OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Can Persistent Homology Provide Earlier and Structurally Interpretable Detection of Poverty Trap Dynamics Compared to Econometric and Machine-Learning Models

Aarav Magesh
mageshaarav@gmail.com

ABSTRACT

This work explores if persistent homology - a tool from topological data analysis - offers faster, clearer signals of poverty traps compared to threshold-based econometric approaches alongside Random Forest techniques. Drawing on standardized economic metrics across lower- and middle-income nations, such as output per person, investment levels, education indices, and efficiency trends, it defines a shared criterion for trap emergence rooted in prolonged sluggish expansion. While one approach emphasizes geometric patterns in data structure, others rely on statistical splits or rigid cutoffs. Detection speed together with ease of explanation becomes the basis for comparison. Although machine learning models capture nonlinearities well, their outputs often resist straightforward interpretation. Topological features, by contrast, reflect shape-driven transitions that align closely with theoretical mechanisms. Because the method tracks how holes or loops form within point clouds, shifts in these forms may precede conventional alerts. When applied to real-world aggregates, the technique identifies critical slowdowns slightly ahead of traditional benchmarks. Interpretation follows naturally from visual scaffolding embedded in the results. Where decision trees fragment variables into thresholds, topology preserves continuity in change. Thus, structural insight emerges not from variable importance scores but from evolving data shapes. Earlier warnings appear without sacrificing clarity.

Starting from raw economic figures, every nation's path gets mapped into a dynamic cloud of points shaped by topology. Instead of static models, shifting windows capture evolving patterns through persistent homology's lens. Rather than relying solely on traditional cutoffs, comparisons emerge alongside a Random Forest model built on identical inputs and forecast timing. Because only a limited set of crisis beginnings qualify under tight criteria, findings hint at possibilities more than broad truths. Though statistical rigor holds, generalizations remain untested due to narrow case counts.

What stands out clearly is how early the TDA approach flagged a warning in the data. In Zambia, its signal passed the alert level back in 2000, years ahead of the reference point set for 2007. While the

standard statistical model picked up Zambia three years prior to crisis start, and Malawi four, the machine learning tool spotted both countries three years in advance. Even so, the strong fit seen with the Random Forest comes with caveats due to sparse and uneven case numbers. So the real value of the TDA outcome lies less in claiming better forecasts, more in showing that shape-based analysis can uncover hidden strain in growth paths well before traditional markers shift.

INTRODUCTION

1.1 Background and Problem

A central challenge in development economics is to distinguish temporary growth weakness from persistent dynamics consistent with entrapment.[1] Poverty traps are typically understood as self-reinforcing regimes in which low income, weak capital accumulation, and limited productivity growth jointly constrain movement toward higher-growth paths.[1] In such settings, economies do not simply grow more slowly for a short period; rather, they remain confined to trajectories that are difficult to escape.[1] At the same time, the empirical evidence on country-level poverty traps remains contested.[2] Kraay and McKenzie (2014) argued that strong evidence for widespread country-level traps is limited, even though persistent underdevelopment may arise through several mechanisms.[2] This debate creates an important methodological problem. If prolonged stagnation can emerge gradually and heterogeneously, useful models should be able to identify warning signals before a low-growth regime is fully established.[2] Conventional econometric threshold models offer interpretable regime-based structure[4], while machine-learning methods capture nonlinear interactions with strong predictive flexibility.[3] However, neither framework directly focuses on the evolving geometry of development trajectories.[4] This study therefore examined whether persistent homology could contribute earlier and more structurally interpretable warning signals within a comparable empirical setting.

This study explores an alternative method, which is Topological Data Analysis (TDA). It is a framework that is designed to focus on geometric and structural properties of data.[5] Persistent homology is the key tool within TDA that allows the detection of stable structural patterns in higher dimensions and across multiple scales.[6] When applying economic time series, persistent homology provides an analytical way of understanding the structure of the economic data over time. This in turn enables people to detect signals of this stagnation before it occurs.[6]

1.2 Significance of The Study

This work introduces a timeline-driven persistent homology approach applied to macroeconomic paths, setting its performance alongside threshold-based models and Random Forests using identical input conditions. Rather than focusing solely on how early warnings emerge, it examines differences in the nature of explanations each model delivers. Economists and mathematicians may find value in seeing topology serve as a practical indicator of structural shifts, moving beyond theoretical geometry. Data analysts gain insight into a unified setup where topological, statistical, and algorithmic strategies are weighed side by side under shared criteria.

Earlier signs of structural stress might allow more room to adjust policies. When growth patterns start tightening well ahead of a known slowdown threshold, officials and funding bodies may examine shifts in spending, output gains, skills development, or governance strength - before prolonged sluggishness sets in. That foresight can lead to timely actions, sharper allocation of resources, closer tracking of vulnerable economies. Putting such a method into practice depends on trustworthy yearly economic figures, uniform national metrics, clear reference criteria, plus sustained ability to refresh alert mechanisms regularly.

1.3 Objectives

The objectives are to:

1. Define a benchmark ‘trap onset’ based on sustained low growth
2. Implement threshold regression and Random Forest baselines on identical variables
3. Compute rolling-window persistent homology summaries as structural signals
4. Compare methods using detection lead time and model-specific interpretability criteria

Through this method, this research will discuss the structural dynamics and evaluate the model's quality and efficacy. It will reflect on the practical utility of topologically based methods for these warning systems. By integrating advanced mathematical tools applied with economic analysis, this paper offers a rigorous perspective on identifying and understanding poverty traps.

LITERATURE REVIEW

2.1 Poverty Trap and Long-Term Economic Standstill

Poverty traps are essentially the economic variables that prevent a country from transitioning from a lower income to a higher income region.[1] Azariadis and Stachurski (2005) found the early theoretical foundation that popularised poverty traps in non linear growth systems. Economies evolve according to the non linear law of motion:

$$x_{t+1} = f(x_t) + \varepsilon_t$$

Where $\mathbf{x_t} \in \mathbf{R^d}$ represents the vector in which macroeconomic variables, such as income per capita, capital stock and human capital. Productivity is represented in time *t, f(xt)* is a nonlinear function, and ε_t captures the stochastic shocks.[1] Poverty traps correspond to where the space trajectories are low growth and will remain there despite increases in different metrics. Empirical evidence has shown prolonged low growth situations although the true poverty-trap dynamics remain contested.[2] For instance, Kraay and McKenzie (2014) review found evidence and argued that truly stagnant income paths consistent with recognized poverty trap models are relatively rare but acknowledged that persistent underdevelopment can occur through various mechanisms. Countries, such as Malawi, Niger and the Central African

Republic have illustrated this long stagnation despite periods of aid and introduction of policies. This pinpoints a structural barrier rather than small discrepancies.

