**FX Anomaly Model – Estimating Required Samples for New BU×CP Pairs**

This document explains how we estimate the **minimum number of examples required** before retraining the anomaly detection model on new **Business Unit × Counterparty (BU×CP)** pairs.

The approach combines **statistical confidence-based estimation** with drift-adjustment factors (**Variance Ratio, PSI, VIF**) while ensuring comparisons are made against the **right reference cohort** of historical data. Example used throughout:  
**new pair = HSBC × ANZ**, recently observed with ~12 deals.

**1. Why this matters**

* Our anomaly detection model is an **autoencoder**, trained to reproduce historical FX deals.
* When new BU×CP pairs appear, the model may not have seen them before.
* Retraining with too few examples risks poor learning and unstable anomaly detection.
* Business needs a clear rule of thumb:  
  *“How many deals of this new type are enough before retraining makes sense?”*

## *Context & Scope (Read Me First)*

## This method is for retraining only. It estimates how many examples we need to safely retrain an already-deployed autoencoder when a new Business Unit × Counterparty (BU×CP) pair starts appearing in an existing client’s data.

## Not for new client onboarding. This is not a general “how many examples to train a new model from scratch” method. New-client training requires broader data sufficiency analysis (feature coverage, categorical diversity, business process differences, class imbalance, etc.) and is out of scope here.

## Range-sensitive. The estimated requirement assumes the future examples for this new pair stay within similar value ranges and patterns as currently observed. If later data arrives with different scales or distributions, the requirement will recalculate upward (we call this “recalibration by drift”).

## No hyperparameter tuning. Retraining uses the existing model parameters and preprocessing. We are deciding when to include the new pair in scheduled retraining, not how to change the model.

## *Executive Summary*

## We estimate how “wobbly” (variable) reconstruction errors are for similar data in history (BU level and CP level).

## We compare the new pair to a matching historical cohort (same instrument & currencies where possible) to see if numeric values are more spread (Variance Ratio, VR) or shifted in shape (Population Stability Index, PSI).

## We roll these into a Variance Inflation Factor (VIF) to inflate the base volatility if the new pair looks harder than history.

## We plug the inflated volatility into a standard confidence formula to get the minimum number of examples required before retraining.

## Because this is retraining, the number reflects how much data we need to integrate this new pair reliably, not how much data is needed to train a fresh model.

## *Prerequisites & Assumptions*

* A **working autoencoder** and **preprocessing pipeline** trained on the client’s historical data.

## The same pipeline is applied consistently (no leakage).

## Historical error variances per BU and CP are representative proxies for unseen pairs.

## For PSI on categorical features, we use reference categories from history with an \_\_OTHER\_\_ bin for unseen values.

## We cap VIF to avoid runaway sample sizes (e.g., VIF ≤ ×3).

## We floor VR at 1.0 (we never reduce required n if data appears “easier”).

**2. Analytical estimation – proactive sample size**

**Step 1: Baseline wobble from history**

**Function:** baseline\_stats(...)

* Run the model on historical data, collect reconstruction errors.
* Compute:
  + **Global σ** (overall spread of errors)
  + **Per-BU variance** and **Per-CP variance**

For HSBC and ANZ in history:

* HSBC variance = 0.64
* ANZ variance = 1.21

**Step 2: Estimate wobble for the new pair**

**Function:** pooled\_sigma(...)

* Combine BU-level and CP-level variances into one estimate.
* Take **square root** to return to σ (standard deviation).

σcell​= sqrt (0.83) ​≈ 0.911

*(We combine at the variance level because variances add cleanly; then square root to return to σ.)*

**3. Data drift checks – making comparisons fair**

At this point, we want to know: *“Do the new HSBC×ANZ deals look similar to historical deals in the same trading context, or do they look different?”*

To make this comparison **fair**, we use a **matching cohort** from history:

* First try strict match on context columns like Instrument, BuyCurr, SellCurr.
* If empty, gradually relax (drop one column, then another).
* If still empty, fallback to same BU, same CP, or global history.

For HSBC×ANZ, suppose the new deals are **Forward trades, Buy=USD, Sell=INR**.  
We select historical deals with the same context (Forward, USD→INR).  
This ensures we’re comparing **apples to apples** when measuring drift.

