Supuestos para la Elección de Modelo

Detección de anomalías en el consumo comercial e industrial de gas

28 de agosto de 2025

1. Datos y alcance

- **Fuente y estructura**: archivo `df_contugas.csv` con **847 946** observaciones y 6 variables (*Fecha, Cliente, Segmento, Presion, Temperatura, Volumen*).
- **Frecuencia esperada**: horaria.
- **Calidad**: no se observaron nulos en las numéricas analizadas.
- **Nota**: la serie agregada mezcla múltiples *Clientes*; por ello, las conclusiones temporales deben verificarse por cliente.

2. Supuestos y evidencia

2.1. Distribución de variables

- **Presión**: histograma y QQ-plot evidencian **bimodalidad** (bandas $\approx$ 3–4 y 17–18) y outliers; se rechaza normalidad.
- **Temperatura**: aproximadamente gaussiana con colas moderadas.
- **Volumen**: no normal, **asimetría positiva** y **cola pesada**, con abundantes ceros y atípicos.

Implicación: evitar umbrales paramétricos globales; preferir métricas robustas (MAD/IQR) y modelos que no requieran normalidad.

2.2. Relaciones entre variables

Se observó correlación moderada entre *Volumen* y *Temperatura* ($\approx +0{,}334$) y correlación negativa con *Presión* ($\approx -0{,}304$).
Implicación: utilizar un **modelo multivariable** con exógenas (Temperatura, Presión) y, de ser posible, un indicador de **régimen de presión**.

2.3. Dependencia temporal y estacionalidad

La ACF agregada se ve plana por mezclar clientes, pero a nivel operativo se espera estacionalidad **diaria (24 h)** y **semanal (168 h)** por *Cliente*.
Implicación: modelar por cliente (o con cliente codificado) e incluir **lags 1/24/168** y variables de calendario (hora del día, día de semana).

2.4. Cero-inflación y outliers

Existen muchos ceros (horas sin operación) y outliers altos en *Volumen*.
Implicación: usar **escalado robusto**, tratar los ceros explícitamente (filtrado/etiqueta) y detectar sobre **scores de error** con umbrales robustos.

3. Consecuencias sobre la elección de modelo

Condición	**Modelo**
Estacionalidad clara, relación casi lineal con exógenas	Regresión **robusta** (Huber/Quantile) con lags
No linealidad relevante e interacciones	**Isolation Forest** multivariable (lags, rolling,
Dinámica temporal no lineal fuerte y mucha historia	**LSTM** (pronóstico o autoencoder)

4. Umbral robusto basado en MAD

Sea el residuo $r_t = y_t - \hat{y}_t$. Para cada *Cliente*:

$$\text{MAD} = \text{mediana}\left(\,|r_t - \text{mediana}(r)|\,\right) \tag{1}$$

$$\tau_{\text{cli}} = \text{mediana}(|r|) + k \cdot 1{,}4826 \cdot \text{MAD}, \quad k \in [3,\, 3{,}5] \tag{2}$$

Se marca anomalía cuando $|r_t| > \tau_{\text{cli}}$. Alternativamente, fijar τ_{cli} en el cuantil alto del error (P99–P99,5).

5. Validación y calibración

- **Backtesting** con inyección de anomalías sintéticas (picos/caídas, cambios de nivel) para medir *precision@k* y *recall*.
- **Estabilidad** del % de anomalías por cliente y semana.
- **Revisión experta** de los top-k eventos por semana.
- **Calibración por cliente** de umbrales (MAD, cuantil o EVT).

6. Checklist de supuestos

- No normalidad en *Volumen* y bimodalidad en *Presión* (✓).
- Outliers frecuentes y cero-inflación (✓).
- Correlación moderada de *Volumen* con *Temperatura* (+) y *Presión* (–) (✓).
- Modelado por *Cliente* con lags 1/24/168 y exógenas (✓).
- Verificación de estacionalidad por *Cliente* con ADF/KPSS y ACF/PACF (□ completar con tablas/figuras por cliente).

7. Resumen ejecutivo

Con base en 847 946 observaciones horarias, *Volumen* presenta asimetría positiva, cola pesada y muchos ceros; *Presión* es **bimodal** y *Temperatura* es casi gaussiana. Se rechaza la normalidad del objetivo y se observan correlaciones moderadas de *Volumen* con *Temperatura* y *Presión*. Bajo estos supuestos, la detección se hará **por Cliente**, con lags estacionales (1, 24, 168 h) y exógenas, comenzando con **Regresión robusta + residuales**, contrastando con **Isolation Forest** y escalando a **LSTM** si la dinámica lo requiere.