UNIVERSITÀ
DEGLI STUDI
DI MILANO

Department of Economics, Management and Quantitative Methods (DEMM)

Functional and Topological Analysis of Australian Weekly Mortality Data

A Case Study using FDA and TDA Techniques

Author: Luca Codeluppi

Student ID: 32289A

Course: Functional and Topological Data Analysis – Data Science for Economics

Abstract

This paper applies Functional Data Analysis (FDA) and Topological Data Analysis (TDA) to weekly mortality data for Australia (AUSmortality database, 2015–2025). We represent mortality profiles as smooth functional curves over time and age, study dominant modes of variation via functional principal components, quantify the relationship between mortality levels and age composition using functional regression, Compare pre- and post-pandemic periods with functional ANOVA, and explore the topological structure of the mortality time series through persistent homology. The aim is to complement standard descriptive analysis with functional and topological tools.

Contents

1 Introduction

Understanding how mortality evolves over time is crucial for public health monitoring and demographic analysis. The AUSmortality dataset provides weekly counts of deaths in Australia disaggregated by age group, sex and year, making it possible to investigate short-term fluctuations and structural changes such as those associated with the COVID-19 pandemic.

Because each observation corresponds not only to a number but to an age profile and to a weekly trajectory, these data naturally lend themselves to Functional Data Analysis (FDA), where observations are treated as smooth curves rather than discrete vectors. FDA enables us to describe weekly mortality as a function of age, extract dominant modes of variation, compare groups of years, and model how age composition influences overall mortality through functional regression.

To complement these functional tools, we also apply Topological Data Analysis (TDA) to the weekly mortality series. Using delay-embedding and persistent homology, TDA allows us to study the "shape" of mortality dynamics—identifying recurrent patterns, structural changes, or differences between time windows that are not detectable using standard linear methods.

So this study integrates FDA and TDA to provide a more general analysis of Australian weekly mortality, with a focus on detecting temporal patterns, age-specific variation, pandemic effects, and the underlying topological structure of mortality fluctuations.

The main goal of this project are:

- Construct functional representations of Australian weekly mortality;
- Identify dominant modes of variation in age-specific mortality via FPCA;
- Model the relationship between mortality levels and age composition using functional regression;
- Compare pre-pandemic and pandemic periods with functional ANOVA.
- Explore the topological structure of weekly mortality dynamics using TDA;
  - Transform the 2020 time series of total weekly deaths into a point cloud via sliding-window embedding (after scaling and detrending);
  - Compute Vietoris–Rips persistence diagrams and barcodes to describe the persistent homological features (H0; H1) of the embedded dynamics.
  - Compare the topology of different sub-periods of 2020 (e.g. weeks 5–15 vs 30–40) by means of the bottleneck distance between their persistence diagrams, assessing whether distinct phases of the year exhibit qualitatively different mortality patterns.

2 Dataset and Preprocessing

2.1 Data description

Briefly describe AUSmortality:

- Country: Australia (ISO code `AUS`);
- Time span: 2015–2025;
- Weekly frequency, with variables Year, Week, Sex, age groups 0–14, 15–64, 65–74, 75–84, 85+, total deaths and proportions.

