Comparative Analysis of Metrics for Model Drift

Vignesh Narayanaswamy

2025-07-22

Introduction

Objective This analysis compares Hellinger Distance and PSI for monitoring input drift and output drift, contrasting the metrics for effectiveness across distinct scenarios of distributional change to discern the strengths and weaknesses of each approach. The objective of this document is to help model owners make informed decisions regarding metric choice in alignment with the specific needs of their monitoring tasks.

Scope To be able to effectively monitor model or heuristic performance, a fundamental challenge is the quantification of performance degradation. Identifying the parameters to track the model performance and defining the thresholds that if breached should raise an alert are fundamental components of continuous monitoring. Drift, defined as the change in a distribution over time, can indicate the degradation of a machine learning model or heuristic.

There are two main types of drift: Data Drift - Change in distribution of X (features) or Y (predictions) E.g. Distribution of incomes of customers changes over time Concept Drift - Drift in P(X|Y) called concept drift. E.g. Definition of what is considered a fraudulent transaction changes

This document only discusses the application of Hellinger Distance and PSI to data drift and does not explore their usage for concept drift.

II. Overview of Drift Metrics

A. Hellinger Distance

Definition

The Hellinger distance $H(P, Q)$ between two discrete probability distributions P and Q is defined as:

$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2}$$

where p_i is the probability of event i in distribution P and q_i is the probability of event i in distribution Q.

Properties

- The Hellinger distance is always in the range $[0, 1]$. A value of 0 indicates that the distributions are identical, while a value of 1 indicates that the distributions are maximally different.
- It is symmetric, meaning $H(P, Q) = H(Q, P)$.

B. Population Stability Index

Definition

The PSI is calculated as:

$$\text{PSI} = \sum \left((p_i - q_i) \times \ln \left(\frac{p_i}{q_i} \right) \right)$$

where: p_i is the proportion of events in a specific bin for the development sample and q_i is the proportion of events in the same bin for the validation sample.

Properties

- A PSI value of 0 indicates no shift in the distribution between the development and validation samples. There is no upper limit to PSI. As the distributions diverge, the value of PSI continues to rise.

Assumptions and Limitations of both approaches

- **Binning**: If used for numeric variables, binning must be performed, which can sometimes be subjective. The choice of bins can impact the PSI value.
- **Size of Samples**: Small sample sizes can lead to instability in the calculations, especially if certain bins have very few observations.

C. Comparison of Hellinger Distance and PSI for distributional shifts

The Population Stability Index (PSI) and Hellinger Distance (HD) measure distributional shifts differently through their respective formulas.

1. **Population Stability Index (PSI):**

$$\text{PSI} = \sum_{i=1}^{n} \left((p_i - q_i) \cdot \ln \left(\frac{p_i}{q_i} \right) \right)$$

   - **Difference in Proportions** $(p_i - q_i)$: This term captures the absolute difference in proportions between the actual and expected distributions for each bin. If the distributions are identical, the difference in proportions is zero, resulting in a PSI of zero.
   - **Logarithm of the Ratio** $\left(\ln \left(\frac{p_i}{q_i}\right)\right)$: The log-ratio amplifies the relative change between the two distributions. It is more sensitive to relative changes, especially when the proportions are small, making the PSI very responsive to distributional shifts.

2. **Hellinger Distance (HD):**

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{n} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2}$$

   - **Square Root of Proportions** $(\sqrt{p_i}$ and $\sqrt{q_i})$: The square root transformation can help dampen the effect of extreme values or outliers. It is less sensitive to large absolute differences in proportions, especially in the tails of the distributions.
   - **Squared Euclidean Distance:** This part measures the "distance" between the two distributions across all bins, emphasizing the geometric difference between the distributions. It is sensitive to differences across the entire range of the distributions, making the Hellinger Distance a balanced measure of distributional shift.

Comparison:

- **Sensitivity:** PSI is often more sensitive to distributional shifts due to the log-ratio, especially when the proportions are small or when there's a significant change in a bin with a small proportion of data. On the other hand, the Hellinger Distance may be less sensitive to such changes due to the square root transformation.

