PM2.5 Taiwan (Mar 2017): Descriptive Analysis

High-Dimensional Time Series Project

Andrea Bianco, Luca Nudo, Emidio Grillo

February 5, 2026

Dataset Setup & Basic Data QA

Data Source & Coverage

- **Source:** Taiwan AirBox $PM_{2.5}$ measurements (hourly), March 2017
- **Spatial selection:** Stations inside Taiwan (lat 21–26, lon 119–123) + Kinmen Island (V46)
- **Raw Taiwan panel:** **T: 744 hours** **N: 511 stations**

Data Structure & Quality

- **Structure:** `time` as hourly index; station series as columns **V2, V3, . . .**
- **Quality checks:** ✓ No missing, non-finite, or negative values
- **Cleaning:** Removed 3 faulty/non-Taiwan stations (V29, V70, V348) → final **N=508** stations

✓ **Clean, fully observed panel with consistent measurements for reliable analysis**

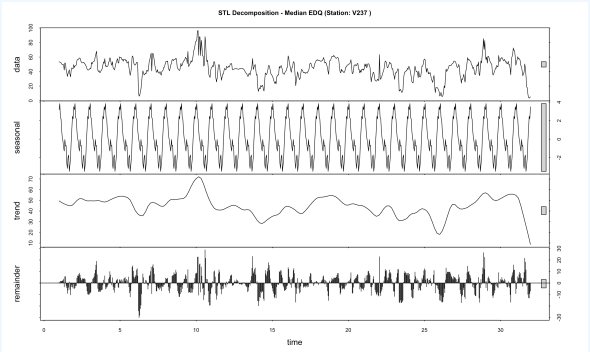
EDQ 500: Representative Station (Median Quantile)

Concept & Methodology

- **Goal:** Identify a representative station minimizing total quantile loss relative to all other stations
- **EDQ 500:** Median quantile ($p = 0.5$) series from the cleaned Taiwan panel
- **Computation:** Selected as the station that minimizes the total quantile loss across all others

✓ **Serves as a single, robust reference series for time series decomposition and seasonality analysis**

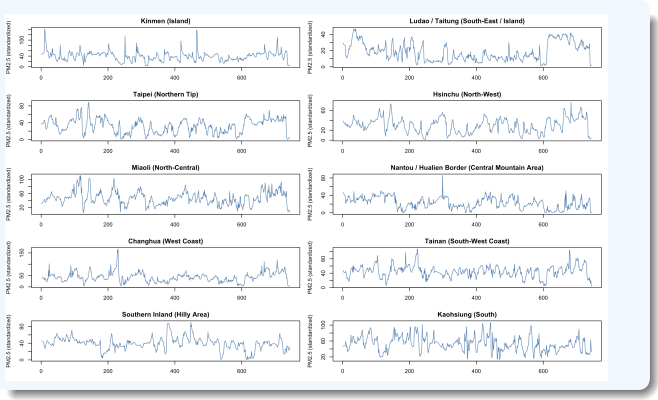
EDQ 500: STL Decomposition



Decomposition Components

- **Data:** Original hourly $PM_{2.5}$ series (top panel)
- **Seasonal:** Daily pattern extracted (freq = 24), regular oscillation
- **Trend:** Long-term evolution, stable 50 $\mu g/m^3$ with peaks and decline
- **Remainder:** Residuals after removing trend and seasonal, irregular fluctuations around zero

Strategic Stations: Raw Historical Series



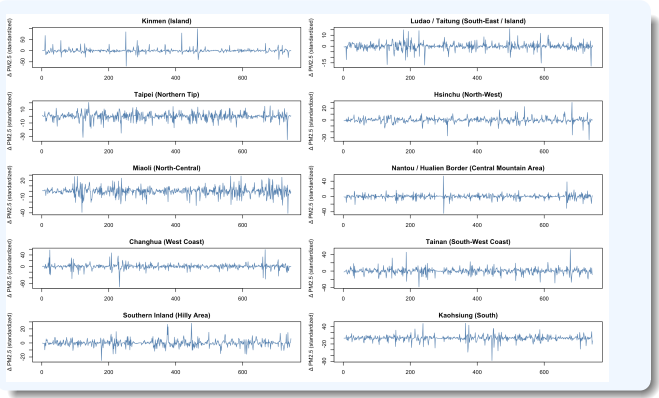
Station Coverage

- **Selection: 10** strategic stations covering islands, coasts, urban and inland areas
- **Data:** Raw (non-differenced) hourly $PM_{2.5}$ series