Basic statistics Summary(min, max, mean, etc.):

```
CountryCode            Year          Week       Sex                 age_0_14       age_15_64      age_65_74
Length:1629        Min.   :2015   Min.   : 1   Length:1629        Min.   : 6.669   Min.   :152.0   Min.   :140.0
Class :character   1st Qu.:2017   1st Qu.:13   Class :character   1st Qu.:12.146   1st Qu.:210.1   1st Qu.:218.0
Mode  :character   Median :2020   Median :26   Mode  :character   Median :14.768   Median :339.6   Median :312.0
                   Mean   :2020   Mean   :26                      Mean   :17.335   Mean   :359.4   Mean   :342.9
                   3rd Qu.:2022   3rd Qu.:39                      3rd Qu.:24.042   3rd Qu.:518.2   3rd Qu.:484.0
                   Max.   :2025   Max.   :53                      Max.   :33.246   Max.   :629.0   Max.   :659.0
  age_75_84       age_85_plus         total        prop_0_14          prop_15_64          prop_65_74
Min.   : 281.0   Min.   : 368.0   Min.   :1277   Min.   :0.0001521   Min.   :0.0009446   Min.   :0.006936
1st Qu.: 400.0   1st Qu.: 598.0   1st Qu.:1559   1st Qu.:0.0002572   1st Qu.:0.0013254   1st Qu.:0.009350
Median : 481.0   Median : 769.0   Median :1750   Median :0.0002885   Median :0.0016980   Median :0.011515
Mean   : 573.2   Mean   : 882.4   Mean   :2175   Mean   :0.0002885   Mean   :0.0017053   Mean   :0.011635
3rd Qu.: 775.0   3rd Qu.:1189.0   3rd Qu.:2999   3rd Qu.:0.0003183   3rd Qu.:0.0020646   3rd Qu.:0.013672
Max.   :1163.0   Max.   :1854.0   Max.   :4221   Max.   :0.0004554   Max.   :0.0025365   Max.   :0.018098
   prop_75_84       prop_85_plus       total_prop          Split          SplitSex    Forecast       week_str
Min.   :0.02203   Min.   :0.1010   Min.   :0.005345   Min.   :1   Min.   :0   Min.   :0.000   Length:1629
1st Qu.:0.02974   1st Qu.:0.1211   1st Qu.:0.006237   1st Qu.:1   1st Qu.:0   1st Qu.:0.000   Class :character
Median :0.03412   Median :0.1311   Median :0.006649   Median :1   Median :0   Median :0.000   Mode  :character
Mean   :0.03464   Mean   :0.1340   Mean   :0.006710   Mean   :1   Mean   :0   Mean   :0.326
3rd Qu.:0.03926   3rd Qu.:0.1441   3rd Qu.:0.007100   3rd Qu.:1   3rd Qu.:0   3rd Qu.:1.000
Max.   :0.05560   Max.   :0.2031   Max.   :0.008893   Max.   :1   Max.   :0   Max.   :1.000
```

Figure 1: Statistic Summary

2.2 Data cleaning and transformations

Some helpful operations:

- Change columns name (`age_0_14`, `total`, ...);
- Select both sex (`Sex = "b"` for both sexes);
- Time variable creation `date` from Year/Week (with `ISOweek2date`);
- No outliers or duplicates were found

The Figure 2 shows weekly mortality in Australia from 2015 to 2025 by sex. All three series exhibit a clear and persistent seasonal pattern, with peaks occurring during the Australian winter and troughs in the summer months. The total mortality naturally lies above the male and female series, while male mortality is consistently higher than female mortality, reflecting well-known demographic differences (more men the women).

Starting around 2020, the total series shows higher volatility and some unusually elevated peaks, likely associated with the indirect and direct effects of the COVID-19 pandemic. Despite these fluctuations, the underlying seasonal structure remains stable throughout the entire period.