- **Outliers and Noise:** Hellinger Distance's square root transformation can dampen the effect of extreme values or noise, making it potentially more robust to such values compared to PSI.
- **Interpretability:** PSI may be less intuitive due to its unbounded nature and reliance on the logarithm of ratios, whereas Hellinger Distance's bounded range (0 to 1) and geometric interpretation can sometimes be more intuitive.

III. Simulation for Comparing Distribution Metrics

This section compares Hellinger Distance and PSI by simulating different scenarios and evaluating their differences.

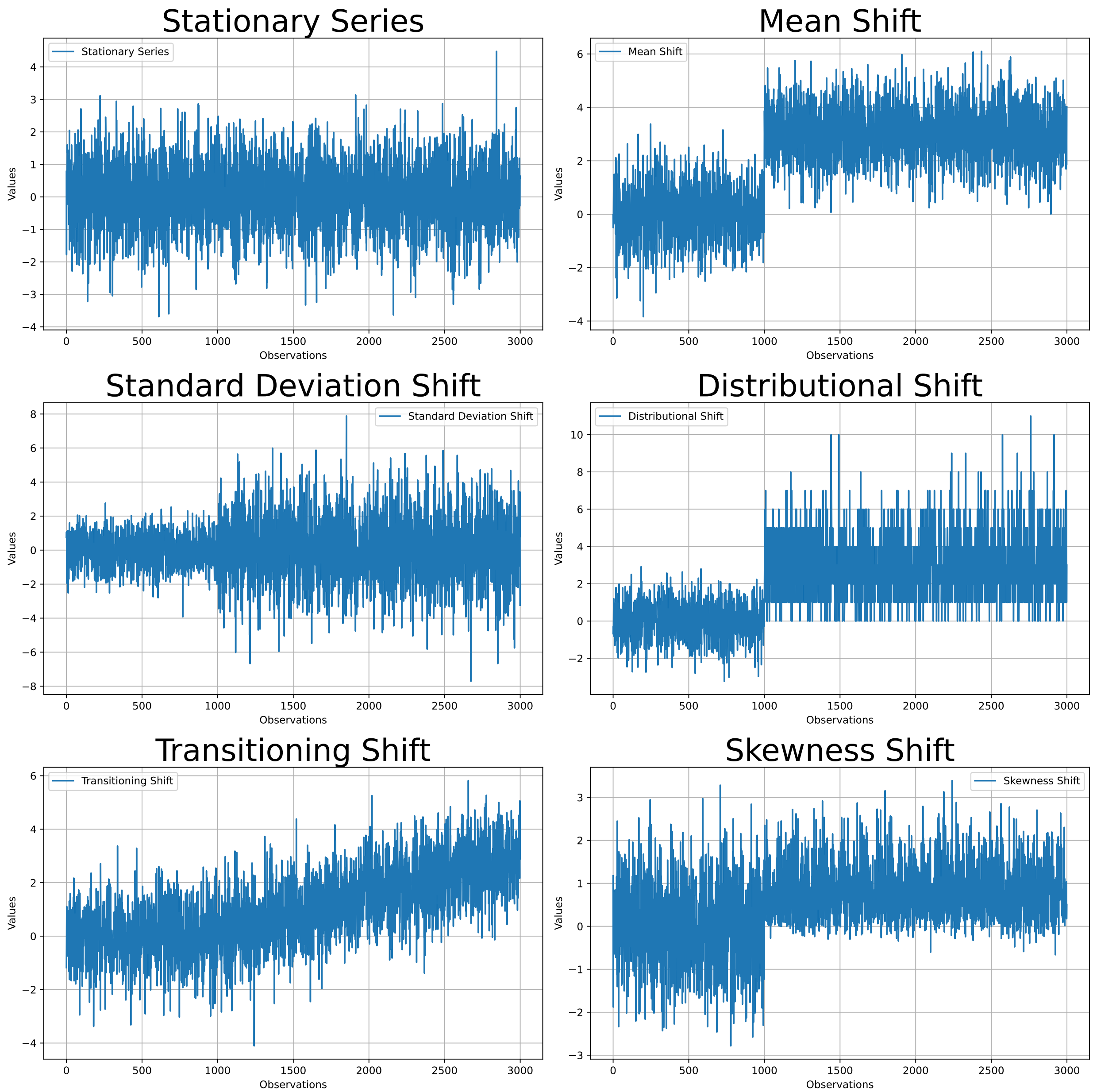
Distributional Shifts Explored

The following shifts are explored in a fictitious scenario, where 5 variables are sampled from standard normal distributions, each with 1000 observations (the training data). Subsequently, an additional 3000 observations are generated (the live data). 1000 observations through the live data, a distributional shift occurs in 4 of the 5 variables. A summary of the 5 variables is below:

