Aprendizaje automático

Series Temporales

FACULTAD DE INGENIERÍA | UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Agenda

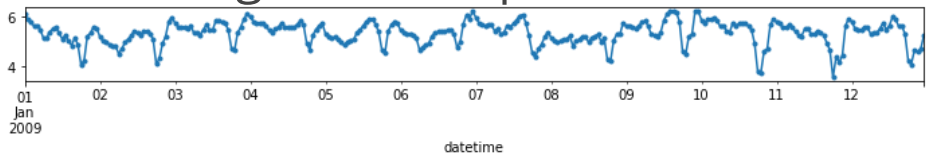
- Series temporales
- Detección de anomalías
- Proyecto IIE-Telefónica
  - Problema
  - Detectores
  - Plataforma
- DC-VAE

Series Temporales

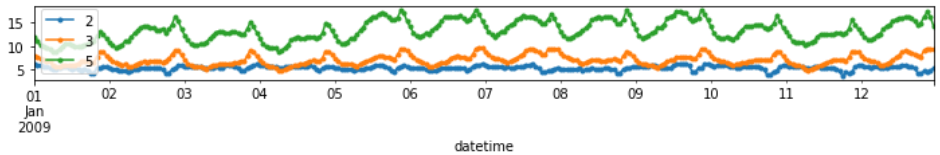
FACULTAD DE INGENIERÍA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Series temporales

- Una serie temporal es un conjunto de datos que se recopilan, registran o generan secuencialmente a lo largo del tiempo, con intervalos regulares o irregulares entre las observaciones.
- Estos datos pueden representar cualquier tipo de fenómeno que varíe con el tiempo, como la temperatura, el precio de las acciones, la demanda de un producto, la velocidad del viento, entre otros.
- Univariadas: una única variable a lo largo del tiempo. Para cada instante solo tengo un valor



- Multivariadas: múltiples variables a lo largo del tiempo. Para cada instante tengo un vector.



Series temporales

Cantidad de publicaciones por año (1985-2020) con las palabras claves: “Time series analysis”

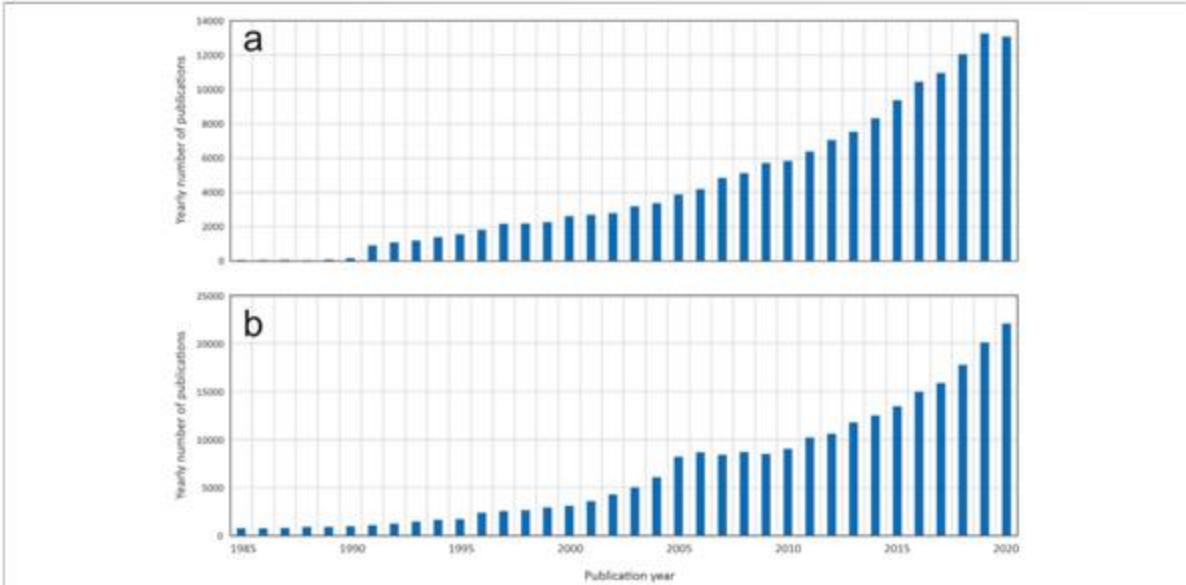


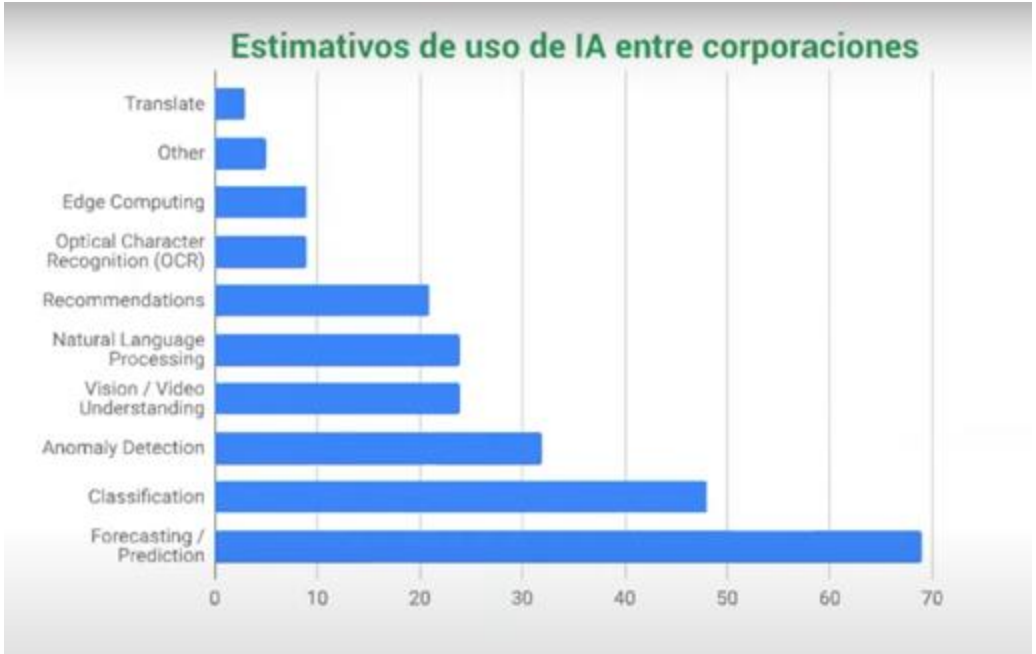
FIGURE 1 | Yearly number of publications with the keyword “time series analysis” per year during 1985–2020 from WoS **(A)** and Scopus **(B)** database.

Flavio Cannavò, Andrea Cannata, Reik V. Donner, y Mikhail Kanevski. "Editorial: Advanced Time Series Analysis in Geosciences". Frontiers in Earth Science, vol. 9, 2021. Disponible en: https://www.frontiersin.org/articles/10.3389/feart.2021.666148. DOI: 10.3389/feart.2021.666148. ISSN: 2296-6463.

Detección de anomalías

Detección de anomalías

- Problemas más comunes de los clientes de Google Cloud



Charla Nicolas Loeff: https://www.youtube.com/watch?v=SYuoTN4iQi8

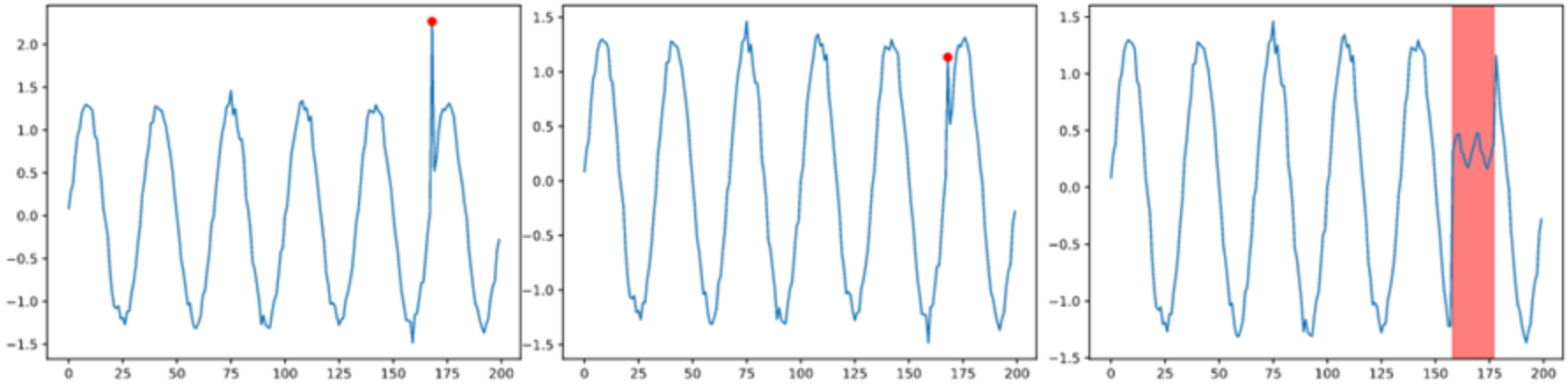
Detección de anomalías

- Anomalía: Las anomalías son patrones en los datos que no se ajustan a una noción bien definida de comportamiento normal.
- Causas: fraudes, ataques al sistema, daños en el sistema, degradación del servicio, etc.
- Retos:
  - El umbral entre la normalidad y lo anormal en general no es preciso.
  - Las anomalías pueden cambiar con el tiempo. Ej: ciberataques.
  - Evolución del comportamiento normal. Lo que hoy es normal quizas en el futuro no sea representante.
  - Etiquetas disponibles.
  - A menudo el comportamiento normal contiene ruido que se puede confundir con las anomalías.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.