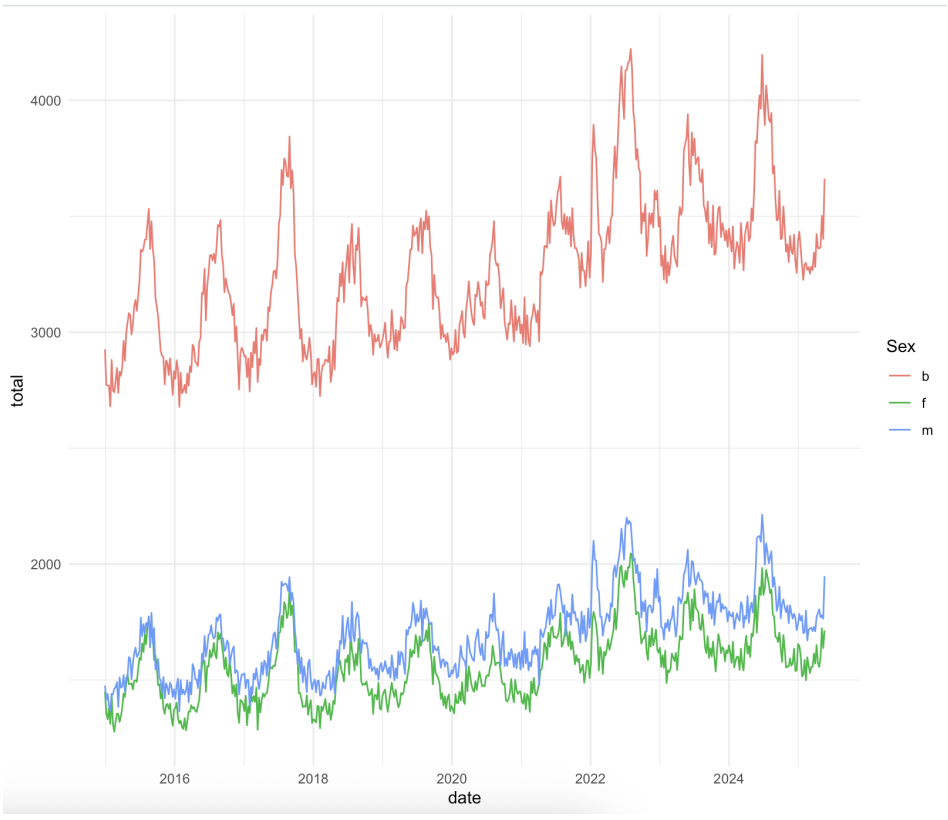


Figure 2: Weekly total deaths in Australia, both sexes, 2015–2025.

3 Functional Data Analysis

3.1 B-spline smoothing of weekly mortality by year

For each year y we consider the vector of total weekly deaths

$$\{D_{y,1}, D_{y,2}, \ldots, D_{y,W_y}\},$$

where W_y denotes the number of observed weeks in year y. Instead of treating these values as isolated points, we model them as a smooth function of the week index,

$$f_y : t \in [1, 53] \longmapsto f_y(t),$$

obtained by B-spline smoothing. This allows us to interpret yearly mortality as a continuous trajectory over the calendar year and to compare curves across years.

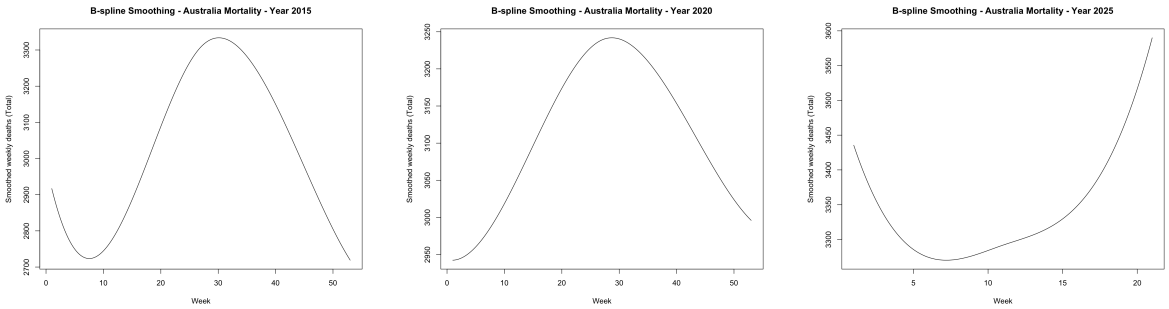
The following R code illustrates the construction of the B-spline basis and the smoothing step for a generic year:

```
[language=R, caption={B-spline smoothing of weekly mortality by year.}]
# time_index: week numbers (1,...,52/53)
# deaths: total weekly deaths for a given year

basis_spline <- create.bspline.basis(
  rangeval = c(1, 53),
  nbasis   = 25,
  norder   = 6
)

fdParobj <- fdPar(basis_spline, Lfdobj = 2, lambda = 0.1)

smooth_fd <- smooth.basis(time_index, deaths, fdParobj)$fd
```



Pre COVID During COVID Post COVID

Figure 3: B-splines curves in three different periods

The first two graphs in the Figure above are pretty similar, while the third one it is different because it's only about the first 5 months of 2025, so not the entire year.