- **Mean Shift**:
  - Mean shifts from 0 to 3.
- **Standard Deviation Shift**:
  - Standard deviation changes from 1 to 2.
- **Distributional Shift**:
  - The data transitions to samples from a Poisson distribution with a mean of 3.
- **Transitioning Series**:
  - The series transitions from an initial mean of 0 to a final mean of 3 over 1000 time steps.
- **Stationary Series**:
  - Remains consistent across the training and live data and does not exhibit any drift.

Graphical representations of the shifts are below:

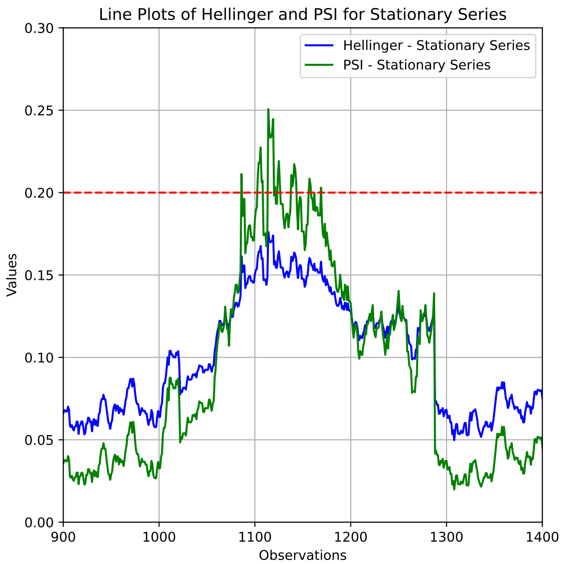
Visualization of Shifts

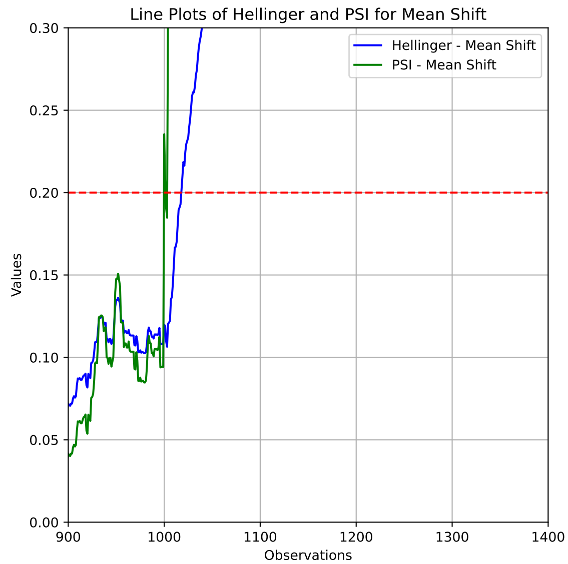


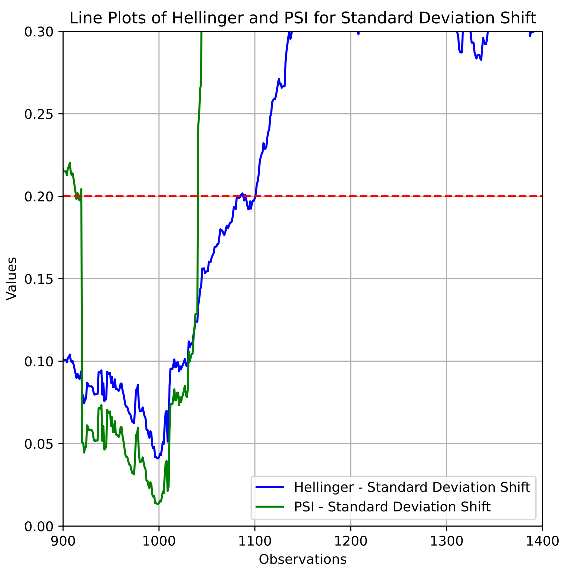
Next, metrics are computed for each variable, using the training data for the reference distribution and using a sliding window over the live data for the comparision distribution. A window size is defined, meaning each segment for analysis includes 150 observations, and a step size of 1 is employed, meaning the window moves forward by one observation for each subsequent analysis.