Detección de anomalías

- Clasificación clásica:

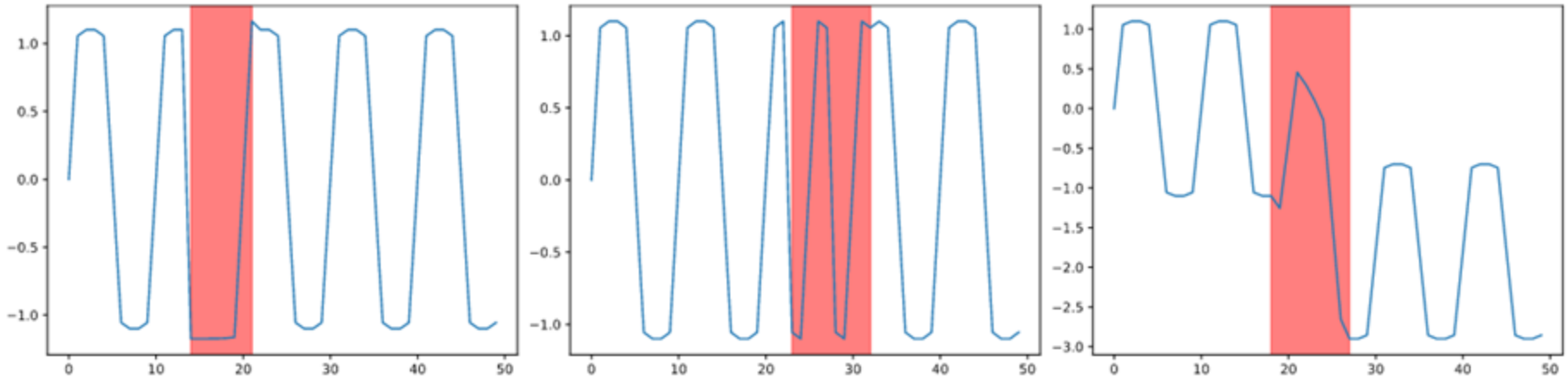


Puntual Contextual Colectiva

Lai, K. H., Zha, D., Xu, J., Zhao, Y., Wang, G., & Hu, X. (2021, June). Revisiting time series outlier detection: Definitions and benchmarks. In Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 1).

Detección de anomalías

- Colectiva:



Forma Estacional Tendencia

Lai, K. H., Zha, D., Xu, J., Zhao, Y., Wang, G., & Hu, X. (2021, June). Revisiting time series outlier detection: Definitions and benchmarks. In Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 1).

Proyecto IIE - Telefónica

FACULTAD DE INGENIERÍA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Proyecto IIE - Telefónica

- Telefónica cuenta con información del desempeño de sus servicios que desea analizar para detectar anomalías que permitan identificar fraudes, ataques o alteraciones en el funcionamiento de los servicios.

Objetivos

- Detección de anomalías de manera automática
- Detección en tiempo real: el proceso de detección no puede durar más que el paso de muestreo de la serie
  - Muestreo cada 5, 10, y 15 minutos.
- Implementación escalable
  - Series de múltiples fuentes.

Motivación

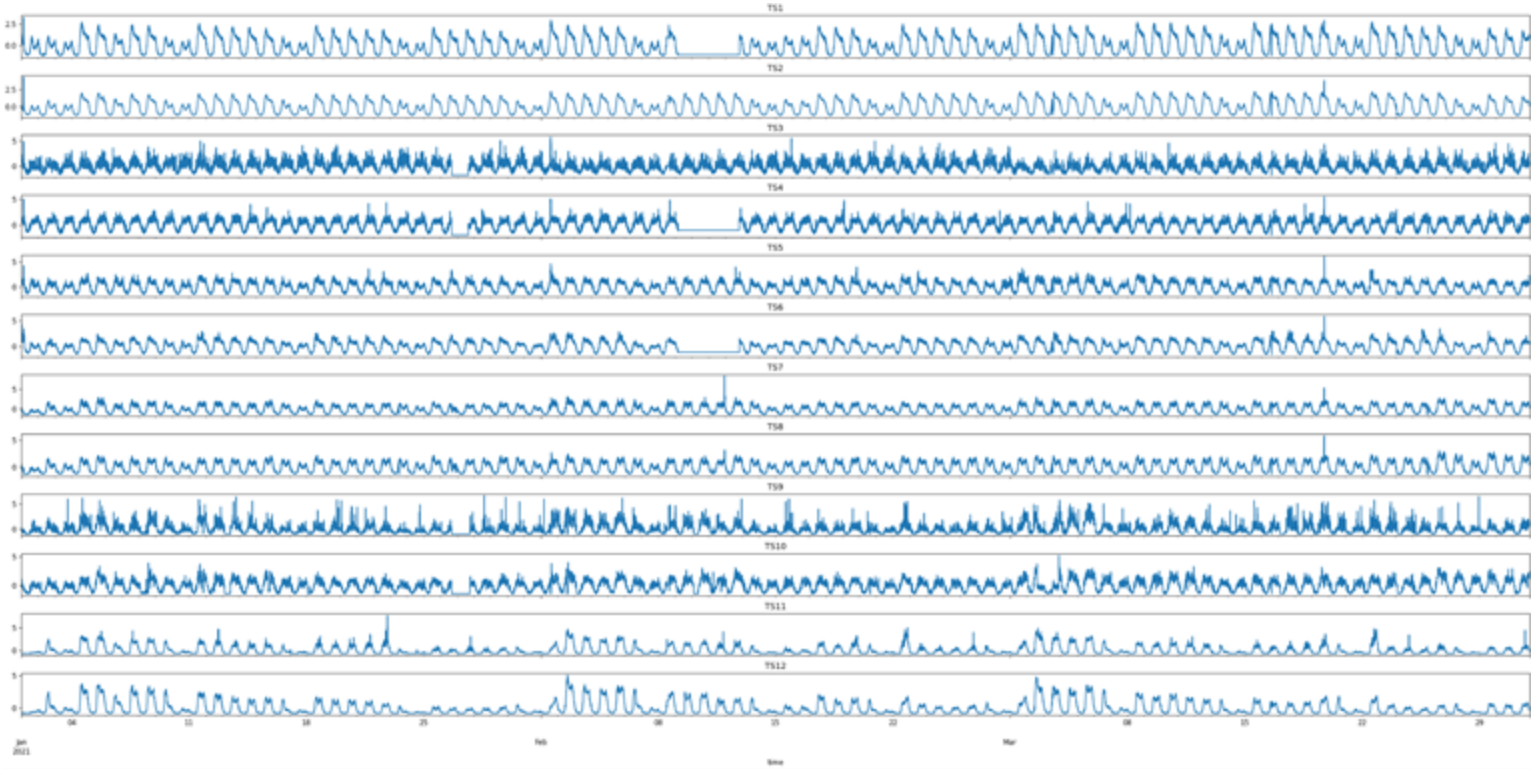
- Plataformas de Big Data que permiten incorporar analítica de datos y visualización. Ej: InfluxDB, Grafana.
- Experiencia en el área.