3.2 Age-specific mortality profiles and FPCA

For every combinations (Year, Week, Sex = b), I have build a function $f_{y,w}(\text{age})$ using the five differents groups of age. Then I have applied the FPCA to identify the modes of variations in the mortality curves:

$$f_{y,w}(\text{age}) = \mu(\text{age}) + \sum_{k=1}^{K} \xi_{y,w,k}\, \phi_k(\text{age}),$$

where ϕ_k are the functional eigenvectors and $\xi_{y,w,k}$ the scores. We can see the results in Figure 4.

```
[language=R, caption={Functional PCA.}]

smooth_fd <- smooth.basis(age_grid, y_values, fdParobj)$fd

pca_results <- pca.fd(weekly_fd_data, nharm = 3)
```

- The **1st Principal Component**, which explains the vast majority of variability (approximately **0.82**), reflects the dominant age-mortality pattern: it peaks around ages 20–40, dips strongly around ages 55–65, and rises again at older ages, indicating that the primary source of variation is a broad shift in mortality levels across age groups, especially in middle adulthood and at the oldest ages.
- The **2nd Principal Component**, accounting for about **0.16** of the variability, shows a clear increasing trend with age, starting slightly negative at younger ages and rising steadily toward very old ages, suggesting a component primarily driven by differences in how mortality changes between young and elderly populations.
- The **3rd Principal Component**, which explains only around **0.016** of the variance, displays a mid-life peak (around ages 40–60) followed by a sharp decline at older ages.

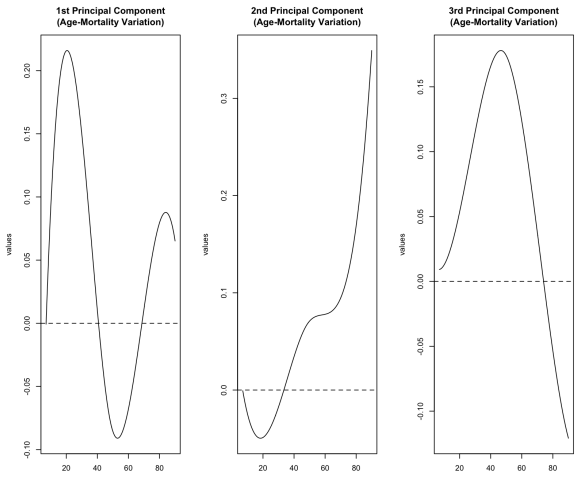


Figure 4: Functional PCA

3.3 Functional regression over weeks

To investigate how the age composition of deaths affects the temporal pattern of mortality, we model weekly total deaths as a functional response over the week index. For each year y we consider the smoothed curve

$$f_y(t), \qquad t \in [1, 53],$$

representing total deaths as a function of the week t, and we construct a corresponding functional covariate $x_y(t)$ given by the weekly proportion of deaths in the oldest age group (85+). Both f_y and x_y are expanded on a common B-spline basis over the domain of weeks, and we estimate the functional linear model

$$f_y(t) \;=\; \beta_0(t) \;+\; \beta_1(t)\, x_y(t) \;+\; \varepsilon_y(t),$$

where $\beta_0(t)$ and $\beta_1(t)$ are coefficient functions. As shown in Figure 5, the seasonal pattern is particularly pronounced at the beginning and at the end of the year, where the coefficient $\beta_1(t)$ reaches its highest values. These peaks coincide with the warmest months in Australia (from November to February), during which elevated temperatures are known to increase mortality risks, especially among older age groups.

```
[language=R, caption={Functional PCA.}]
norder_week <- 6
nbasis_week <- 25
fregress_time <- fRegress(
  y = deaths_fd_combined,
  xfdlist = xfdlist,
  betalist = betalist
)
plot(
  fregress_time$betaestlist$prop85,
  xlab = "Week",
  ylab = expression(beta[1](t) ~ "for prop85(t)"),
  main = "Functional regression coefficient over weeks"
)
```