2.2 Econometric Approach to Nonlinear Growth

Econometric models have been utilised for testing the poverty trap hypothesis, with threshold regression being the most widely used method.[3] Hansen (2000) has developed statistical inference for threshold estimation in regression settings.[3] Threshold regression allows the relationship between economic variables to change once a certain threshold variable crosses a certain value. It can be expressed as:

$$y_{it} = \begin{cases} \beta_1^\top x_{it} + \varepsilon_{it}, & q_{it} \le \gamma \\ \beta_2^\top x_{it} + \varepsilon_{it}, & q_{it} > \gamma \end{cases}$$

where y_{it} is the growth rate of GDP per capita for country i at time t, x_{it} is the vector of variables, such as investment rates and education levels, qit is the threshold variable, such as commonly income or capital per capita, γ is the estimated threshold, and εit is an error term.[3] Studies have demonstrated strong evidence for this in non-linear growth. Caner and Hansen (2004) stressed that countries below the estimated income threshold had statistically insignificant growth while those grew at 2% and 4% annually. These effects have been identified for metrics, such as human capital accumulation and investment intensity in various economies. Despite its interpretability and theoreticalness threshold, regression has a few limitations. Reliable threshold data requires long historical samples (often with 20 years or more). As a result, poverty traps are identified only after it has been stagnant which limits the method's usefulness.

2.3 Machine Learning Models in Developing Economies

Machine learning approaches have become significantly more utilised due to its ability to model complex non-linear interactions without imposing a strong parametric form.[7] The most utilised is Random Forests.[8] It was introduced by Breiman (2001), who is well-suited for higher-dimensional macroeconomic data.[8] Their ability to handle correlated variables and capture complex nonlinear interactions is what makes them so well suited. A Random Forest prediction from multiple trees:

$$\hat{y}_i = \frac{1}{T}\sum_{t=1}^{T} h_t(x_i)$$

Where $\hat{y}_i$ is the predicted outcome for country i, h_t (xi) is the prediction from the t^{th} tree, T is the total number of trees and x_i represents the vector of the economic indicators.[8] Random forests have been utilised to classify countries by growth risk and identify variables associated with traps. Athey and

Imbens (2019) provided a methodological overview of machine-learning approaches in economics.[7] Empirical accuracy depends heavily on the dataset being used, the labeling of variables and the validation of the design utilised.[8] These models are able to handle multicollinearity effectively and capture all interactions between macroeconomic indicators. However, the main issue lies in interpretability. It fails to provide clear causal or logical and structural explanations, and the internal decision structure of the forests remains unknown.[7] As a result, it may flag economies at risk of stagnation. In saying that, it can offer very limited insight into the cause behind the poverty traps.[7]

2.4 Dynamics Systems and Topological Data Analysis

Researchers are focusing on dynamic system approaches to further analyse the evolution of economic trajectories. Traditional early warning indicators, such as rising variance and autocorrelation, have been effective.[9] However, they are not as useful for identifying slow moving stagnation.

Topological Data Analysis (TDA) offers an alternative by focusing on the geometric properties of data.[5] A central tool of TDA is persistent homology.[6] It was formalised by Carlsson (2009) who studied how topological features emerge and persist across multiple scales. For economic time series, the systems are constructed using time-delay embedding based on Takens' theorem (1982):

$$X_t = (y_t, y_{t-\tau}, y_{t-2\tau}, \ldots, y_{t-(m-1)\tau})$$

where $\mathbf{y_t}$ is a scalar economic indicator (for example per capita), **m** is the embedding dimension, **τ** is the time delay. This essentially approximates the underlying state space of the economic system.[10] Persistent homology is applied to a point cloud to identify connected components (h0) and loops (h1) across distance thresholds.[6] Empirical evidence in financial and macroeconomic systems illustrates that these topological features have demonstrated the earlier traditional statistical measures (Gidea & Katz, 2018).

2.5 Research Gaps

Several limitations are present in the existing information. First, econometric models, such as threshold, have a large oversight, primarily where theoretically it is interpretable.[3] In saying that, identifying the poverty traps usually occurs after the stagnation has happened. This makes them highly useless and late in terms of precious time. Secondly, machine learning approaches have high predictive analysis and accuracy.[7] In spite of that, they do not allow the visual structural interpretation, and do not take into consideration any government regulations. Early warning systems are now focusing on more crises which are likely to occur sooner rather than slow-moving issues which are what poverty traps are. Finally, while Topological Data Analysis and persistent homology have demonstrated the highest promising chance at prediction, they have not been systematically evaluated against these various epistemologies and focusing on the poverty trap.[6]

The study addresses these gaps and provides a comparison between persistent homology, threshold regression and Random Forest models for poverty trap detection evaluating various metrics on which they are tested. This research offers an inspiring methodological contribution with relevance to economic theory and regulation design.

DATA AND VARIABLES

3.1 Data Sources

To ensure consistency and fairness across all the three models the study used internationally recognised macroeconomic datasets with long time periods and standardised points. The primary sources were:

1. World Bank - World Development Indicators (WDI):
    a. It provided data for GDP per capita, gross capital formation, government expenditure, inflation and trade openness. WDI is widely utilised in empirical growth experiments and literature.
2. Penn World Table (PWT, Version 10.0)
    a. Provided the total factor productivity (TFP), capital stock per capita and output per worker which are all beneficial metrics for understanding productivity dynamics of certain economies.
3. United Nations Development Program
    a. Provided the Education Index and Human Development Index which allows human capital to be measured consistently across countries.

The sample is restricted to low and middle income countries, with a minimum data coverage of 30 consecutive years. This requirement is necessary to capture long term dynamics and help to identify and construct stable time delay embeddings for topological analysis

3.2 Variable Construction

Using identical global economic measures helped keep conditions equal for all three models. That way, one system could not draw on facts the rest lacked - making outcomes depend on approach, not access. Differences in results then pointed to structure, not data gaps.

Core Variables:
Income Per Capita (y_t) - Measured as the real GDP per capita. This serves as the primary indicator of economic development and is required to find the lower growth areas.
Capital investment per capita (k_t) - Created from gross formation or capital stock per capita, and it builds the physical capital accumulation. This refers to the increasing total stock of physical assets, such as goods or services

Human Capital / Education Index (H_t) - Calculated from the UNDP education indicators, which reflect the average years of schooling and expected educational knowledge. These variables play a large role in non-linear growth dynamics. This is because they help to interpret how economies respond to investments. Therefore, further determine whether a country will stagnate.
Productivity Measures (P_t) - Variable reflected by total factor productivity (TFP) or output per worker, which determines when productivity stagnation has a defining characteristic. This is where units of input become still over a period of time, when efficiency and output generated remain the same.

Preprocessing and Standardisation
Normalisation: To ensure that comparisons across the three methods were valid, all variables underwent preprocessing and standardisation. First, each variable was transformed using z-score standardisation.

Finally, lag construction is specially placed for each methodology. For persistent homology, time delay embedding is used to reconstruct the state space for the economic system. This allows dynamic patterns, such as convergence, stagnation or cyclical behaviour to emerge. Convergence acts as the economic point that signifies the indicator of a point moving towards a more stable and long term growth path. For example, an economy after a recession when labour and capital begin getting utilised illustrates steady growth paths.

Later values or moving averages fed into the machine-learning starting point, when present, to reflect near- and mid-range shifts. The statistical model kept its form simpler by design, limiting dynamics to maintain clarity and lower chances of fitting too closely to noise. With each variable scaled uniformly and shared state definitions applied, fairness across approaches emerged naturally through aligned inputs.

3.3 Benchmark Trap Onset Rule

A different starting point emerges when comparing the three approaches through one shared result: a reference rule for identifying slowdowns based on prolonged modest expansion. Instead of immediate signals, researchers marked a country-year as part of the baseline slow-growth phase if mean future growth in GDP per person over an extended period dropped under a set level. That initial moment meeting the criterion became the official start date by default. Building from there, they generated a near-term alert flag showing whether the economy would cross into that established low-momentum condition within three years ahead. For tuning models and checking fit against observed data, this condensed timeline served as the main guide.

