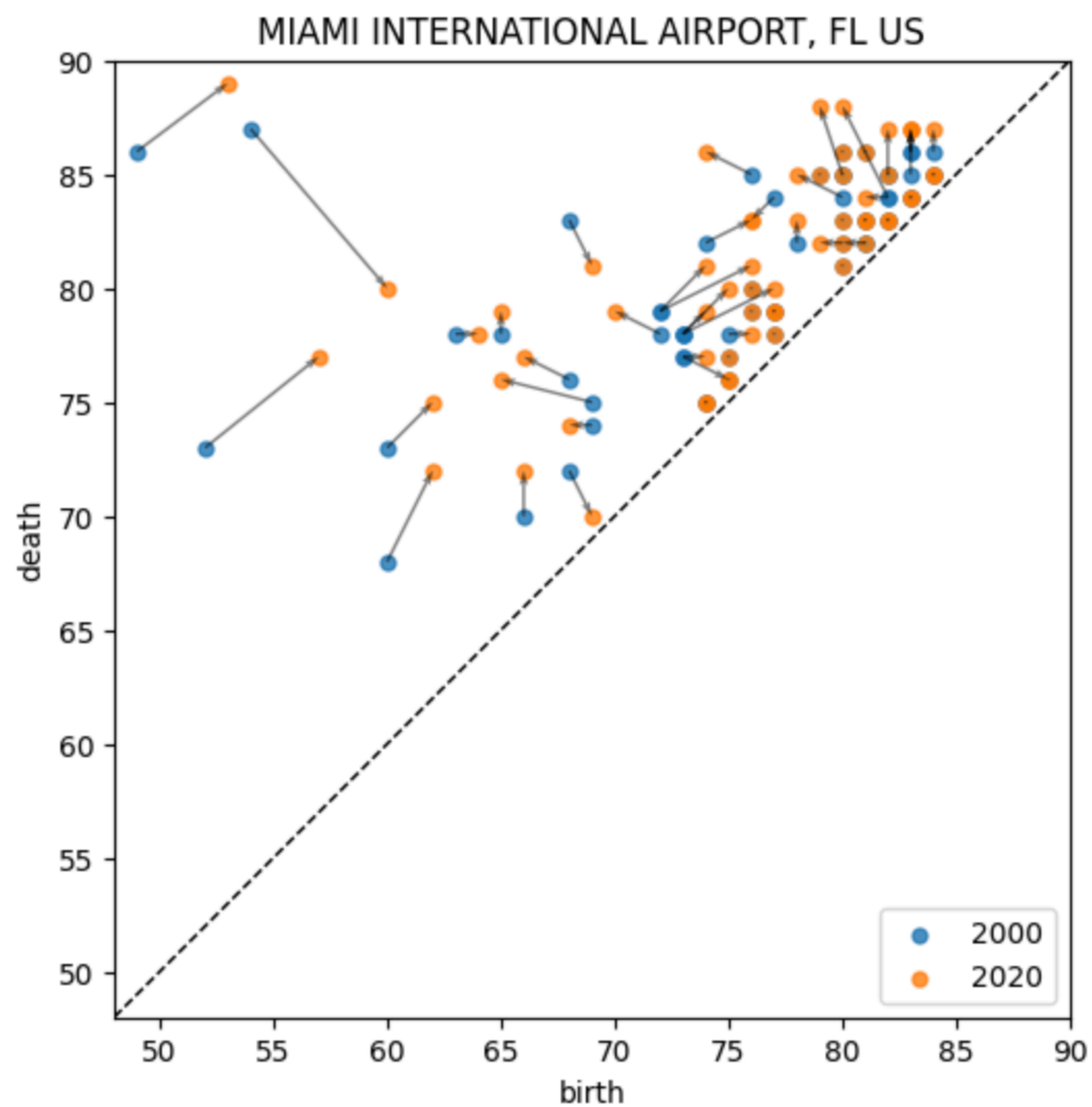




Numerical Studies

Preliminary Analysis

- Pairing using bottleneck distance between persistence diagrams



Numerical Studies

Preliminary Analysis

- Pairing using bottleneck distance between persistence diagrams

Numerical Studies

Result & Analysis

- Estimated parameter : $\hat{T}(x) = Ax + b$

  where $A = \begin{pmatrix} 0.9562 & 0.0377 \\ 0.0165 & 0.9455 \end{pmatrix}$ and $b = \begin{pmatrix} 0.1785 \\ 3.006 \end{pmatrix}$.

- Interpretation for A : $\|A\|_{op} = 0.9786 < 1.$
  - From 2000 to 2020, the point clouds in the PDs has moved down.
- Interpretation for b
  - Compared to the scale of persistence diagrams, the effect of b is minor.

Future Research

- Prediction Tasks
  - Try more data examples ex. image dataset, financial time-series
  - Model evaluation using metrics
  - Parametrize $T(x)$ with more complicated methodologies ex. Neural Network
- Build theories
  - Stability theory, Asymptotic theories

Thank you!