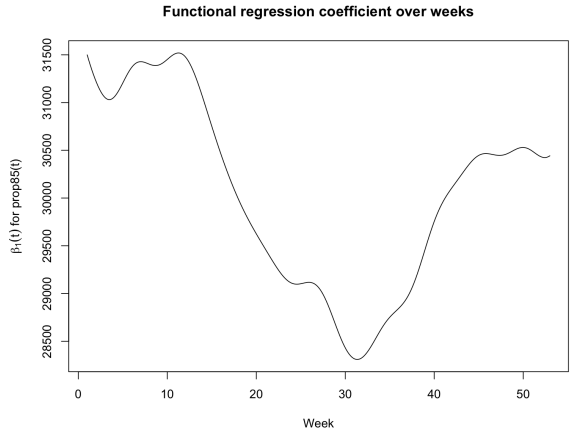
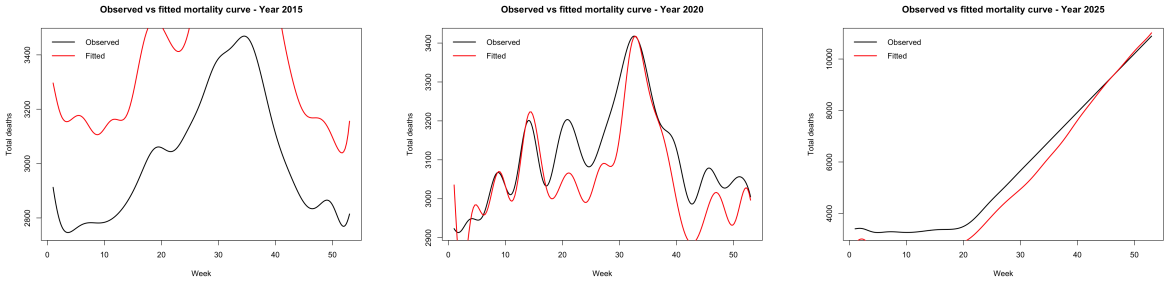


Figure 5: 85+ mortality over weeks

For each year we compare the observed mortality curve $f_y(t)$ with its fitted counterpart $\hat{f}_y(t)$ obtained from the regression model. Plotting the observed and fitted curves together provides a visual assessment of the goodness of fit over time, highlighting years or specific weeks where the model succeeds or fails in capturing the observed mortality dynamics.



Pre COVID During COVID Post COVID

Figure 6: Observed vs Fitted Mortality curve

3.4 Functional ANOVA (FANOVA)

Two different groups:

- Group A: 2015–2019 (pre-covid),
- Group B: 2020–2025 (post covid).

The two in Figure 7 curves display a similar overall shape, with mortality increasing sharply at younger ages, decreasing around middle ages, and rising again among the oldest groups. However, Group B shows consistently higher mortality levels across almost all ages, with a particularly pronounced increase in the 20–40 and 80+ age ranges. This indicates a structural shift in weekly mortality profiles during the pandemic years, with elevated deaths both among older individuals—as expected—and among some younger adult groups, suggesting broader demographic impacts beyond the oldest age classes.