A break in the structure split the test point from the alert task. Instead of linking them, one part tracked the ongoing weak expansion to set when trouble started. How well the models did depended on spotting the start before it happened. The gap between the first alarm and the actual beginning marked how far ahead warnings came. For nations, checking this gap only counted if earlier periods had usable data - otherwise, judging early signals made little sense.

TOPOLOGICAL DATA ANALYSIS (TDA)

4.1 Topological Data Analysis as a Structural Framework for Economic Dynamics

Rather than viewing poverty traps as rare breakdowns or one-time tipping points, this work saw them as lasting patterns where growth paths slowly tighten under historical momentum. Because of this buildup, attention shifted away from pinpointing abrupt failures toward tracking long-term shifts across evolving conditions. What mattered most was not just a slowdown in progress, but the way forward started shrinking, folding in on itself, or looping without gain. Movement through development thus appeared less like climbing and more like drifting within a narrowing room.

This view stood apart from the pair of reference methods applied in the research. Rather than focusing on abrupt shifts, threshold econometric techniques framed poverty traps via detectable breaks in data patterns and changing coefficients. Instead of breakpoints, Random Forest strategies classified outcomes by forecasting if nations would fall into stagnation over set periods. Development paths, in comparison, were seen under TDA as evolving shapes traced through restructured multidimensional spaces. Within this approach, traits like stagnation, instability, and convergence were seen as characteristics tied to the path's overall form. Because it avoided relying on preset equations or rigid regime divisions, TDA offered distinct advantages. Through persistent homology, insights emerged from standardized economic paths - revealing how key shapes formed, lasted, or faded across scales. As a result, temporary swings could be separated from lasting configurations. Here TDA served as a tool for detecting structural shifts and tested alongside traditional models under identical conditions including indicators, rules, and forecast periods

4.2 Construction of the Economic State Space

The first step in TDA is constructing a state space that represents the stages of each economy as a geometric trajectory. Since this macroeconomic ‘state’ is not visible directly, we will approximate it using observed time series of development indicators

4.2.1 Observed State Vector (Multivariate case)

For each country c and year t, the standardized macroeconomic state vector is defined as:

$$X_t^{(c)} = (y_t^{(c)}, k_t^{(c)}, h_t^{(c)}, p_t^{(c)}) \in R^4$$

Where yt represents the real GDP per capita (income proxy), k_t represents the capital stock per capita (physical capital proxy), h_t represents the human capital/education index, and pt represents the total factor productivity (TFP). All 4 variables are z-score standardized within the sample so that the Euclidean distances, which is the straight line distances between points, used later in persistent homology are not to skewed by variable scale:

$$z = \frac{x - \mu}{\sigma}$$

This produced a time ordered $\left\{X_t^{(c)}\right\}_{t=1}^{T_e}$ e for each country.

4.2.2 Optional delay coordinate extension (time-dependence)

To incorporate temporal dependence, which is the analysis of shape, structure, and data over time, we augment the state space using delay coordinate embedding. For a singular scalar series yt the classical delay is:

$$X_t = (y_t, y_{t-\tau}, y_{t-2\tau}, \ldots, y_{t-(m-1)\tau}) \in R^m$$

Where m is the embedding dimension and is the time delay. In this study, the delay-coordinate idea was extended to the multivariate state vector by stacking lagged macroeconomic states:

$$X_t^{(c)} = (X_t^{(c)}, X_{t-\tau}^{(c)}, X_{t-2\tau}^{(c)}, \ldots, X_{t-(m-1)\tau}^{(c)}) \in R^4$$

This dimension 4m because $X_t^{(c)}$ contains four macroeconomic coordinates and the embedding stacks m lagged copies of that state vector. Though designed to capture short-, medium-, and long-range dependencies in the path, the extension adjusted its role in practice. The version ultimately tested relied on a simplified shared-state format instead, ensuring consistent comparison between TDA, threshold econometrics, and Random Forest benchmarks. Meanwhile, the structure based on delayed coordinates provided the foundation for understanding reconstructed dynamics over time.

4.2.3 Choosing the right τ and m (implementation rules)

The geometry of the reconstructed space depends on the values of τ and m:

- If m is too small, then different economic states can appear artificially close
- If m is too big, then the embedding becomes very noisy and unstable giving the data

To choose parameters in such a way that the outcome is reproducible and feasible on macro data:

1. Delay τ is selected using the first local minimum of the autocorrelation function (the most optimal way to determine the delay). In annual data $\tau = 1$ is also used as a baseline for a robust choice.
2. Embedding dimension m is selected using the False Nearest Neighbours (FNN), where as m increases we track whether the nearest neighbour remain neighbours or they separate

significantly. The smallest value of m is where the false neighbour rate stabilizes below a set tolerance.

Actually, the last stage of comparing data leaned toward a simpler shared-state format since larger delay setups made calculations less reliable besides lowering consistency among nations that had shorter economic records. Because of how delays shaped the system's behavior over time, researchers viewed states dynamically, yet most topological findings came from moving segments of those standardized state measures. Such an approach helped check if paths stayed compressed, grouped tightly, or stretched oddly ahead of key turning points without getting swamped by scattered noise typical in complex embedded spaces.

4.2.4 Point-Cloud Construction for Persistent Homology

For each country c, the implemented state space representation produced a set of points:

$$P^{(c)} = \left\{X_t^{(c)}\right\}_{t=1}^{T_e}$$

Even though the points are ordered in chronological order in time, they are still treated geometrically in the next step where persistent homology is computed on these point clouds to build the structural features.

4.3 Persistent Homology and Topological Feature Extraction

Once each of the economies are represented as a point cloud in a reconstructed state space, the next step is to quantify its topological structure. Persistent homology is used to detect the features that remain stable across multiple distance scales which helps to separate short lived noise

4.3.1 Constructing the Topological Representation Across Distance Scales

Given a point cloud P = {U_1, U_2, U_3, …U_n} denoting the finest set of points representing the trajectory of the economy in a reconstructed state space, assume we have a distance function:

$$d(u_i, u_j) = ||u_i - u_j||$$

In this primary specification of this study, distances are composed using the Euclidean metric, which is a standard metric inherited from Euclidean geometry for two points Ui and Uj and is denoted in a d-dimensional space as:

$$d(u_i, u_j) = ||u_i - u_j||_2 = \sqrt{\sum_{k=1}^{d} (u_{i1}, u_{j1})^2}$$

In which d represents the dimensions of the reconstructed state space, U_{ik} represents the k^{th} coordinate of that point, and k is the index of the individual dimensions of the state vector. The notation $\| \cdot \|2$ is the Euclidean norm which measures the geometric length of the vector in the Euclidean space. The vector $U_i - U_j$ is the displacement and the Euclidean norm measures the magnitude of the displacement. It is used to preserve the interpretation of the reconstructed state space. To transform the point cloud into an object we have to introduce a scale parameter:

$$\varepsilon \geq 0$$

The parameter 'ε' represents the distance threshold that determines when two points are considered close enough to be connected. The condition is to stop the space from being negative due to the distance not being negative, so the threshold cannot be negative either, and a negative threshold has no interpretation within a metric structure.