Tareas

- Investigación
- Desarrollo conjunto
- Transferencia

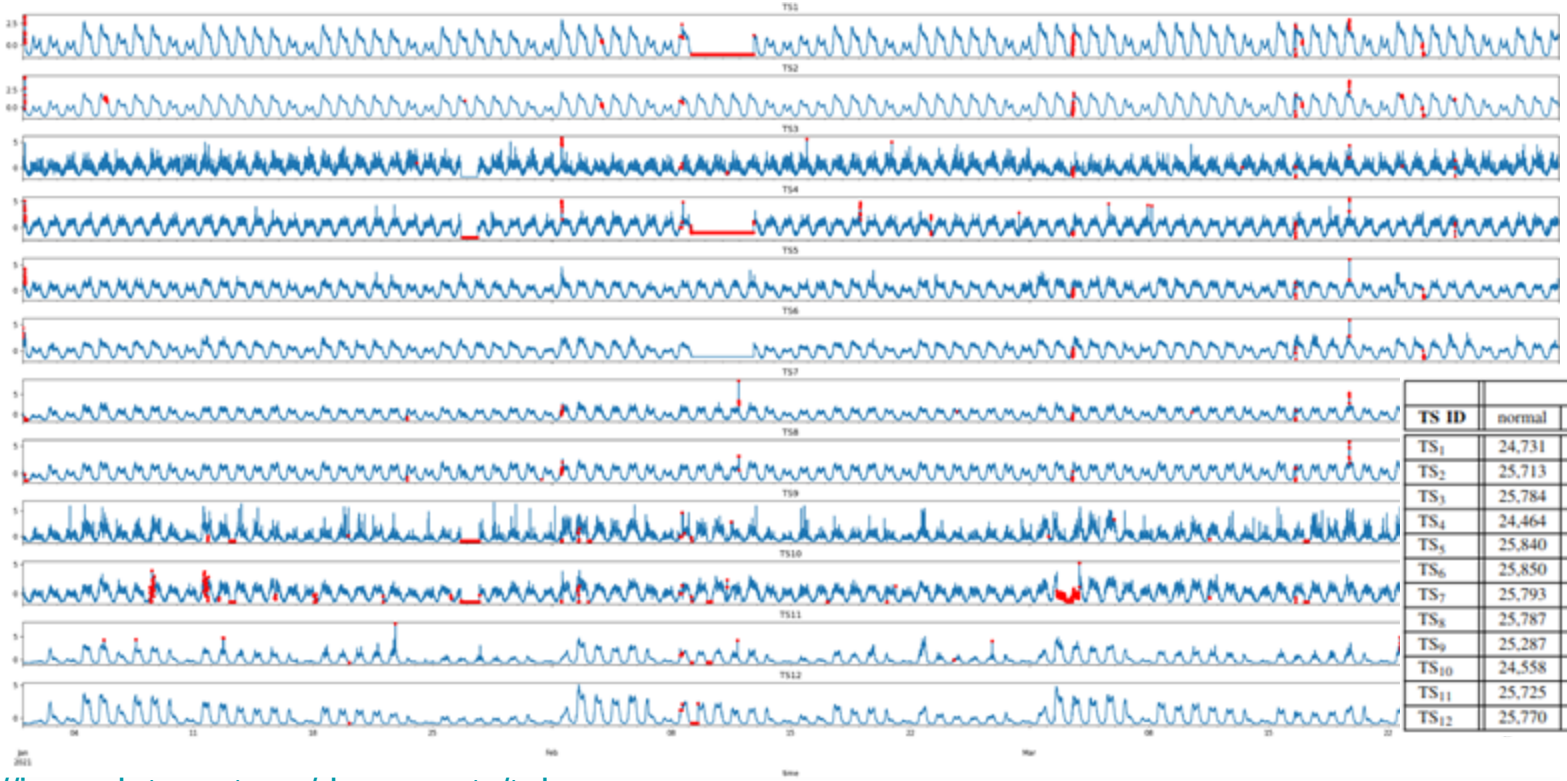
Proyecto IIE - Telefónica

Datos



Proyecto IIE - Telefónica

Datos



	training		
TS ID	normal	anomalous	%
TS_1	24,731	1,183	4,6%
TS_2	25,713	201	0,8%
TS_3	25,784	130	0,5%
TS_4	24,464	1,450	5,6%
TS_5	25,840	74	0,3%
TS_6	25,850	64	0,2%
TS_7	25,793	127	0,5%
TS_8	25,787	127	0,5%
TS_9	25,287	627	2,4%
TS_{10}	24,558	1,356	5,2%
TS_{11}	25,725	189	0,7%
TS_{12}	25,770	144	0,6%

https://ieee-dataport.org/documents/telco

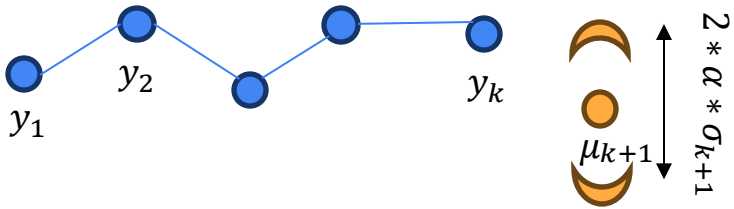
Detectores: SARIMA + Kalman

- La estrategia consiste en aprender un modelo estocástico de la serie a partir de datos históricos del comportamiento de la misma y luego comprobar si las nuevas muestras o secuencias de muestras caen en un espacio de baja probabilidad de ocurrencia.
- Dada una serie de tiempo hasta el timpo k, el proceso se basa en predecir la distribución de la serie en el tiempo k+1 y compararla con el valor que llegue en ese instante

$$\boxed{t = k}$$

$$Y_k = (y_1, \dots, y_k)$$

$$p(y_{k+1}|Y_k) = \mathcal{N}(\mu_{k+1}, \sigma_{k+1})$$



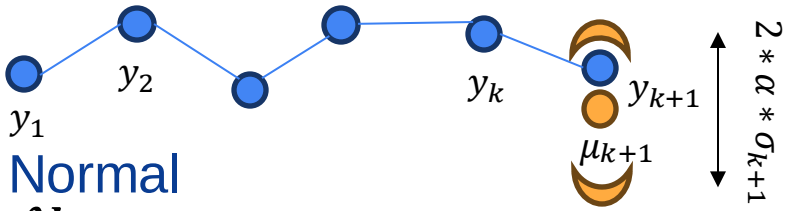
Tagliafico, S. M., González, G. G., Fernández, A., Gómez, G., & Acuna, J. (2018). Real time anomaly detection in network traffic time series. *Proceedings ITISE*, 1417-1428.

Detectores: SARIMA + Kalman

- La estrategia consiste en aprender un modelo estocástico de la serie a partir de datos históricos del comportamiento de la misma y luego comprobar si las nuevas muestras o secuencias de muestras caen en un espacio de baja probabilidad de ocurrencia.
- Dada una serie de tiempo hasta el timpo k, el proceso se basa en predecir la distribución de la serie en el tiempo k+1 y compararla con el valor que llegue en ese instante

$$t = k + 1$$

$$||y_{k+1} - \mu_{k+1}|| < \alpha * \sigma_{k+1}$$



y_{k+1}

Tagliafico, S. M., González, G. G., Fernández, A., Gómez, G., & Acuna, J. (2018). Real time anomaly detection in network traffic time series. *Proceedings ITISE*, 1417-1428.

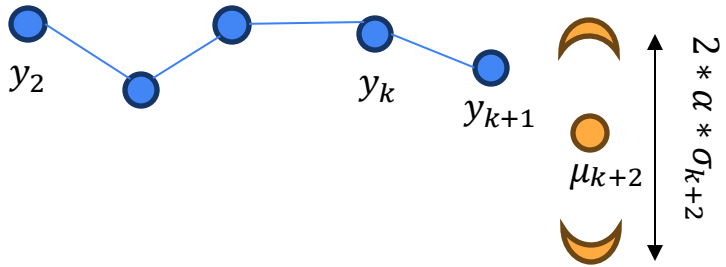
Detectores: SARIMA + Kalman

- La estrategia consiste en aprender un modelo estocástico de la serie a partir de datos históricos del comportamiento de la misma y luego comprobar si las nuevas muestras o secuencias de muestras caen en un espacio de baja probabilidad de ocurrencia.
- Dada una serie de tiempo hasta el timpo k, el proceso se basa en predecir la distribución de la serie en el tiempo k+1 y compararla con el valor que llegue en ese instante

$$t = k + 1$$

$$Y_{k+1} = (y_2, \dots, y_{k+1})$$

$$p(y_{k+2}|Y_{k+1}) = \mathcal{N}(\mu_{k+2}, \sigma_{k+2})$$

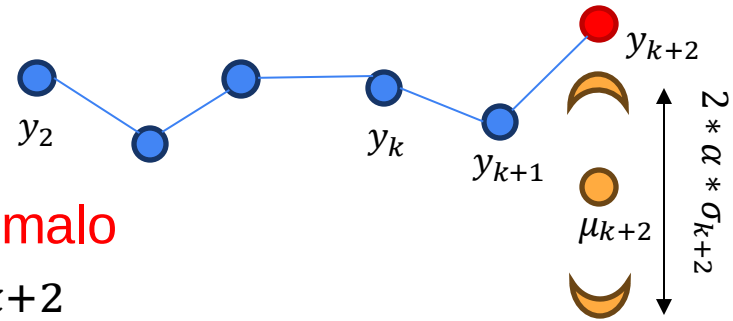


Tagliafico, S. M., González, G. G., Fernández, A., Gómez, G., & Acuna, J. (2018). Real time anomaly detection in network traffic time series. *Proceedings ITISE*, 1417-1428.