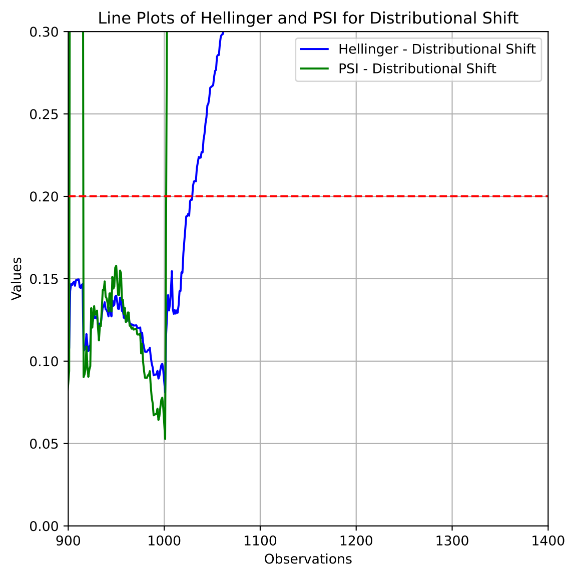
Visual Comparison of Drift Metrics

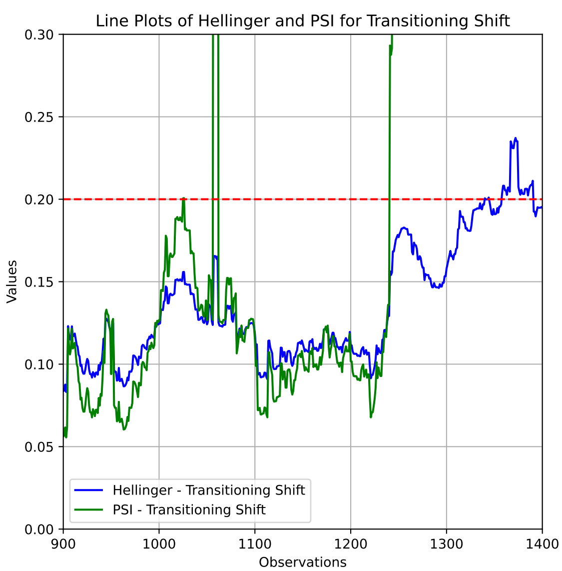
Metrics for the previously described distributions are plotted below. PSI and Hellinger distance respond to all distributional shifts examined. However, PSI tends to be more reactive, detecting changes faster. While this is advantageous in scenarios where this is true drift presence, PSI may be overly reactive to noise, or where there is a small number of observations in some bins. Note that PSI has false positives for drift for the Stationary Series, where there is no drifts.

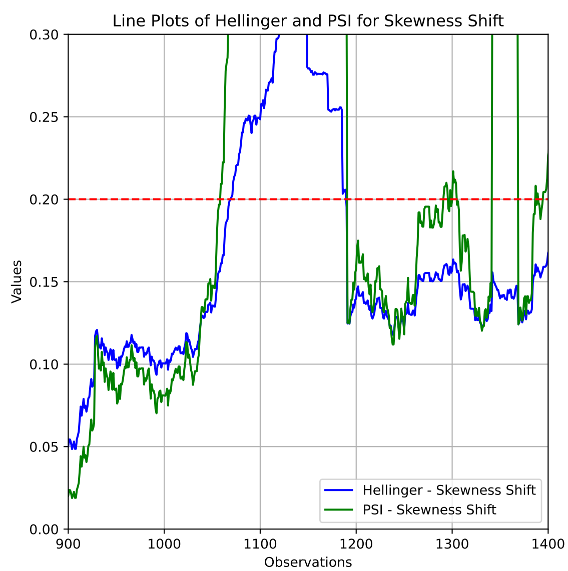

Line Plots of Hellinger and PSI for Stationary Series
Hellinger - Stationary Series
PSI - Stationary Series
Values
Observations
0.00
0.05
0.10
0.15
0.20
0.25
0.30
900
1000
1100
1200
1300
1400