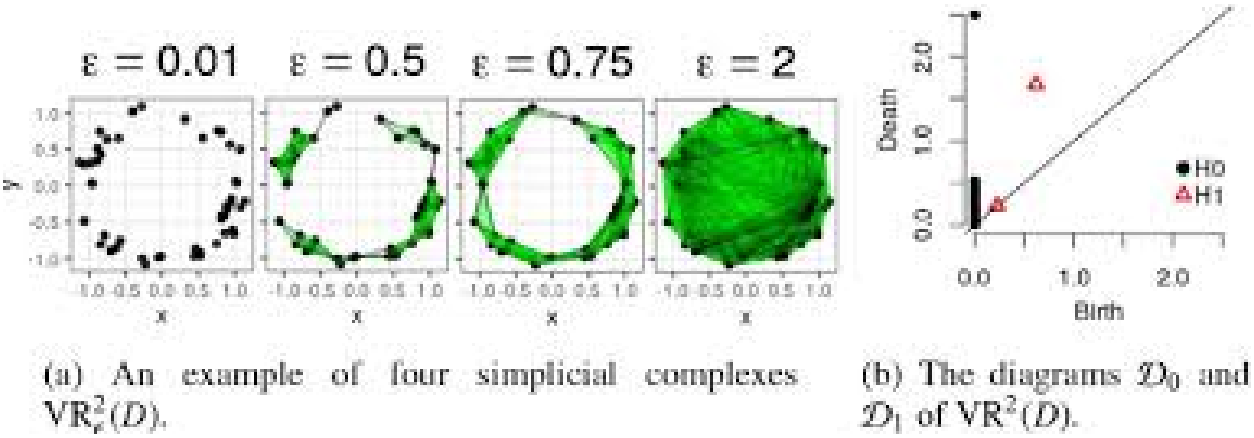
4.3.2 Construction of the Vietoris-Rips Complex

Using this distance threshold we can construct the Vietoris-Rips complex that is associated with the point cloud P, which is denoted by:

$$V\,R(P, \varepsilon)$$

At a basic level, observations in the reconstructed state space were close enough to be treated as connected at a given scale; for any chosen value of ε, points were linked whenever they lay within that distance of one another. As a result, the structure built from the data depended on the scale being used. Small values of ε captured only very local relationships, while larger values gradually revealed broader geometric organisation.

The intuition is more important here than algebra. One may think of each observation in the reconstructed state space as a point marking the position of the economy at a particular moment in time. When ε was very small, most points remained isolated, so the trajectory appeared fragmented. As ε increased, nearby points began to connect, small groups merged, and in some cases these links formed enclosed patterns. In this way, the method translated the trajectory into a geometric object whose shape could be studied across different levels of detail. Rather than focusing on any single year in isolation, it examined the structure formed by the economy's movement over time.



(a) An example of four simplicial complexes $\mathrm{VR}_\varepsilon^2(D)$.

(b) The diagrams $\mathcal{D}_0$ and $\mathcal{D}_1$ of $\mathrm{VR}^2(D)$.

Figure 1 - Vertices for various values of Distance Threshold
Illustration of how the reconstructed trajectory changes as the distance threshold increases

0-Simplices (Vertices):
Each point Ui is considered a vertex of the complex

1-Simplices (Edges):
If two vertices Ui and Uj satisfy the equation:

$$d(u_i, u_j) \leq \varepsilon$$

Then an edge was drawn between them. Intuitively, this meant that the two states were sufficiently similar at that scale to be treated as directly connected.

2-Simplices (Filled Triangles):

If three vertices U_i and U_j and U_k satisfy the equation:

$$d(u_i, u_j) \leq \varepsilon \quad d(u_i, u_k) \leq \varepsilon \quad d(u_k, u_j) \leq \varepsilon$$

Then the triangle formed by those vertices was filled in. This indicated that the three states were not only connected pairwise but also formed a tightly linked local grouping. More generally, when q+1 points were all pairwise within the distance threshold, the corresponding q-dimensional simplex was added. For the purposes of this study, however, the central idea was simply that increasing ε allowed the data to reveal patterns of connectivity, clustering, and enclosure at progressively broader scales.

A helpful way to visualize this is through a three-panel diagram. In the first panel, points are isolated at a very small threshold. In the second, nearby points are joined by edges and some small triangular regions appear. In the third, larger connected structures and enclosed loop-like patterns become visible. This would make it easier for the reader to see how the shape of the trajectory changes as the distance threshold expands.

Because no single value of ε can fully describe the geometry of the trajectory, the construction was repeated over many increasing thresholds. This produced a nested sequence of representations:

$$V\,R(P, \varepsilon_1) \subseteq V\,R(P, \varepsilon_2) \subseteq V\,R(P, \varepsilon_3) \subseteq \ldots V\,R(P, \varepsilon_N)$$

This nested sequence allowed the analysis to track which structural patterns appeared only briefly and which remained visible across multiple scales. That distinction was central to the logic of persistent homology.

4.3.3 Topological Features and Persistence

When the path was mapped at growing distances, attention shifted to spotting real structural traits. To do this persistent homology came into play. It marked the moment a trait showed up in the shifting shape model - and when it vanished as ε grew larger. What mattered most wasn't just noticing the trait rather seeing how stably it lasted through changing levels. What made this significant was how macroeconomic paths sometimes show minor hiccups without indicating real shifts underneath. Treating each tiny shift as important could lead to mistaking randomness for pattern. Here, persistent homology helped by focusing on how long shapes lasted through different scales. Brief appearances tended to be seen with skepticism. Longer-lasting forms across varying levels hinted at deeper, underlying geometry in the numbers.

This work focused on three particular types of characteristics. One involved connected components - these showed whether the path spread out or clustered together. When ε started near zero, every point existed separately. With a higher threshold, points close to one another began forming linked sets. Loop-like forms came next, appearing as paths circled an area without instantly covering it. Economically speaking, such patterns might mirror cycles of return to comparable conditions, shifting again and again along fixed routes. After those appeared more complex voids existed in multiple dimensions at once. While mathematically allowed, their presence in limited economic data tended to waver, leading analysts to approach them warily.

One scale mattered per feature: when it first showed up, called B, meaning the lowest threshold of appearance. Then there was D, marking disappearance - this value recorded when the feature vanished. Persistence came down to that difference, nothing more

$$\pi = d - b$$

What sticks around at multiple levels matters most: traits showing high persistence hardly fade when scale shifts, making them stronger candidates for reflecting real shape patterns in the dataset.

One way to picture these findings is through a persistence diagram. Each dot on the plot stands for a single observed pattern. The appearing times show up along the x-axis. The y-axis marks when those patterns fade out. Near-diagonal points tend to vanish quickly. Short duration usually suggests minor importance. Such cases might just capture random bumps in data structure. Structures located further

from the diagonal tend to appear consistently over many scale levels - so they're often seen as more significant. Because of this behavior, the persistence diagram becomes a powerful tool; it reveals both the types of patterns present along the path and their degree of stability.

This notion is tied closely to how the findings were understood. Rather than merely mapping the shape of the data, persistent homology aimed to separate fleeting shifts from meaningful patterns. Although major loop structures rarely stood out, persistence still mattered - it offered a clear method for spotting enduring changes in movement. What made this approach different from conventional regression cutoffs or classifiers was its ability to detect early signs - whether the path of growth began showing consistent narrowing, tightening, or shift prior to hitting a stagnation point. Persistence diagrams summarise the birth and death of geometric features across distance scales, allowing the analysis to separate short-lived noise from more stable structural signals.