Observations

- **Behavior:** Strong variability, visible trends and level shifts
- **Insight:** ! Series exhibit clear **non-stationary** patterns

Strategic Stations: Differenced Series



Transformation Applied

- **Transformation:** First differences $\nabla Y_t = Y_t - Y_{t-1}$ of $PM_{2.5}$ series
- **Effect:** Removal of trends and level shifts

Results

- **Behavior:** Fluctuations centered around zero with more stable variance
- **Insight:** ✓ Differencing improves **stationarity**, enabling dependence analysis

Stationarity Tests: ADF and KPSS

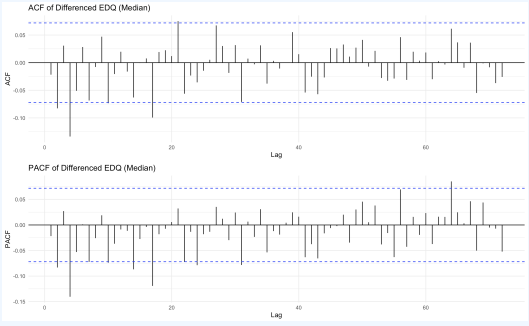
Testing Framework

- **Augmented Dickey–Fuller (ADF):** Tests the null hypothesis of a unit root, i.e. non-stationarity
- **KPSS:** Tests the null hypothesis of level stationarity, providing a complementary perspective to ADF
- **Joint interpretation:** Consistency between ADF rejection and KPSS non-rejection supports stationarity

Series	ADF (Reject unit root)	KPSS (Stationary)
Raw levels	100% ✓	30% !
First differences	100% ✓	100% ✓

✓ **First differencing removes trends and level shifts, yielding panel-wide stationarity**

ACF & PACF of Differenced EDQ 500



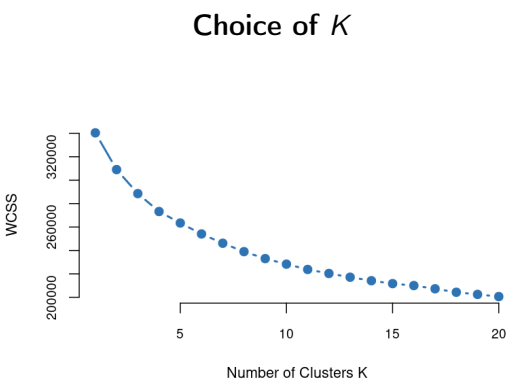
Why ACF and PACF?

- **Goal:** Quantify how dependence decays over time lags
- **Range:** Up to 72 lags (3 days) to capture persistence beyond 24h cycles

How to read the plots

- **ACF:** Overall correlation with past values at lag h (persistence pattern)
- **PACF:** Direct correlation at lag h after removing intermediate lags
- **Bands:** Spikes outside the confidence bands indicate non-negligible dependence

Clustering Strategy & Preliminary Selection of K

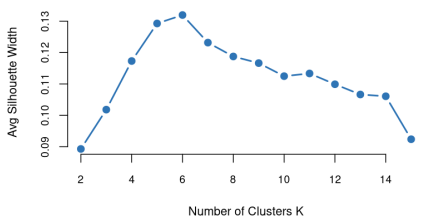


The k-means Approach

We partitioned the 508 stations into K groups by minimizing the **Within-Cluster Sum of Squares (WCSS)**.

- **Standardization**: Z-score normalization to focus on the **shape** of temporal profiles rather than absolute PM2.5 levels.
- **Metric**: Standard Euclidean distance.
- **Elbow ambiguity**: Primary elbow between $K = 5$ and $K = 6$, requiring a second decision criterion.

Final Selection of K & Interpretability



- The **Average Silhouette Width** reaches its absolute maximum at $K = 6$.
- This mathematically confirms $K = 6$ as the partition with the highest internal cohesion.

The Final Choice: $K = 6$

- **Optimal Trade-off**: Point of maximum statistical precision without over-fragmenting the dataset.
- **Physical Interpretability**: Balances detail and parsimony, mapping diverse local dynamics (**urban**, **industrial**, and **coastal/mountain** zones) into meaningful regional profiles.