Line Plots of Hellinger and PSI for Mean Shift
Hellinger - Mean Shift
PSI - Mean Shift
Values
Observations
0.00
0.05
0.10
0.15
0.20
0.25
0.30
900
1000
1100
1200
1300
1400


Line Plots of Hellinger and PSI for Standard Deviation Shift
Hellinger - Standard Deviation Shift
PSI - Standard Deviation Shift
Values
Observations
0.00
0.05
0.10
0.15
0.20
0.25
0.30
900
1000
1100
1200
1300
1400

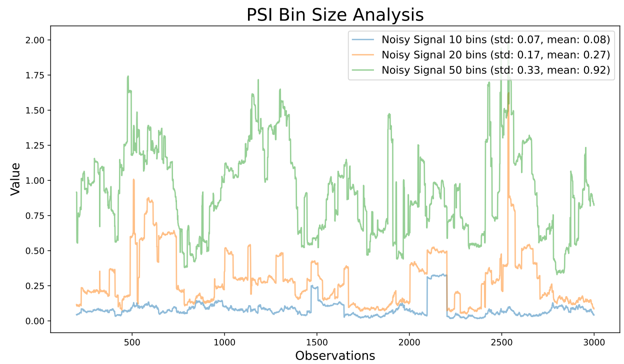

Line Plots of Hellinger and PSI for Distributional Shift
Hellinger - Distributional Shift
PSI - Distributional Shift
Values
Observations
0.00
0.05
0.10
0.15
0.20
0.25
0.30
900
1000
1100
1200
1300
1400

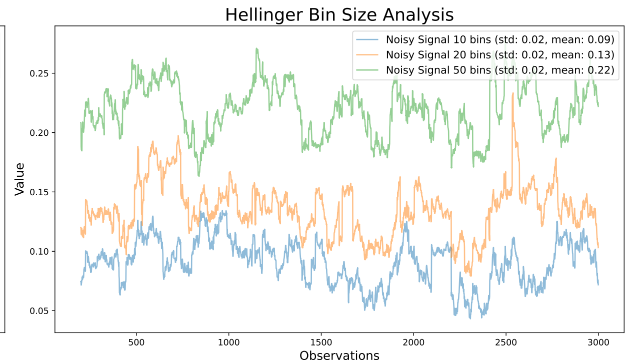

Line Plots of Hellinger and PSI for Transitioning Shift
Hellinger - Transitioning Shift
PSI - Transitioning Shift
Values
Observations
0.00
0.05
0.10
0.15
0.20
0.25
0.30
900
1000
1100
1200
1300
1400


Line Plots of Hellinger and PSI for Skewness Shift
Hellinger - Skewness Shift
PSI - Skewness Shift
Values
Observations
0.00
0.05
0.10
0.15
0.20
0.25
0.30
900
1000
1100
1200
1300
1400

Effect of Bin Number

To further demonstrate the effect of noise and bin number on PSI and Hellinger Distance, consider the following example which uses different bin numbers and a variable with no distributional shift, but is noisier than the previous distributions observed (mean zero with standard deviation of 10). Note that while both the average value of PSI and Hellinger respond monotonically to bin number (more bins leads to larger metric values), the standard deviation of PSI increases as the number of bins goes up while the standard deviation of Hellinger Distance stays constant.





Choosing between PSI and Hellinger Distance to Monitor Drift

Property/Scenario	Population Stability Index (PSI)	Hellinger Distance (HD)
General Sensitivity	High sensitivity to distributional shifts due to the log-ratio, especially when the proportions are small.	Less sensitive to large absolute differences in proportions, especially in the tails of the distributions due to square root transformation.
Outliers and Noise	May be more affected by outliers and noise.	The square root transformation can help dampen the effect of extreme values or outliers, making HD potentially more robust to such values compared to PSI.
Interpretability	May be less intuitive due to its unbounded nature and reliance on the logarithm of ratios.	More intuitive due to bounded range (0 to 1) and geometric interpretation.
Binning Requirement	Binning must be performed, which can be subjective. The choice of bins can impact the PSI value.	Binning must be performed, which can be subjective. The choice of bins can impact the HD value.
Size of Samples	Small sample sizes can lead to instability in the calculations, especially if certain bins have very few observations.	Small sample sizes can lead to instability in the calculations, especially if certain bins have very few observations.