Detectores: SARIMA + Kalman

- La estrategia consiste en aprender un modelo estocástico de la serie a partir de datos históricos del comportamiento de la misma y luego comprobar si las nuevas muestras o secuencias de muestras caen en un espacio de baja probabilidad de ocurrencia.
- Dada una serie de tiempo hasta el timpo k, el proceso se basa en predecir la distribución de la serie en el tiempo k+1 y compararla con el valor que llegue en ese instante

$$t = k + 2$$



$$||y_{k+2} - m_{k+2}|| > \alpha * \sigma_{k+2} \Rightarrow \text{Anómalo } y_{k+2}$$

Tagliafico, S. M., González, G. G., Fernández, A., Gómez, G., & Acuna, J. (2018). Real time anomaly detection in network traffic time series. *Proceedings ITISE*, 1417-1428.

Detectores: SARIMA + Kalman

Entrenamiento + validación:

Definir el mejor modelo que aproxima el comportamiento normal de la serie

- No supervisado
- Datos normales
- Hiperparámetros $(p, d, q) \times (P, D, Q, s)$

→

Punto de operación:

Definir el valor de α

- Supervisado
- Etiquetas o experto en los datos

→

Operación / Test

https://www.statsmodels.org/stable/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html#statsmodels-tsa-statespace-sarimax-sarimax

Proyecto IIE - Telefónica

Detectores: SARIMA + Kalman

Ventajas:
- Interpretable
- Pocos parámetros
- Tiempo real

Desventajas
- No captura comportamientos a largo plazo
- Se necesita conocer sobre filtros y espacio de estados
- Univariado solamente

Proyecto IIE - Telefónica

Detectores: ADTK

- Detección de anomalías sobre series univariadas y multivariadas con un conjunto de modelos.
- La implementación del modelo ADTK se basa en el módulo Python Anomaly Detection Toolkit (ADTK)
- En este caso el proceso de detección utiliza en forma simultánea hasta 12 modelos a elección y para la decisión final sobre si un punto en la serie es anómalo se utiliza el criterio de votación entre los resultados individuales de cada detector.

0 - OutlierDetector (LocalOutlierFactor)
1 - OutlierDetector (IsolationForest)
2 - QuantileAD
3 - InterQuartileRangeAD
4 - GeneralizedESDTestAD
5 - PersistAD

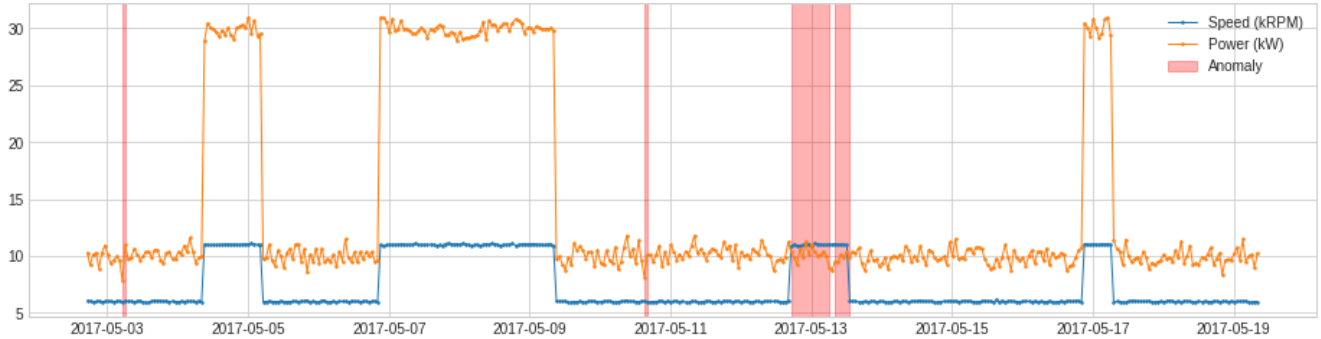
6 - LevelShiftAD
7 - VolatilityShiftAD
8 - ClassicSeasonalDecomposition
9 - SeasonalAD
10 - PcaAD
11 - MinClusterDetector

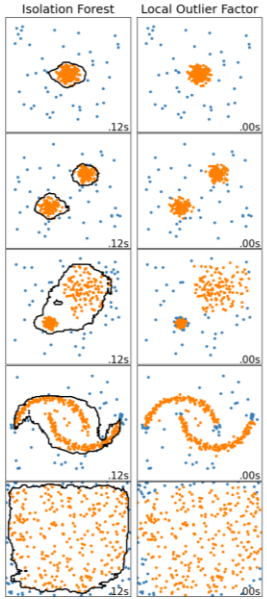
https://adtk.readthedocs.io/en/stable/index.html

Proyecto IIE - Telefónica

Detectores: ADTK

- 0 - OutlierDetector (LocalOutlierFactor) – Multivariado
  - Basado es estimación local de la densidad de probabilidad con kNN, adecuado para dimension de datos moderada.
- 1 - OutlierDetector (IsolationForest) - Multivariado
  - Basado en árboles aleatorios, eficiente con datos de alta dimensión.



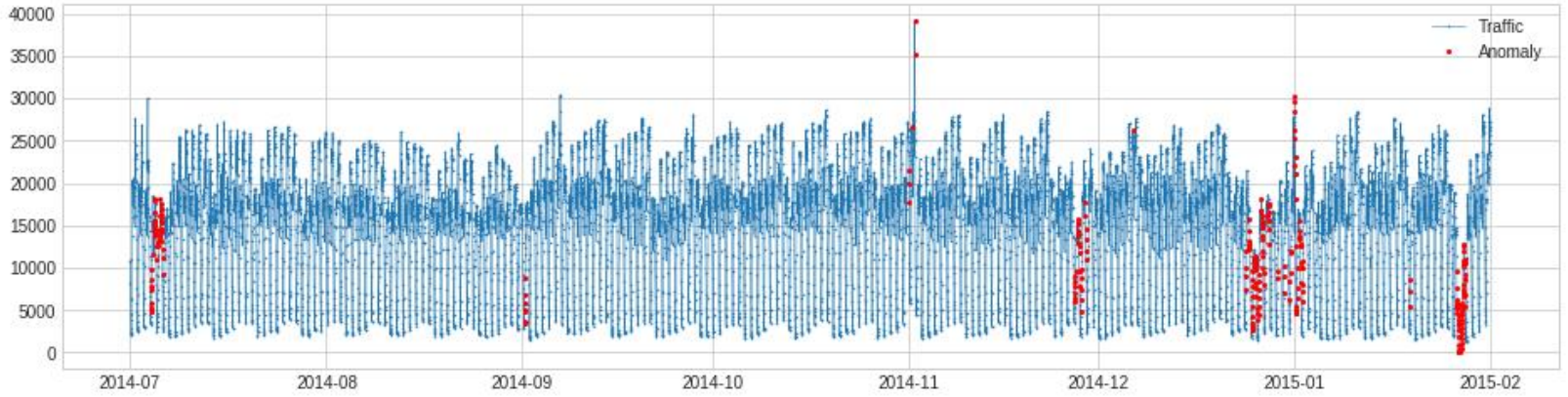


https://adtk.readthedocs.io/en/stable/index.html

Proyecto IIE - Telefónica

Detectores: ADTK

- 9 – SeasonalAD – Univariado
  - Detección de cambio en el patrón estacionario

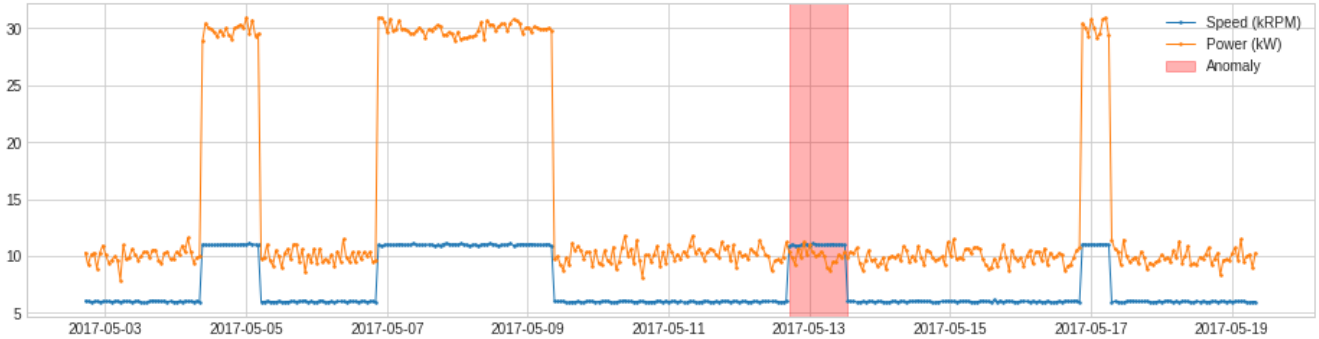


https://adtk.readthedocs.io/en/stable/index.html

Proyecto IIE - Telefónica

Detectores: ADTK

- 10 – PcaAD – Multivariado
  - Detección basada en el análisis de componentes principales (PCA) en series de tiempo multivariadas (se toma cada punto de tiempo como un vector en un espacio de alta dimensión). Las anomalías se determinan a partir del error de reconstrucción de eso vectores.