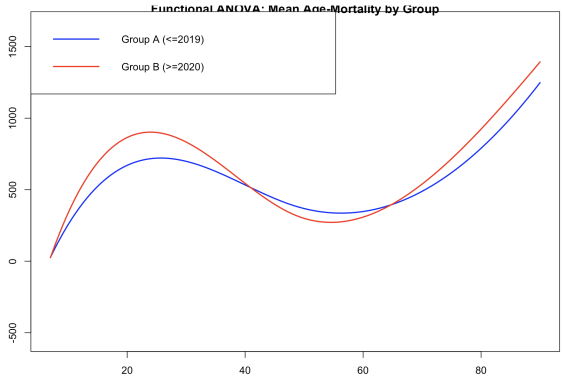


Figure 7: Functional ANOVA

Something interesting to see is represented in Figure 8, the F-statistic curve shows two pronounced peaks around ages 75 and 90, indicating that the mortality structure changed most in older and very old individuals. This pattern is consistent with the demographic impact of the COVID-19. A **global p-value of 0** indicates a highly significant difference between the pre-2020 and post-2020 functional mortality profiles.

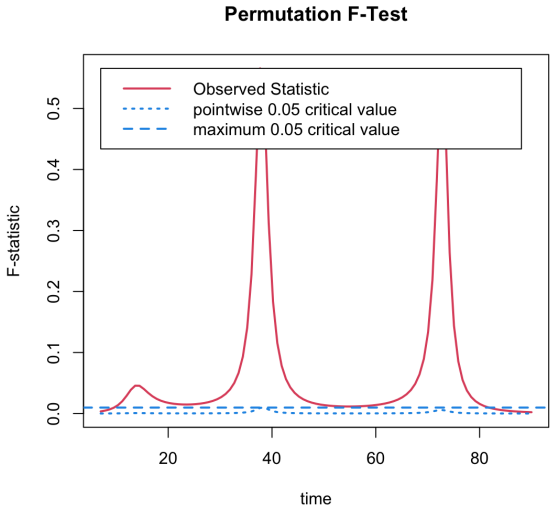


Figure 8: Permutation Test

4 Topological Data Analysis

4.1 Delay embedding and point cloud construction

To analyse the temporal structure of weekly mortality as a geometric object, the time series of total deaths is transformed into a point cloud through a sliding–window (delay) embedding. Given the sequence $\{z_t\}_{t=1}^{T}$, the embedding maps each time index t into a d–dimensional vector $(z_t, z_{t+\tau}, \ldots, z_{t+(d-1)\tau})$, where d is the embedding dimension and τ the delay. This procedure reconstructs the underlying dynamics of the series in a higher–dimensional space, allowing topological features such as clusters or loops to emerge in the geometry of the embedded trajectory.

```
sliding_window <- function(ts, dim = 3, delay = 1)
point_cloud <- sliding_window(ts_detr, dim = 5, delay = 2)
```

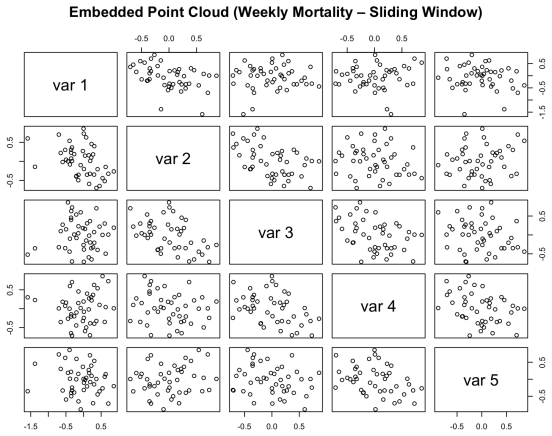


Figure 9: Embedded Point Cloud

As we can see from Fivure 9 the point cloud appears well spread across the embedding space, without forming clear clusters or geometric structures such as loops. This suggests that the short–term dynamics of weekly mortality do not follow a strongly recurrent or periodic pattern at this resolution.

4.2 Persistence diagrams

Once the point cloud is obtained, we compute its persistent homology using the Vietoris–Rips filtration. As the scale parameter increases, simplicial complexes are constructed and topological features (connected components and loops) appear and disappear. The persistence diagram summarises these features as points (b, d), where b denotes the scale at which a feature is born and d the scale at which it dies. Long–lasting features correspond to points far from the diagonal and indicate meaningful geometric or dynamical structure.