**📊 Variance Ratio (VR) – “How wide is the spread?”**

* **Definition:** Ratio of the spread (std) in new vs matched history.
* **Formula:**

A black text on a white background

Description automatically generated

* **Example (BuyAmount):**
  + Matched history std = 20
  + New HSBC×ANZ std = 30
  + VR = 30/20 = **1.5**

**📦 Population Stability Index (PSI) – “Did the shape change?”**

* **Definition:** Measures how the distribution of values has shifted.
* **How it works:**
  1. Break the **reference cohort** into bins (quantiles for numeric; categories for categorical).
  2. qi​ = fraction of history in bin i, pi​ = fraction of new data in bin i.
  3. Compute fractions in each bin for reference (qᵢ) and new (pᵢ).
  4. PSI :

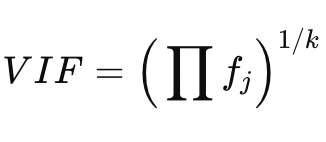
A black text with a white background

Description automatically generated

* **Example (SellCurr, quartiles):**
  1. History: 25%, 25%, 25%, 25%
  2. New HSBC×ANZ: 10%, 20%, 40%, 30%
  3. PSI ≈ **0.23** → moderate shift

**⚖️ Variance Inflation Factor (VIF) – “Roll it all up”**

* **Definition:** A multiplier ≥1 that inflates σ when new data looks harder to learn.
* **How it’s built:**
  + For each numeric feature:
    - Compute VR
    - Convert PSI into a factor (e.g., PSI=0.23 ⇒ ×1.2)
    - Take the larger of the two
  + Aggregate across features with **geometric mean**

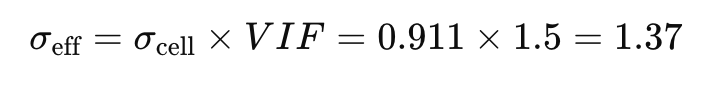


* + Cap at vif\_cap (e.g., ×3)

**Example (HSBC×ANZ):**

* BuyAmount: VR=1.5, PSI→1.2 ⇒ factor=1.5
* SellAmount: VR=1.4, PSI→1.5 ⇒ factor=1.5
  + VIF = √(1.5×1.5) = **1.5**

So:



**4. Sample size formula**

Finally, we use σ\_eff in the confidence-based formula:

A close-up of a mathematical equation

Description automatically generated

* z = 1.96 (95% confidence)
* δ = 0.25 (target precision band)
* σ\_eff = 1.37

n= [(1.96×1.37/0.25)2] ≈ 120

**Result:** HSBC×ANZ requires ~120 examples before retraining.  
With only 12 available now, shortfall = 108.

**5. Assumptions**

* Historical BU and CP variances are representative for unseen pairs.
* Matching cohorts are found via context columns; if empty, fallback to broader history.
* VR is floored at 1 (never reduce required n).
* PSI bins are stable; unseen categories go into *"\_\_OTHER\_\_".*
* VIF is capped (e.g., ×3) to prevent runaway values.

**6. Watch-outs**

* **Small samples (<10 new rows):** VR/PSI unstable; rely more on analytical formula.
* **Completely new categories:** PSI can spike; controlled by *"\_\_OTHER\_\_"* bin + VIF cap.
* **No cohort found:** fallback may make comparisons less precise → flag as lower confidence.
* **Conflicts between VR and PSI:** VIF takes the **conservative max**.

**7. Takeaway**

* **VR** shows if new deals are more volatile.
* **PSI** shows if deal distributions shifted.
* **Matching cohort** ensures these are compared against the right slice of history.
* **VIF** inflates σ to reflect extra difficulty.
* **Analytical formula** turns σeff into a safe, proactive sample size.

**For HSBC×ANZ today:** ~120 required; 12 available; retraining deferred.

PS:

**Example (numeric SellAmount)**

Reference values (8 rows): 5, 6, 7, 8, 9, 10, 11, 12

* Bins from reference quartiles:
  + Bin1: [5,7)
  + Bin2: [7,9.5)
  + Bin3: [9.5,11)
  + Bin4: [11,12]

Reference distribution q :

* Bin1: 2/8 = 0.25
* Bin2: 3/8 = 0.375
* Bin3: 1/8 = 0.125
* Bin4: 2/8 = 0.25

New values (8 rows): 6, 7, 7, 8, 11, 12, 12, 12  
New distribution p :

* Bin1: 1/8 = 0.125
* Bin2: 3/8 = 0.375
* Bin3: 0/8 = 0 (use ε=1e-6 to avoid log(0))
* Bin4: 4/8 = 0.50

A black text with a white background

Description automatically generated

* Bin1: (0.125–0.25) ln(0.125/0.25) = 0.0866
* Bin2: (0.375–0.375) ln(1) = 0
* Bin3: (0–0.125) ln(ε/0.125) ≫ positive contribution
* Bin4: (0.50–0.25) ln(0.50/0.25) = 0.1733

Total PSI ≈ **>0.25** (major shift).