https://adtk.readthedocs.io/en/stable/index.html

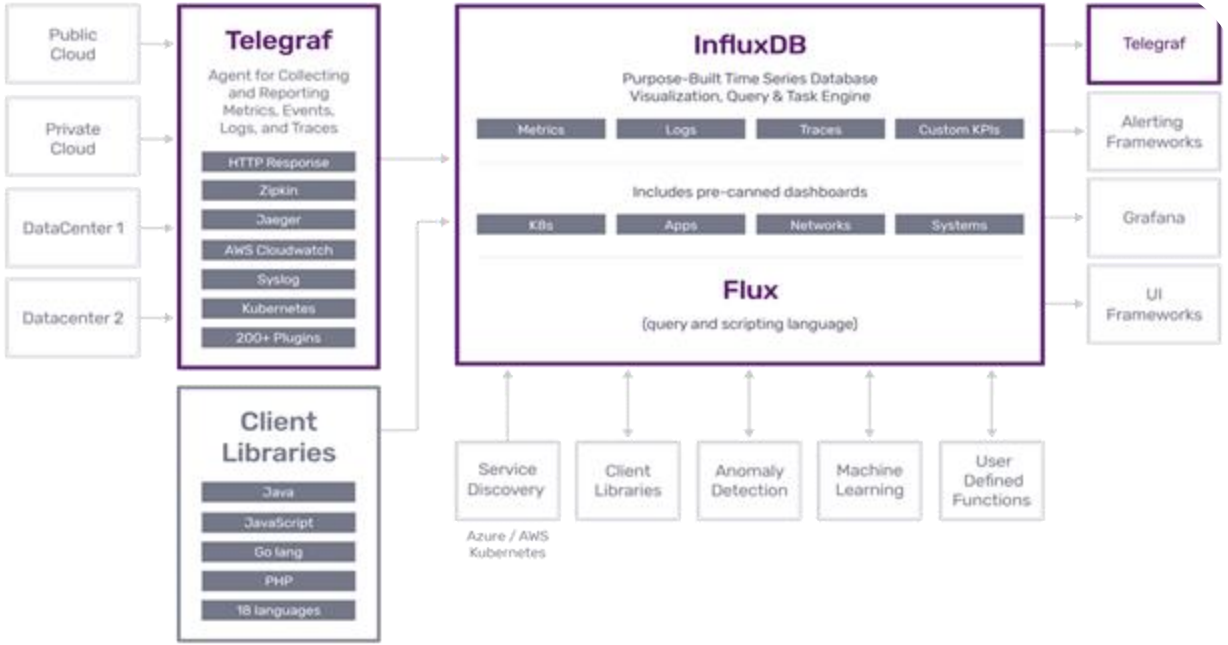
Detectores: ADTK

Ventajas:
- Métodos sencillos
- Interpretable
- Pocos parámetros
- Tiempo real
- Modelos univariados y multivariados
- Capacidad de detectar anomalías de diferentes tipos

Desventajas
- No captura comportamientos a largo plazo
- Modelos multivariados no utilizan la componente tiempo
- Pueden ser complejo determinar la combinación de detectores correcta.

Proyecto IIE - Telefónica

Plataforma

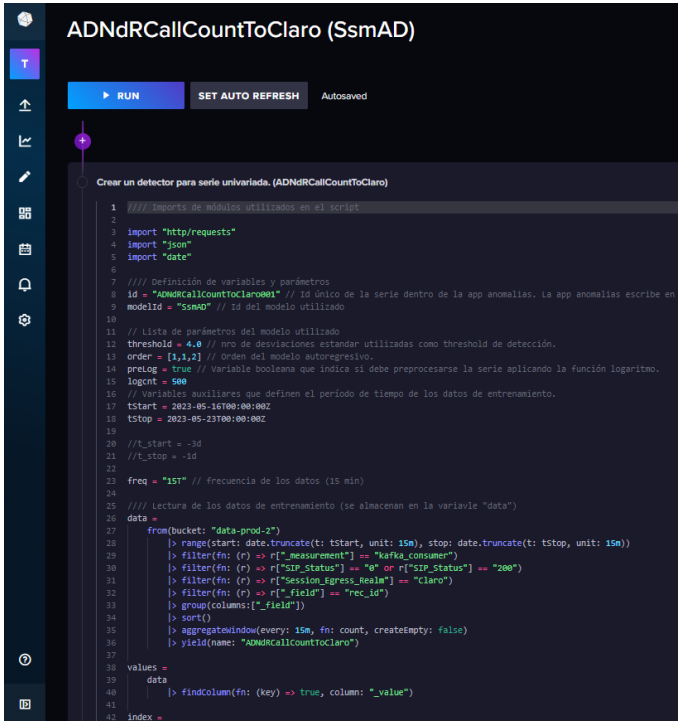


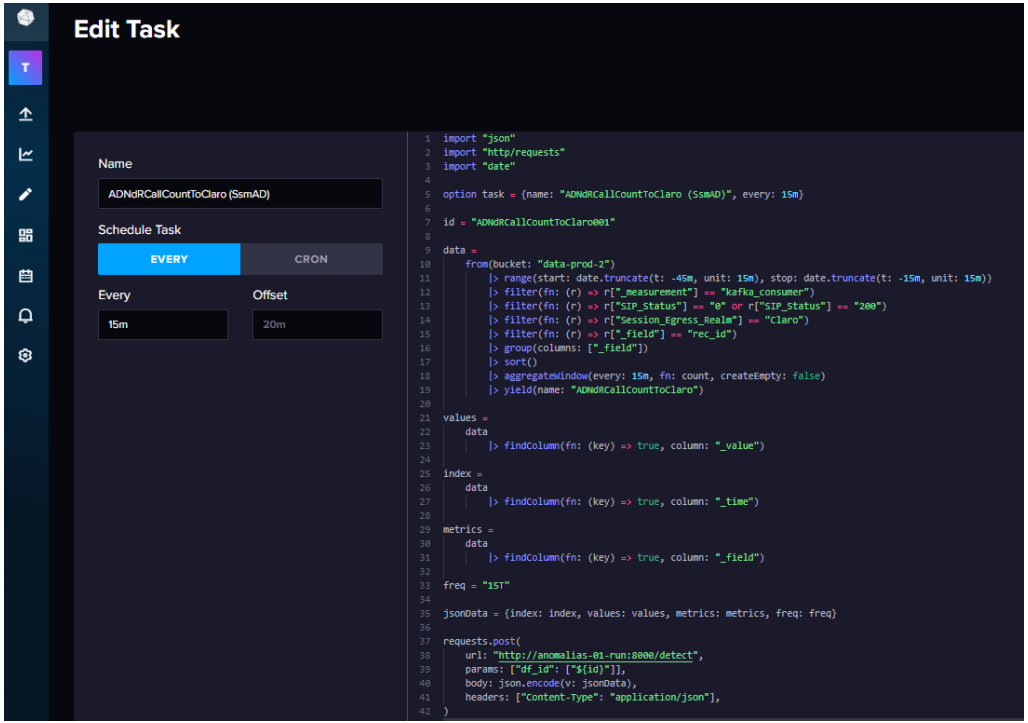
https://www.influxdata.com/

Proyecto IIE - Telefónica

Plataforma

- Crear un caso (Influx)