Listing 1: Persistent homology from the embedded mortality series.

```
diag <- ripsDiag(
  X            = point_cloud,
  maxdimension = 1,
  maxscale     = 1,
  dist         = "euclidean"
)
```

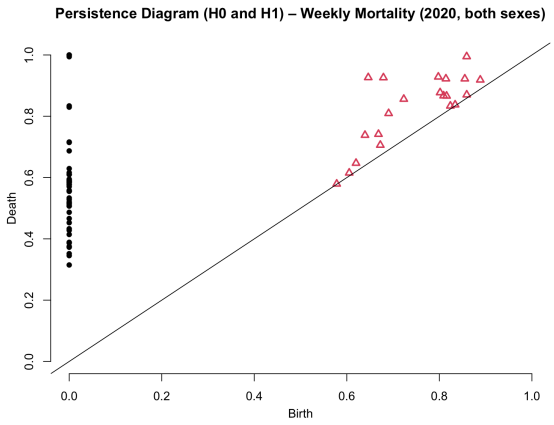


Figure 10: Persistence Diagram (H0 and H1)

As we can see from Figure 10

- The black dots correspond to H_0 features (connected components), all of which are short–lived, indicating noise.
- The red triangles represent H_1 are suggesting a non-negligible one-dimensional topological structure

4.3 Persistence Barcode Diagrams

An equivalent visual representation is provided by the persistence barcode, where each topological feature is displayed as a horizontal segment spanning from its birth scale to its death scale. This representation highlights the lifespan of features more directly than the diagram. Short bars typically correspond to noise, while long bars reflect persistent structures in the geometry of the embedded mortality trajectory.

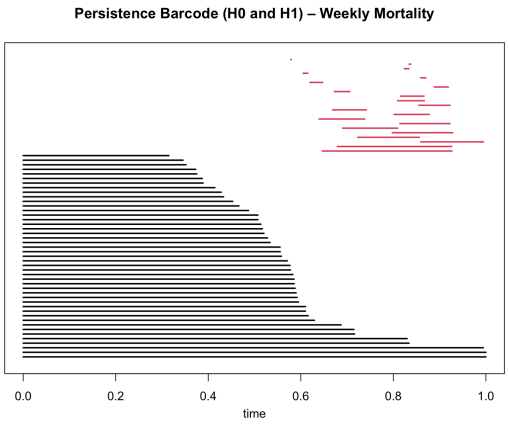


Figure 11: Persistence Barcode (H0 and H1)

4.4 Bottleneck distance between periods

To compare the topological structure of different segments of the mortality series (e.g., early vs. late weeks of 2020), we compute the bottleneck distance between their persistence diagrams. This metric quantifies the minimum amount of perturbation required to match the features of one diagram to those of the other. A bottleneck distance of **0.5659** indicates that the two time windows share similar topological patterns.

Listing 2: Bottleneck distance between two mortality windows.

```
bdist <- bottleneck(diag1[["diagram"]], diag2[["diagram"]])
```

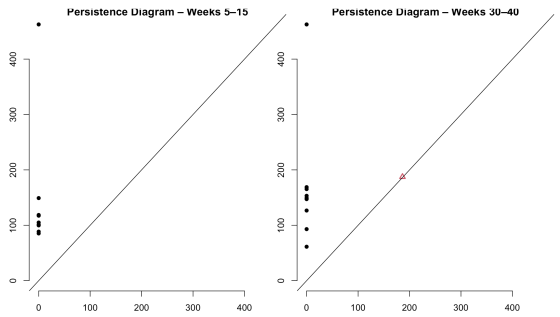


Figure 12: Bottleneck distance weeks 5-15 vs 30-40

Figure 12 compares the persistence diagrams obtained from two distinct temporal windows of the 2020 mortality series: Weeks 5–15 and Weeks 30–40. In both diagrams, the H_0 features (black dots) are short-lived and clustered very close to the diagonal, indicating that neither window exhibits meaningful topological separation or clustering structure.