Spatial Distribution of PM2.5 Clusters

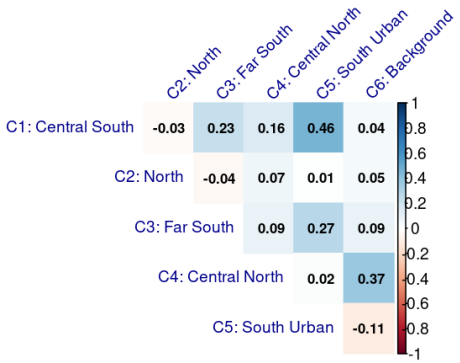
The Western Footprint

The 508 stations are not distributed randomly: they follow the **Western Plains' human activity footprint**. The **Central Mountain Range** creates a stark contrast between the industrial/urban West and the near-empty East.

- **Striking Spatial Coherence**: PM2.5 shocks are driven by **large-scale regional dynamics**.
- **Regional Pockets**: The partition identifies key physical barriers:
  - **North**: Isolated by the *Taipei Basin effect*.
  - **West Coast**: A continuous industrial corridor for pollutant transport.
  - **South**: Specific zones of air stagnation against the foothills.
- **Topographic Synergy**: A physical mapping of how topography traps and channels air quality shocks, reflecting the interaction between local emission sources and mountain barriers.

Click here to open the Interactive Map

Cluster Correlation Analysis (CCM)



- **Industrial Axis (C1-C5)**: Highest correlation (**0.46**), western coast acts as atmospheric corridor.
- **North Isolation (C2)**: Near-zero correlation, strong **basin effect**.
- **Central-Rural Link (C4-C6)**: Moderate coupling (**0.37**), easier dispersion.

Signal Enhancement

The transition to cluster averages performed a **denoising** process, highlighting regional dynamics over local micro-emissions.

Modeling Options Given Our Data

Data Context

- **Setting:** High-dimensional multivariate time series
- **Structure:** Strong temporal dependence with potential cross-sectional interactions

Candidate Modeling Approaches

- **Classical VAR:** Natural multivariate baseline, but infeasible when the number of series is large relative to sample size
- **Regularized VAR (LASSO-based):** Enable estimation in high dimensions by enforcing sparsity and controlling overfitting
- **Low-rank / Factor-based VAR:** Capture common latent dynamics shared across many series, reducing dimensionality

Why a VAR-LASSO Model?

Chosen Approach

- **Model:** Regularized VAR framework (LASSO-based) for high-dimensional multivariate time series
- **Objective:** Estimate cross-dependencies while controlling complexity and overfitting

Temporal Dependence & Interpretability

- **Dynamics:** VAR explicitly models lagged interactions, preserving the time-series structure
- **Readability:** Selected coefficients can be interpreted as a parsimonious dependence network

✓ **VAR-LASSO is a scalable compromise: it retains dynamic structure while remaining stable and interpretable in high dimensions**

VAR-LASSO: Model Assumptions

Core Assumptions

- **Linearity:** Joint dynamics follow a linear VAR structure
- **Stationarity:** Stable second-order properties after preprocessing
- **Sparsity:** Only few cross-lagged interactions are non-zero

Physical Justification

Regional transport and topographic barriers imply **localized interactions**, supporting the sparsity assumption.

✓ **VAR-LASSO exploits linearity, stationarity, and sparsity**

VAR-LASSO: High-Dimensional Challenge

The Problem

- **Parameters:** $K^2 \times p$ coefficients
- **Classical OLS:** Fails due to overfitting and ill-conditioning

The Solution

ℓ_1 penalty shrinks weak coefficients to zero:

- Automatic variable selection
- Stable estimation
- Improved generalization

✓ **LASSO regularization enables reliable high-dimensional estimation**

VAR-LASSO: Model Formulation

VAR(p) Model

$$Y_t = c + \sum_{\ell=1}^{p} A_\ell Y_{t-\ell} + \varepsilon_t$$

where $A_\ell \in \mathbb{R}^{K \times K}$ are coefficient matrices.

LASSO Objective

$$\min_{A_1,\ldots,A_p} \left\{ \text{RSS} + \lambda \sum_{\ell=1}^{p} \|A_\ell\|_1 \right\}$$

Interpretation

Non-zero $A_\ell^{ij} \neq 0$ means cluster j at lag ℓ influences cluster i.

Model Results: Predictive Performance

Error Metrics

- **RMSE: 8.127** **MAE: 5.313**
- **MAE significantly lower than RMSE**: Model is stable on average error, but penalized by extreme spikes (weather events, traffic peaks).