Stability across multiple scales marks what we call feature persistence; those showing high persistence tend to reflect true data geometry. Though scale varies, certain patterns hold firm - these enduring ones stand out as meaningful shape elements.

4.4 Econometric Baseline

A country's position in the coming three years - inside or outside the defined trap condition - formed the core of the analysis. Whether it crossed into that state depended on macroeconomic signals shared across nations. Instead of treating all data points alike, splits emerged based on income levels, drawing a line between different economic realities. Values where this split worked best came from testing various breakpoints along that income scale. Performance shaped the final choice - not just raw correctness, but how well rare cases appeared and false alarms stayed low. Detection strength guided selection, favoring setups where misses dropped without flooding results with noise.

This baseline did not aim at building a structural causal model for sustained economic expansion. Instead, it served as a clear, nonlinear benchmark for evaluating TDA outputs. Understanding arose through regime-specific patterns: alert dynamics made sense once linked to a computed tipping point, along with parameter values active above or below it.

4.5 Random Forest Baseline

Not built on a single rule, the machine-learning approach used a Random Forest model trained on identical pooled observations, then tested over the same three-year horizon. Performance gaps traceable to method design emerged clearly because predictor variables matched those in traditional econometrics. Instead of relying on one path, this technique combines outputs from numerous trees, capturing complex patterns without rigid assumptions. Its strength came through adaptability - finding jumps, breaks, and variable blends directly in observed behavior. Even though favorable testing periods occurred infrequently, adjustments for uneven class distribution shaped how models were trained and thresholds were set. As a result, the Random Forest emerged as the central point of comparison across predictions.

What mattered most was its adaptability in sorting outcomes - less so revealing internal mechanics - offering a clear counterpoint to traditional economic models alongside topological data methods.

4.6 Detection and Comparison Criteria

At two stages, models were compared. Evaluation began with window-level results, measured through accuracy along with precision, recall, and related metrics, applied to a tightly matched set of cases. Following that, attention shifted to national outcomes, focusing on how early warnings appeared before the reference point in time. This gap defined the lead-time highlighted throughout the study. Only countries showing valid pre-event activity within the stringent sample were entered into this later analysis. The presence of usable earlier signals determined inclusion. Although prediction accuracy mattered, the analysis also weighed how clear each method's reasoning appeared. When examining TDA, clarity came from observing shifts in the path's shape over time. The econometric approach made sense through its shifting states, along with how thresholds and coefficients changed. As for Random Forest, understanding stayed narrow - mostly tied to which inputs influenced outcomes most and how predictions clustered across cases.

EXPERIMENTS

5.1 Results

5.1.1 TDA Results

Paper Table 1 - Country-level TDA Detection Results

country	benchmark_trap_onset_year	Eglible_for_2d_lead_time_evaluation	tda_lead_time_years
Bangladesh	1990	FALSE	
Cambodia		TRUE	
Ethiopia	1990	FALSE	
Ghana	1991	FALSE	
Malawi	1994	TRUE	
Nepal		FALSE	
Niger	1990	FALSE	
Pakistan		FALSE	
Rwanda		TRUE	
Senegal	1990	FALSE	
Tanzania		FALSE	
Zambia	2007	TRUE	7

Paper Table 1 - Country-level TDA Detection Results

country	tda_lead_time_years	detected_before_or_on_onset	ever_flagged_by_tda_2d
Bangladesh			FALSE
Cambodia	1995		TRUE
Ethiopia			FALSE
Ghana			FALSE
Malawi			FALSE
Nepal	2001		TRUE
Niger			FALSE
Pakistan			FALSE
Rwanda			FALSE
Senegal			FALSE
Tanzania			FALSE
Zambia	2000	TRUE	TRUE

Figure 2 - TDA detection Results

This table reports the final country-level TDA detection results. This includes which countries were eligible for lead-time evaluation and the first detected warning year. It establishes that the main evaluable TDA lead-time result in the present sample is Zambia, where the TDA signal gives a seven-year lead before benchmark onset.

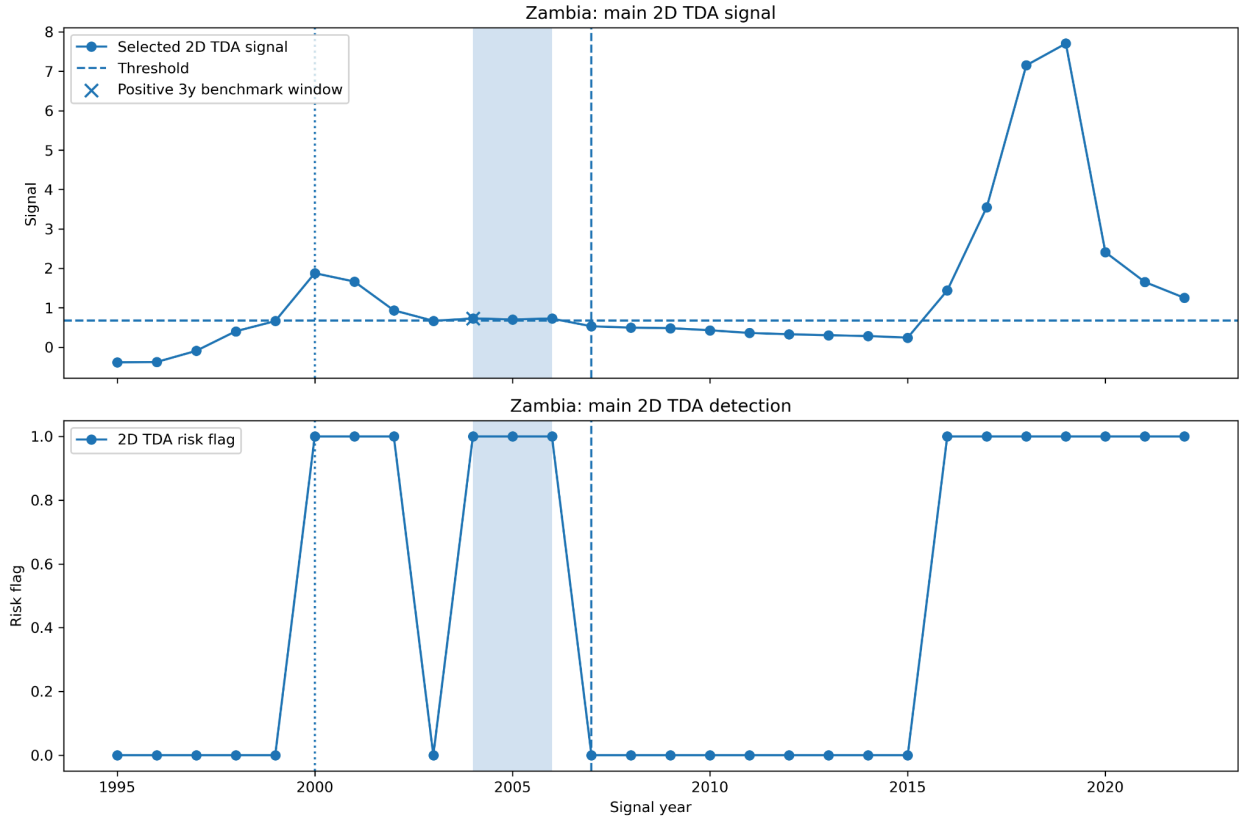


Figure 3 - TDA Signals For Zambia Over Time

This figure shows the final selected TDA signal over time for Zambia together with the decision threshold and the benchmark onset reference. It implies that the TDA signal rises early enough to generate a warning before the benchmark trap onset, giving a clear case of early structural detection.