The key difference lies in the H_1 features. In the early-season window (Weeks 5–15), no persistent loops emerge, suggesting a relatively smooth and evolution of mortality levels. Conversely, the later-season window (Weeks 30–40) shows a single H_1 feature with moderate persistence (red triangle), indicating the presence of a weak cyclic or recurrent pattern. This contrast suggests that mid-to-late season mortality may display more structured short-term oscillations compared to the early season.

5 Discussion and Conclusions

In this study, we applied a combined **Functional Data Analysis (FDA)** and **Topological Data Analysis (TDA)** framework to the Australian weekly all-cause mortality dataset (AUS, 2015–2025). Our objective was to extract both smooth functional structures and deeper geometric patterns from a high-dimensional demographic time series. The joint use of FDA and TDA allowed us to characterize mortality dynamics across years, age groups, and different epidemiological phases, most notably the COVID-19 period.

5.1 Insights from Functional Data Analysis

FDA provided a rich statistical representation of mortality patterns.

- **B-spline smoothing** enabled the reconstruction of yearly mortality trajectories, removing noise while preserving seasonal structure. These smoothed curves revealed clear annual cycles, with summer peaks and winter troughs, consistent with climatic and epidemiological seasonality;
- **Functional PCA** identified the main modes of variation in age-specific mortality. PC1 captured the overall mortality level across all ages, while PC2 highlighted contrasts between younger and older age groups. These components offered a compact representation of demographic shifts over time;
- **Functional regression** over weeks quantified how the weekly age distribution of mortality explains fluctuations in total mortality. The estimated coefficient functions showed that the contribution of older age groups varies over the year, and fitted-vs-observed comparisons confirmed that weekly functional predictors capture meaningful structure in the seasonal cycle;
- **Functional ANOVA** revealed systematic differences between the pre-pandemic period (2015–2019) and the COVID-affected years (2020+). Group B displayed consistently higher mortality at advanced ages and larger curvature in the functional mean, indicating an intensification of age-specific mortality disparities;

5.2 Insights from Topological Data Analysis

TDA provided a complementary geometric perspective on the temporal structure of mortality. Using delay-embedding, we reconstructed the 2020 weekly mortality series into a high-dimensional point cloud, capturing its dynamical behaviour in phase space.

- The **persistence diagrams** showed many short-lived H_0 features—interpreted as noise—but also a few longer-lived structures indicating more stable patterns in week-to-week dynamics. Persistence barcodes reinforced this interpretation by highlighting limited long-range topological persistence.
- The **Bottleneck distance** was used to compare the topological signatures of two mortality periods (weeks 5–15 vs. 30–40). The near-zero distance suggests that although the absolute mortality levels differ, the underlying dynamical "shape" of fluctuations remains similar across these two seasonal windows, indicating structural stability within the same year.

5.3 Final Remarks

Overall, combining FDA and TDA proved highly effective for understanding mortality dynamics. FDA delivered smooth functional models, interpretable decompositions, and inferential comparisons across years. TDA added a geometric dimension, revealing the persistence and similarity of temporal patterns beyond what classical statistics alone can capture.

Together, these methods provide a multi-layered quantitative description of Australian mortality, capturing trend, seasonality, age structure, and dynamical complexity—an approach that may be extended to other countries, demographic processes, and epidemiological contexts.

Future Developments

Several directions may extend the present work. First, the FDA–TDA framework could be applied to mortality data from other countries, allowing for cross-country comparisons and the identification of common or divergent functional patterns.

Second, future analyses could include external covariates such as temperature, influenza activity or COVID-19 indicators, integrating them into functional regression models to better explain variations in weekly mortality.

Third, the topological component may be enriched by using persistence landscapes, silhouettes or other stable representations, enabling more robust statistical inference on topological features.

Finally, functional and topological descriptors could be combined in predictive models for real-time mortality monitoring, supporting early detection of unusual seasonal or epidemiological dynamics.