Computed on differenced $PM_{2.5}$ — relative error is more informative than absolute scale

Interpretation

On **differenced series**, these values are plausible and indicate:

- Controlled average prediction error
- Presence of local outliers (meteorological episodes, unobserved emissions)

Model Results: Signal Extraction

Predictive Association

- **Correlation (Predicted vs Observed): 0.355**

Analysis

For **differenced $PM_{2.5}$**, correlation naturally drops because:

- Differencing removes trend, leaving only volatile changes
- $PM_{2.5}$ is noisy (unobserved wind, rain, sudden emissions)
- High-dimensional + rolling window estimation

Model Results: Validation and Baseline

Uncertainty Calibration

- **95% CI Coverage: 88.2%** (target: 95%)
- **Slight under-coverage**: typical of penalized high-dimensional VAR

Coverage 85–90% is acceptable in high-dimensional contexts

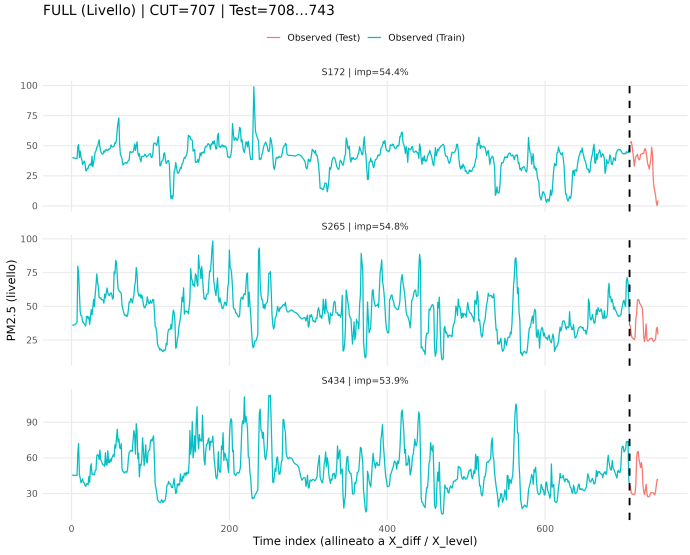
Baseline Comparison

- **Baseline**: Zero-change predictor (predict 0 on differenced series)
- **Result**: VAR-LASSO reduces RMSE/MAE and improves correlation across majority of stations

Key Message

The model extracts **cross-sectional and temporal dependencies** not captured by a naïve approach, justifying VAR-LASSO over baseline.

Rolling Prediction Strategy



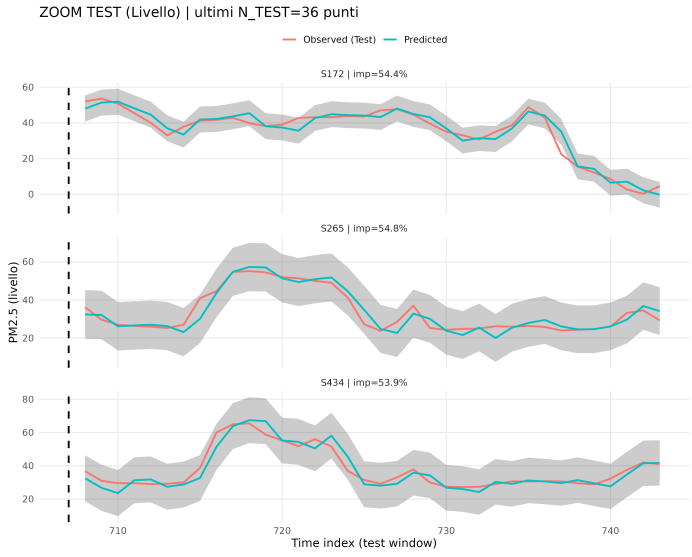
Short-Term Memory

- Uses **last 36 observations**
- **Short-memory** dynamics
- Recent patterns focused

Rolling Procedure

- Fit on 36-obs window
- Predict next point
- Shift forward, repeat

Model Fit: Test Window Performance



Tracking Dynamics

- Predictions follow observed values
- Grey band = uncertainty
- Limited lag on peaks

Key Takeaway

- Adaptive forecasting
- Extracts short-run signal
- Stable despite volatility

Why These Three Stations?

Selection Criterion

The three displayed stations were selected based on **maximum improvement over baseline**:

$$\text{Improvement} = 100 \times \frac{\text{MSE}_{\text{baseline}} - \text{MSE}_{\text{model}}}{\text{MSE}_{\text{baseline}}}$$

where baseline predicts **zero change** on differenced series.

Interpretation

- These stations show where VAR-LASSO **captures cross-sectional and temporal dependencies most effectively**
- Highest improvement indicates locations where **inter-cluster dynamics matter most**
- Model adds maximum value compared to naïve prediction strategy