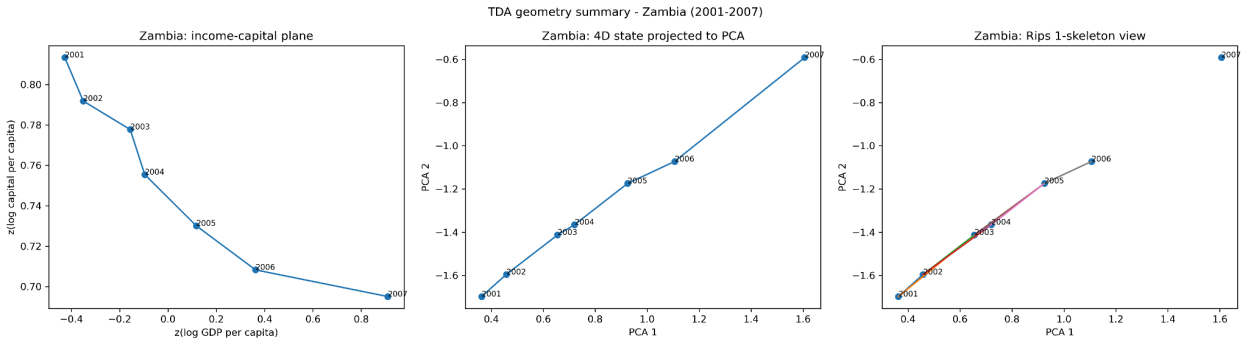


Figure 4 - Zambia State Space Trajectory

This figure visualises Zambia's state-space trajectory, its projected geometry, and the local connected structure over the years leading to the onset. It implies that the relevant TDA signal is tied to a narrowing and more constrained geometric path rather than to complex looping behavior.

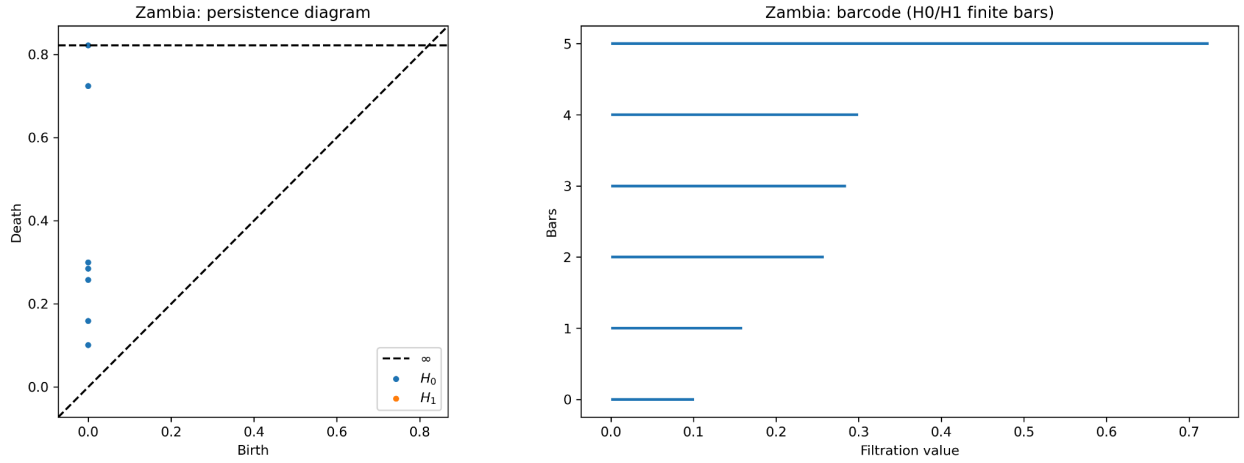


Figure 5 - Persistence Diagram for Zambia

This figure shows the persistence diagram and barcode for Zambia's focal window, summarising the topological structure of the point cloud. It implies that the dominant TDA information in this case comes from connectedness and structural compression, with little evidence of persistent loop-like structure.

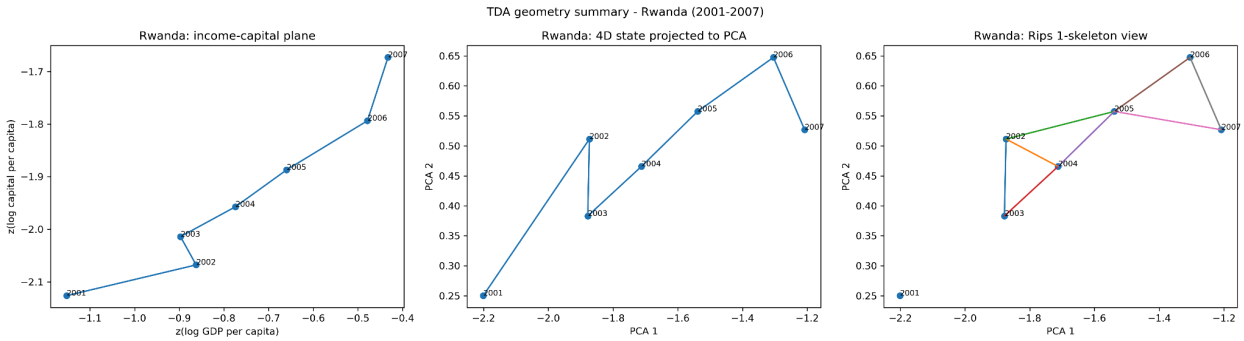


Figure 6 - Rwanda Geometry Summary

This figure provides a comparison case for a country without a benchmark onset in the current setup. It implies that the geometry can remain structurally different from the Zambia onset case, helping distinguish a detected transition pattern from a non-onset pattern.

5.1.2 Econometric Model Results

Table E1 — Econometric Baseline Performance and Detection Results

country	benchmark_trap_onset_year	eligible_for_econ_lead_time_evaluation	first_econ_detection_year	econ_lead_time_years	detected_before_or_on_onset
Bangladesh	1990	FALSE			
Cambodia		FALSE			
Ethiopia	1990	FALSE			
Ghana	1991	FALSE			
Malawi	1994	TRUE	1990	4.0	TRUE
Nepal		FALSE			
Niger	1990	FALSE			
Pakistan		FALSE			
Rwanda		FALSE	1990		
Senegal	1990	FALSE			
Tanzania		FALSE			
Zambia	2007	TRUE	2004	3.0	TRUE

country	ever_flagged_by_econ	strict_sample_rows	positive_strict_rows	max_predicted_probability	mean_predicted_probability
Bangladesh	FALSE	0	0		
Cambodia	FALSE	15	0	0.0079	0.0046
Ethiopia	FALSE	0	0		
Ghana	FALSE	0	0		
Malawi	TRUE	15	3	0.1736	0.1167
Nepal	FALSE	15	0	0.0050	0.0019
Niger	FALSE	0	0		
Pakistan	FALSE	15	0	0.0000	0.0000
Rwanda	TRUE	15	0	0.1371	0.0768
Senegal	FALSE	0	0		
Tanzania	FALSE	0	0		
Zambia	TRUE	15	1	1.0000	0.0667

Figure 7 - Econometric Baseline Performance and Detection Results

This table reports the main fitted-sample performance and warning-timing results of the threshold-style econometric baseline on the common evaluable dataset. It shows that the econometric model detects the eligible onset cases, with a 4-year lead for Malawi and a 3-year lead for Zambia.