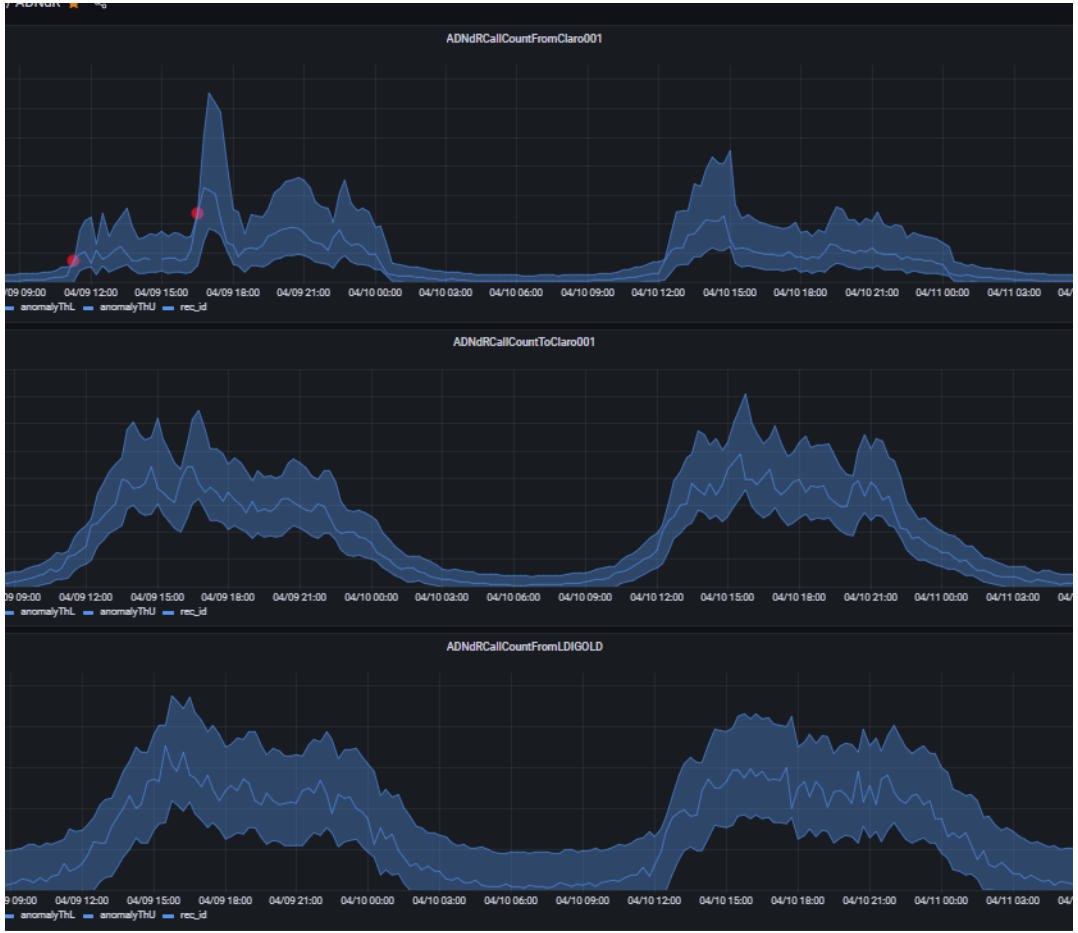




Proyecto IIE - Telefónica

Plataforma

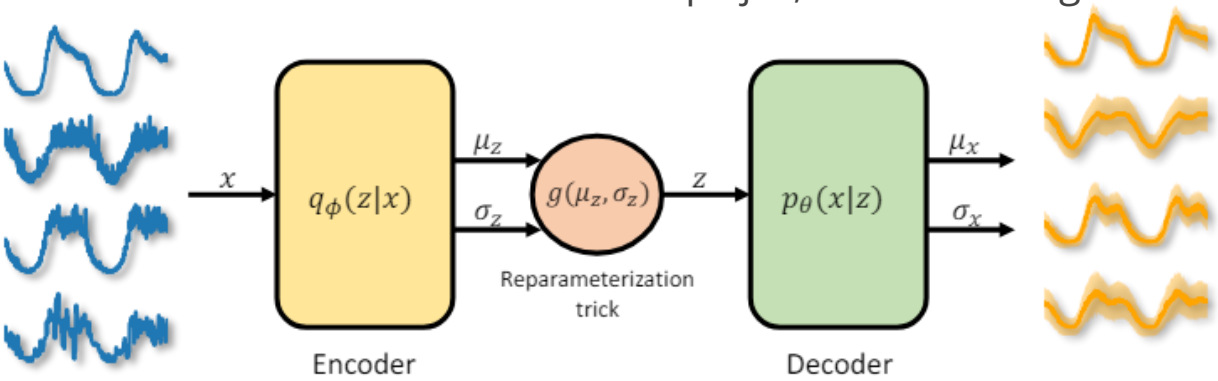
- Crear un caso (Grafana)

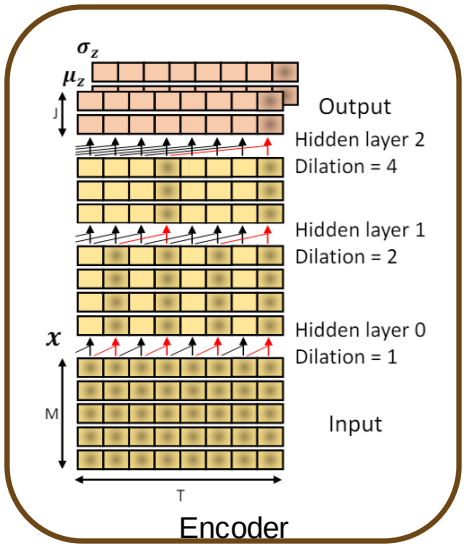


DC-VAE

DC-VAE

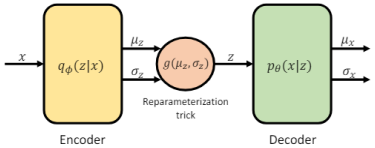
- Modelo de redes neuronales multivariado que explota tanto la relación entre variables como la correlación temporal
- Arquitectura: Redes neuronales convolucionales de 1 dimensión con dilatación (DC).
- Modelo: Variational Auto-Encoders (VAE). Modelos generativos con capacidad de modelar distribuciones complejas, como las imágenes.



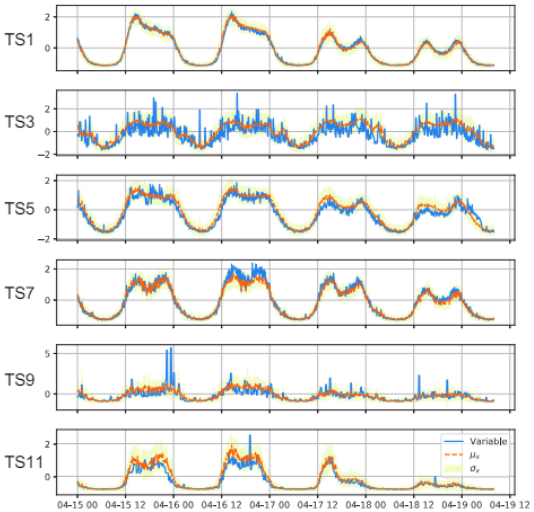


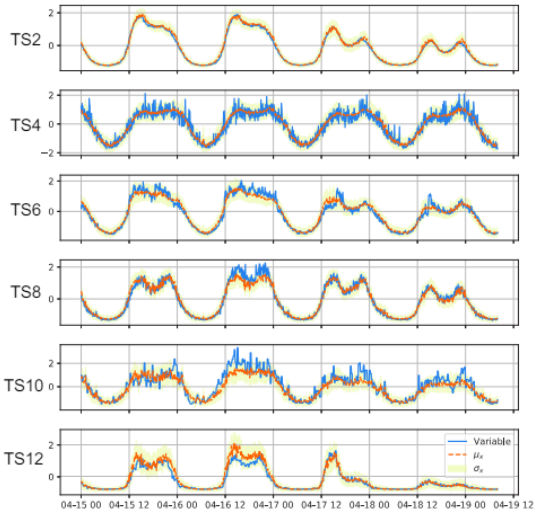
González, G. G., Tagliafico, S. M., Fernández, A., Gómez, G., & Casas, P. (2023). One Model to Find Them All Deep Learning for Multivariate Time-Series Anomaly Detection in Mobile Network Data. IEEE Transactions on Network and Service Management.

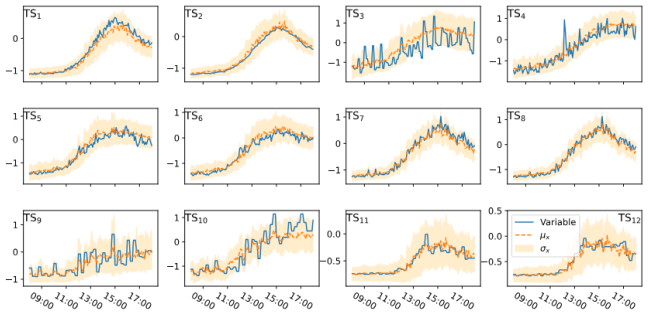
DC-VAE



- Desempeño

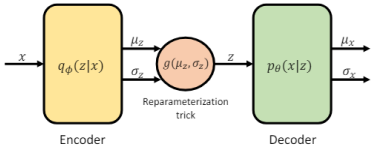






González, G. G., Tagliafico, S. M., Fernández, A., Gómez, G., & Casas, P. (2023). One Model to Find Them All Deep Learning for Multivariate Time-Series Anomaly Detection in Mobile Network Data. IEEE Transactions on Network and Service Management.