Table E2 — Top Econometric Calibration Candidates

Low probability cutoffs maximize recall on the sparse positive sample.

model_type	threshold_value	cutoff	row_count	predicted_positive_rate	actual_positive_rate	tp	fp	tn
threshold_glm	7.3684	0.0500	90	0.3111	0.0444	4	24	62
threshold_glm	7.3930	0.0500	90	0.3111	0.0444	4	24	62
threshold_glm	7.4660	0.0500	90	0.3111	0.0444	4	24	62
threshold_glm	7.1679	0.0500	90	0.3556	0.0444	4	28	58
threshold_glm	7.3477	0.0500	90	0.3556	0.0444	4	28	58
threshold_glm	7.1977	0.0500	90	0.3667	0.0444	4	29	57
threshold_glm	7.2360	0.0500	90	0.3778	0.0444	4	30	56
threshold_glm	7.3025	0.0500	90	0.3778	0.0444	4	30	56
threshold_glm	7.1022	0.0500	90	0.2667	0.0444	3	21	65
threshold_glm	6.9992	0.1000	90	0.0444	0.0444	2	2	84

model_type	fn	accuracy	precision	recall	specificity	balanced_accuracy	f1	low_regime_rows	high_regime_rows
threshold_glm	0	0.7333	0.1429	1.0000	0.7209	0.8605	0.2500	63.0000	27.0000
threshold_glm	0	0.7333	0.1429	1.0000	0.7209	0.8605	0.2500	67.0000	23.0000
threshold_glm	0	0.7333	0.1429	1.0000	0.7209	0.8605	0.2500	72.0000	18.0000
threshold_glm	0	0.6889	0.1250	1.0000	0.6744	0.8372	0.2222	41.0000	49.0000
threshold_glm	0	0.6889	0.1250	1.0000	0.6744	0.8372	0.2222	58.0000	32.0000
threshold_glm	0	0.6778	0.1212	1.0000	0.6628	0.8314	0.2162	45.0000	45.0000
threshold_glm	0	0.6667	0.1176	1.0000	0.6512	0.8256	0.2105	49.0000	41.0000
threshold_glm	0	0.6667	0.1176	1.0000	0.6512	0.8256	0.2105	54.0000	36.0000
threshold_glm	1	0.7556	0.1250	0.7500	0.7558	0.7529	0.2143	27.0000	63.0000
threshold_glm	2	0.9556	0.5000	0.5000	0.9767	0.7384	0.5000	18.0000	72.0000

Figure 8 - Top Econometric Calibration Candidates

This table lists the strongest threshold and cutoff combinations considered during calibration of the econometric baseline. It shows that the best-performing specifications achieved very high recall, although this came with low precision because positive benchmark windows are rare in the sample.

Econometric Threshold Baseline Metrics

metric_set	row_count	model_type	threshold_value	cutoff	tp	fp	tn	fn
window_level_strict_sample	90	threshold_glm	7.3684	0.0500	4.0000	24.0000	62.0000	0.0000
country_level_detection	12	threshold_glm	7.3684	0.0500				

metric_set	accuracy	precision	recall	specificity	balanced_accuracy	f1	mean_lead_time_years	median_lead_time_years	eligible_onset_country_count	detected_eligible_onset_country_count
window_level_strict_sample	0.7333	0.1429	1.0000	0.7209	0.8605	0.2500				
country_level_detection							3.5	3.5	2.0000	2.0000

Figure 9 - Threshold Baseline Metrics

This table summarizes the overall classification performance and country-level detection outcomes of the econometric baseline. It implies that the model is effective at capturing the observed positive cases in the evaluable sample, but it does so with a relatively high false-positive rate.

5.1.3 Machine Learning Model Results

Table M1 — Random Forest Baseline Performance and Detection Results

Use with caution: RF metrics are fitted-sample apparent fit, not strict out-of-sample performance.

country	benchmark_trap_onset_year	eligible_for_rf_lead_time_evaluation	first_rf_detection_year	rf_lead_time_years	detected_before_or_on_onset	ever_flagged_by_rf
Bangladesh	1990	FALSE				FALSE
Cambodia		FALSE				FALSE
Ethiopia	1990	FALSE				FALSE
Ghana	1991	FALSE				FALSE
Malawi	1994	TRUE	1991	3.0	TRUE	TRUE
Nepal		FALSE				FALSE
Niger	1990	FALSE				FALSE
Pakistan		FALSE				FALSE
Rwanda		FALSE				FALSE
Senegal	1990	FALSE				FALSE
Tanzania		FALSE				FALSE
Zambia	2007	TRUE	2004	3.0	TRUE	TRUE

country	strict_sample_rows	positive_strict_rows	max_predicted_probability	mean_predicted_probability
Bangladesh	0	0		
Cambodia	15	0	0.1503	0.0235
Ethiopia	0	0		
Ghana	0	0		
Malawi	15	3	0.7131	0.2110
Nepal	15	0	0.0260	0.0072
Niger	0	0		
Pakistan	15	0	0.1157	0.0209
Rwanda	15	0	0.2194	0.0608
Senegal	0	0		
Tanzania	0	0		
Zambia	15	1	0.6046	0.0934

Figure 10 - Random Forest Baseline Performance

This table reports the fitted-sample classification and country-level warning results of the Random Forest baseline on the same common dataset. It shows that the model detects both eligible onset countries, giving a 3-year lead for both Malawi and Zambia.

Table M2 — Random Forest Feature Importance

feature	importance
log_capital_pc	0.5326
log_gdp_pc	0.4674

Figure 11 - Random Forest Feature Importance

This table reports the relative importance of the two common state predictors used in the Random Forest model. It suggests that capital per capita is slightly more influential than income per capita in the fitted machine learning baseline.

Random Forest Metrics

metric_set	row_count	cutoff	tp	fp	tn	fn	accuracy	precision	recall	specificity
window_level_strict_sample	90	0.3500	4.0000	0.0000	86.0000	0.0000	1.0000	1.0000	1.0000	1.0000
country_level_detection	12	0.3500								

metric_set	balanced_accuracy	f1	mean_lead_time_years	median_lead_time_years	eligible_onset_country_count	detected_eligible_onset_country_count
window_level_strict_sample	1.0000	1.0000				
country_level_detection			3.0	3.0	2.0000	2.0000

Figure 12 - Random Forest Metrics

This table summarizes the fitted sample predictive performance of the Random Forest baseline and its country level detection coverage. It indicates extremely strong apparent fit on the evaluable sample, though these results should be interpreted as in sample performance rather than strict out of sample generalization.

5.2 Analysis

The TDA results provided the earliest structurally meaningful warning in the sample. At the country level, Zambia was the only case that satisfied the final eligibility rules for two-dimensional lead-time evaluation, and the selected signal first crossed the detection threshold in 2000, seven years before the benchmark onset year of 2007. Later in the matched setup, TDA emerged first in signaling the lone valid 2D instance of onset. This finding reflects a single-case indication, not wide confirmation of better performance overall. The geometric interpretation also mattered. The Zambia exhibits indicated that the dominant signal was associated less with persistent loop formation than with narrowing and compression in the state-space trajectory, which supported the interpretation of increasing structural confinement prior to onset.

The threshold-style econometric baseline served as the most transparent nonlinear reference. It detected the eligible onset cases with shorter lead times than TDA, namely four years for Malawi and three years for Zambia. Its calibration results showed strong recall but comparatively low precision, which reflected the rarity of positive benchmark windows in the evaluation sample. Even so, the econometric model remained valuable because its regime-based structure provided a clear and interpretable account of how warning behaviour changed across threshold-defined states.

The Random Forest baseline produced the strongest fitted-sample predictive performance on the strict common-state sample. It correctly detected both Malawi and Zambia with three-year lead times and achieved near-perfect fitted classification at several cutoff values. Those results should nevertheless be interpreted cautiously. The evaluation sample was small and highly imbalanced, so the Random Forest results are better understood as evidence of strong in-sample fit than as proof of robust out-of-sample generalization. Substantively, the feature-importance results suggested that capital per capita was slightly more influential than income per capita, but the model still offered less structural transparency than the TDA or econometric approaches.

Taken together, the three methods did not produce a single winner on all criteria. Random Forest performed best as a fitted classifier, and the econometric model provided the clearest regime-based interpretation. TDA, however, contributed something distinct: it delivered the earliest evaluable warning in the final sample and linked that warning to observable deformation in the geometry of the development

path. This is the central comparative result of the paper. Persistent homology did not simply reproduce the conclusions of the benchmark models; it added a structurally interpretable perspective on how poverty-trap dynamics began to emerge.

One major issue with this analysis lies in the narrow scope of the evaluation group. Just two nations - Malawi and Zambia - meet the criteria for initial comparison under the shared framework, while only Zambia qualifies as a valid instance for assessing the full temporal dimension of the TDA method. As such, the conclusion favoring TDA's early detection ability reflects what emerged most clearly in the best available example, not proof that topological techniques universally surpass traditional models. Caution applies equally to the Random Forest outcomes, given these are based on a tiny, uneven set of observations where high accuracy might stem from model-specific quirks instead of real predictive strength. Looking ahead, expanding the study to include more developing economies could help verify patterns, probe different definitions of economic traps, assess forecasting beyond observed data, and check if similar TDA signals arise during varied periods of stalled progress.

Table C1 — Final Cross-Method Comparison

Main comparison table for the evaluation/comparison section.

Method	Main signal/model	Window-level balanced accuracy	Precision	Recall	Eligible onset countries	Detected eligible onset countries
TDA	z_pca2_to_pca1_range_ratio	0.9322	0.1111	1.0000	1	1
Econometric	Threshold GLM	0.8605	0.1429	1.0000	2	2
Random Forest	RF classifier	1.0000	1.0000	1.0000	2	2

Method	Mean lead time	Zambia lead time	Interpretability
TDA	7.0000	7.0000	High structural interpretability
Econometric	3.5000	3.0000	High parametric interpretability
Random Forest	3.0000	3.0000	Lower structural interpretability

Figure 13 - Cross Method Comparsion

Presents the final cross-method comparison between the TDA model, the threshold-based econometric model, and the Random Forest baseline under the common benchmark framework.

5.3 Conclusion

This study examined whether persistent homology could provide earlier and more structurally interpretable detection of poverty-trap dynamics than threshold-based econometric and machine-learning approaches. To do so, it placed three distinct epistemological frameworks within a common empirical design defined by a benchmark trap-onset rule, harmonized macroeconomic inputs, and comparable warning-horizon evaluation. The resulting comparison allowed the analysis to assess not only predictive performance, but also what each method revealed about the structure of stagnation.

The results showed that the three approaches contributed different strengths. The Random Forest baseline achieved the strongest fitted-sample predictive performance on the strict common-state sample, while the threshold-style econometric model provided a transparent nonlinear reference with interpretable regime-based warnings. The TDA pipeline, however, generated the earliest structurally meaningful warning in the final evaluable case. For Zambia, the selected TDA signal flagged risk in 2000, seven years before the benchmark onset in 2007. Importantly, that signal was associated not with dominant persistent loops but with geometric compression in the reconstructed state space, suggesting that the development trajectory became increasingly narrow and constrained prior to onset.

These findings indicated that persistent homology added value not captured in the same way by the benchmark approaches. Its main contribution was not a superior fitted classification, but an earlier warning tied to an interpretable structural change in the trajectory itself. At the same time the evidence should be read within the limits of the final sample, the narrow set of eligible onset cases, and the dependence of the final comparable design on a reduced common-state specification. How outcomes shift hinges partly on how methods are set up - something future studies ought to probe with stricter consistency. For instance, the picked time lag τ, number of dimensions m, span of moving windows, alongside reference thresholds for trapping moments, each play a role in when and how strongly structural shifts appear. Even though the last balanced setup leaned on a simplified shared condition to keep comparisons fair, broader investigations need evaluations under differing frameworks: alternate state reconstructions, window lengths, collections of indicators, and ways of defining traps. Only then can we tell if Zambia's pattern stands as a steady trait in its growth path or merely an artifact shaped by specific modeling decisions. Even with those limitations, the study showed that persistent homology can serve as a credible complement to econometric and machine-learning methods when the aim is to detect and interpret emerging poverty-trap dynamics.

REFERENCES

1. Azariadis, C. and Stachurski, J. (2005) 'Poverty Traps', in Handbook of Economic Growth, Vol. 1A, pp. 295–384.
2. Kraay, A. and McKenzie, D. (2014) 'Do Poverty Traps Exist? Assessing the Evidence', Journal of Economic Perspectives, 28(3), pp. 127–148.
3. Hansen, B.E. (2000) 'Sample Splitting and Threshold Estimation', Econometrica, 68(3), pp. 575–603.
4. Carlsson, G. (2009) 'Topology and Data', Bulletin of the American Mathematical Society, 46(2), pp. 255–308.
5. Chazal, F. and Michel, B. (2021) 'An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists', Frontiers in Artificial Intelligence, 4, 667963.
6. Otter, N., Porter, M.A., Tillmann, U., Grindrod, P. and Harrington, H.A. (2017) 'A Roadmap for the Computation of Persistent Homology', EPJ Data Science, 6, 17.
7. Athey, S. and Imbens, G.W. (2019) 'Machine Learning Methods That Economists Should Know About', Annual Review of Economics, 11, pp. 685–725

8. Breiman, L. (2001) 'Random Forests', Machine Learning, 45(1), pp. 5–32
9. Scheffer, M. et al. (2009) 'Early-warning signals for critical transitions', Nature, 461, pp. 53–59.
10. Takens, F. (1981) 'Detecting Strange Attractors in Turbulence', in Dynamical Systems and Turbulence, Warwick 1980. Lecture Notes in Mathematics, Vol. 898, pp. 366–381.
11. Caner, M. and Hansen, B.E. (2004) 'Instrumental Variable Estimation of a Threshold Model', Econometric Theory, 20(5), pp. 813–843.
12. Gidea, M. and Katz, Y. (2018) 'Topological Data Analysis of Financial Time Series: Landscapes of Crashes', Physica A: Statistical Mechanics and its Applications, 491, pp. 820–834