DC-VAE

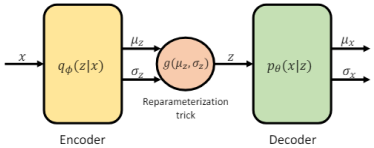


- Desempeño

	ENS-15			S-EXPS			ARIMA			S-VAE			*DC-VAE*		
TS ID	R_r	P_r	$F1_r$	R_r	P_r	$F1_r$	R_r	P_r	$F1_r$	R_r	P_r	$F1_r$	R_r	P_r	$F1_r$
TS_1	45%	50%	48%	45%	88%	60%	64%	92%	**75%**	23%	56%	32%	58%	71%	**64%**
TS_2	37%	100%	54%	70%	96%	**81%**	59%	95%	**73%**	16%	92%	27%	74%	20%	**67%**
TS_3	78%	33%	47%	78%	58%	**67%**	78%	46%	58%	71%	50%	59%	86%	47%	**60%**
TS_4	75%	59%	**66%**	67%	41%	**51%**	58%	38%	46%	63%	25%	36%	63%	21%	**32%**
TS_5	73%	73%	**73%**	45%	63%	53%	64%	64%	**64%**	50%	20%	29%	75%	50%	60%
TS_6	88%	62%	**72%**	63%	63%	63%	75%	50%	60%	14%	100%	25%	57%	83%	68%
TS_7	77%	63%	**69%**	69%	53%	60%	69%	46%	56%	45%	100%	63%	72%	90%	**80%**
TS_8	67%	44%	53%	56%	36%	43%	56%	56%	**56%**	57%	35%	43%	44%	80%	**57%**
TS_9	10%	17%	**12%**	5%	5%	5%	19%	9%	12%	6%	4%	4%	17%	11%	**13%**
TS_{10}	8%	18%	11%	48%	44%	46%	48%	38%	42%	39%	81%	**52%**	52%	59%	**55%**
TS_{11}	58%	21%	31%	50%	32%	**39%**	67%	26%	37%	67%	17%	27%	100%	25%	**40%**
TS_{12}	0%	0%	0%	100%	67%	**80%**	100%	24%	**38%**	0%	0%	0%	100%	11%	22%
mean	51%	45%	45%	58%	54%	**54%**	63%	49%	51%	38%	48%	33%	67%	47%	**52%**
median	63%	47%	51%	60%	55%	**57%**	64%	46%	56%	42%	43%	31%	68%	49%	**59%**

González, G. G., Tagliafico, S. M., Fernández, A., Gómez, G., & Casas, P. (2023). One Model to Find Them All Deep Learning for Multivariate Time-Series Anomaly Detection in Mobile Network Data. IEEE Transactions on Network and Service Management.

DC-VAE



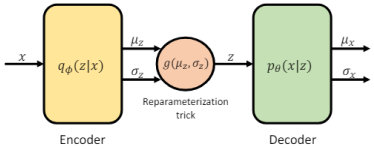
Ventajas:

- Capacidad de capturar relaciones amplias de tiempo
- Escalable: multivariado
- Explota la correlación temporal
- Visual

Desventajas

- Complejo de entrenar y buscar hiperparámetros
- Mucha cantidad de parámetros
- Complejo de adaptar

DC-VAE

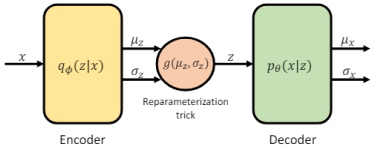


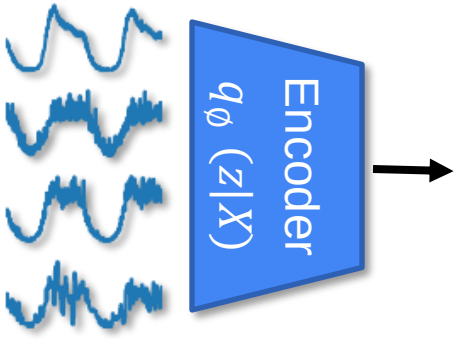
Dada la complejidad del entrenamiento, se necesita encontrar una manera de adaptar el modelo de manera más efectiva, que volver a entrenar todo el modelo frente a:

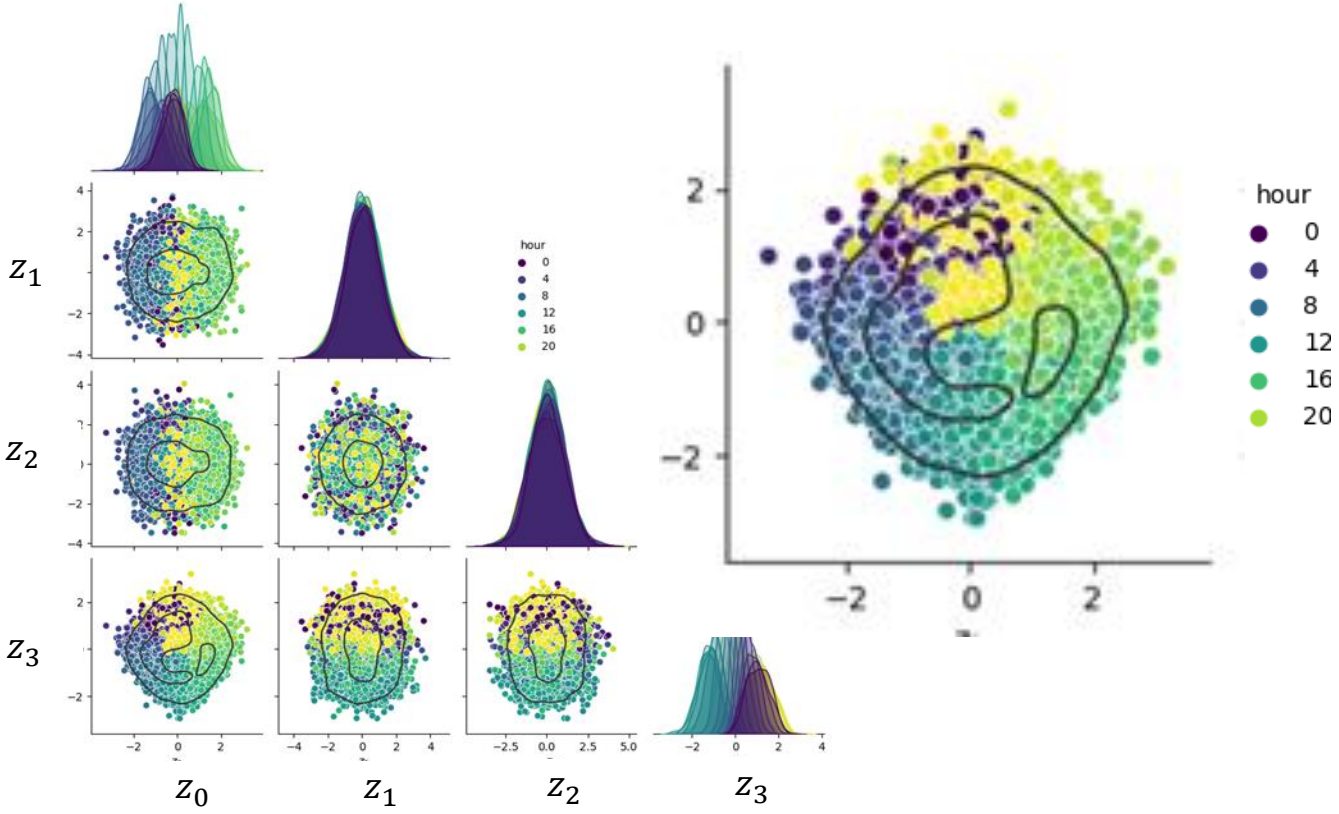
- Cambios en la distribución de las series
- Incorporar nuevas series a monitorear
- Generative Replay

DC-VAE

- Generative Replay

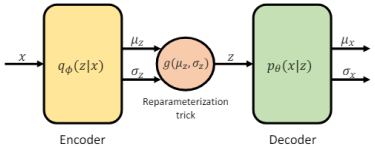


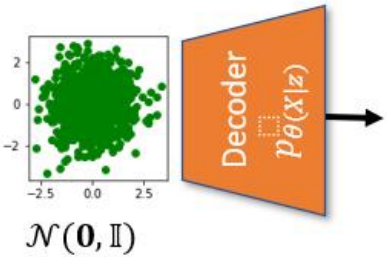


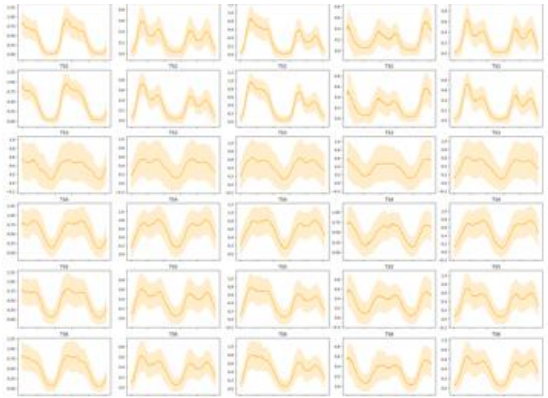


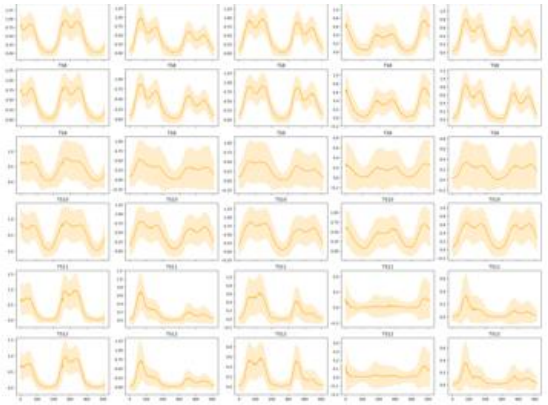
DC-VAE

- Generative Replay

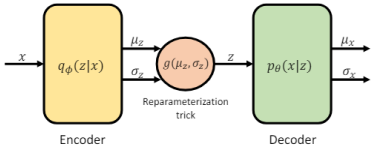




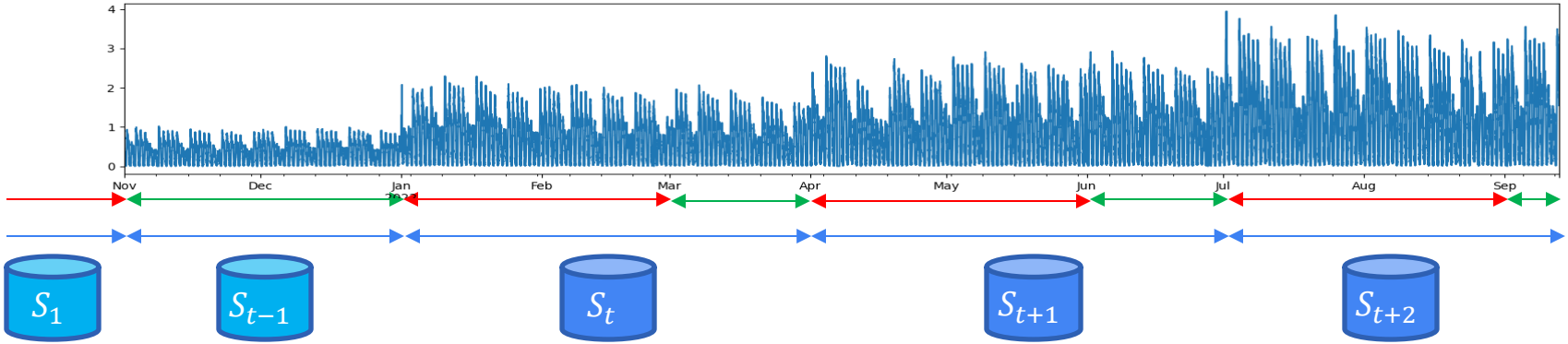




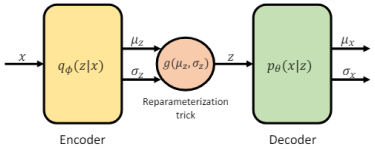
DC-VAE



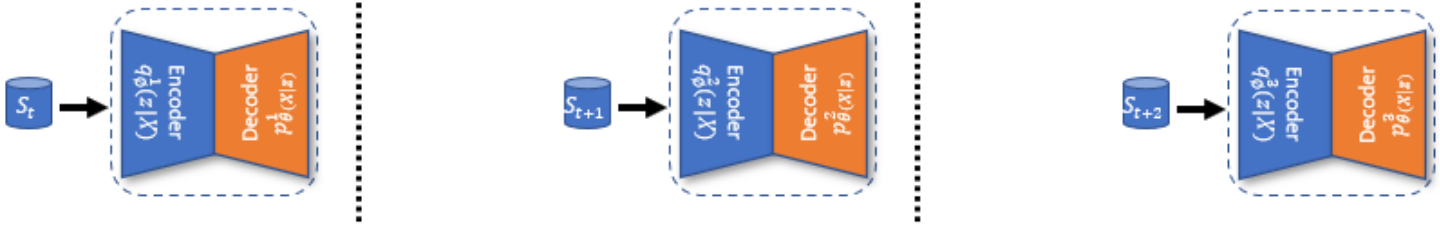
- Generative Replay



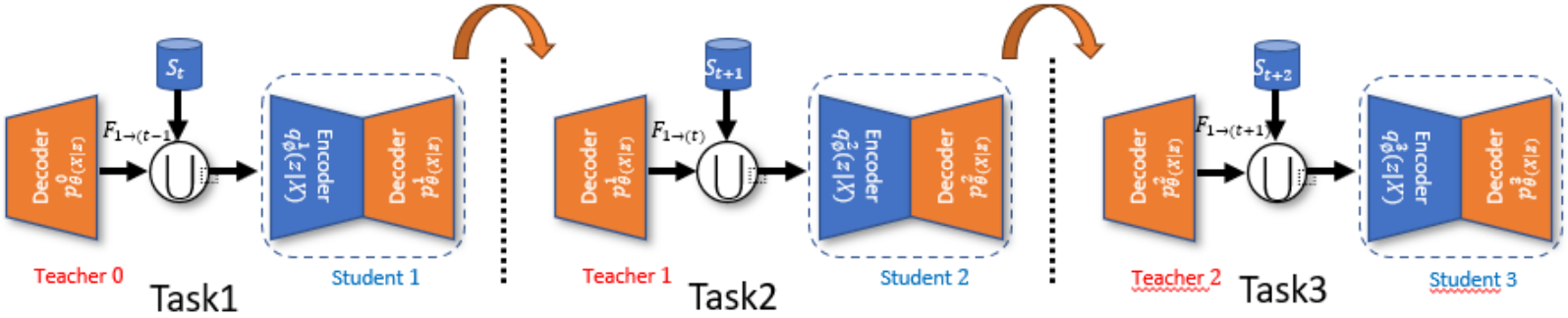
DC-VAE



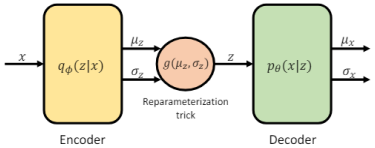
No GenDeX



GenDeX



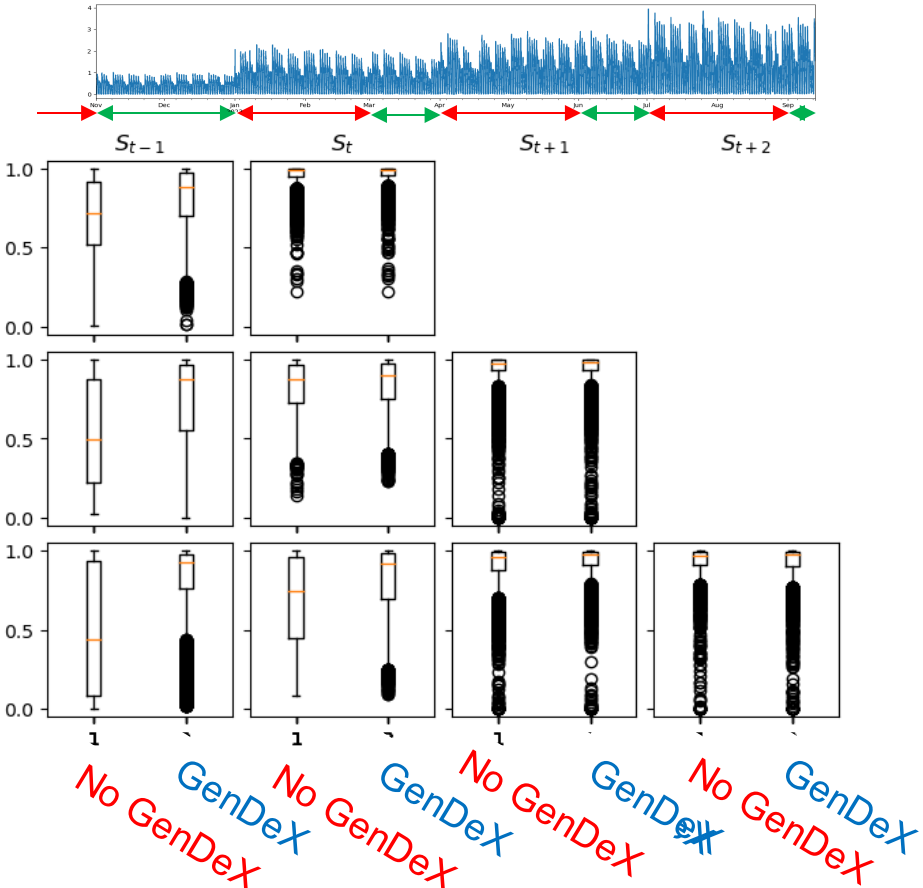
DC-VAE



DC-VAE1 (Task 1)

DC-VAE2 (Task 2)

DC-VAE3 (Task 3)



DC-VAE

- Trabajo a futuro
  - Comparación entre modelo multivariado y global
  - Adaptación a los cambios de dominio
  - Generalización a múltiples dominios.