Online Detection of Emergent Phenomena in Streaming Data

Edward Austin[1] Idris Eckley[1] Lawrence Bardwell[1] Peter Willis[2]

[1]STOR-i CDT Lancaster University [2] BT Group PLC

Motivation

BT monitor the amount of data passing through their network at any given moment so that faults can be detected in real time.

The nature of the data means that over each day an expected shape exists.

When this shape is deviated from, it either represents a surge in customer demand, or an outage on the network.

It is these outages that BT seek to detect, in real time, so that faults can be fixed faster and customers reconnected sooner.

Functional Data Representation

A functional observation is a smooth curve defined over an interval $\mathcal{T} \subset \mathbb{R}$.

Here the observation $X(t)$ represents a function for the amount of data recorded at any time $t \in \mathcal{T}$.

In practice, functional data is not observed in real time but on a discrete, dense, uniform grid $\mathcal{T} = \{1, \ldots, T\}$.

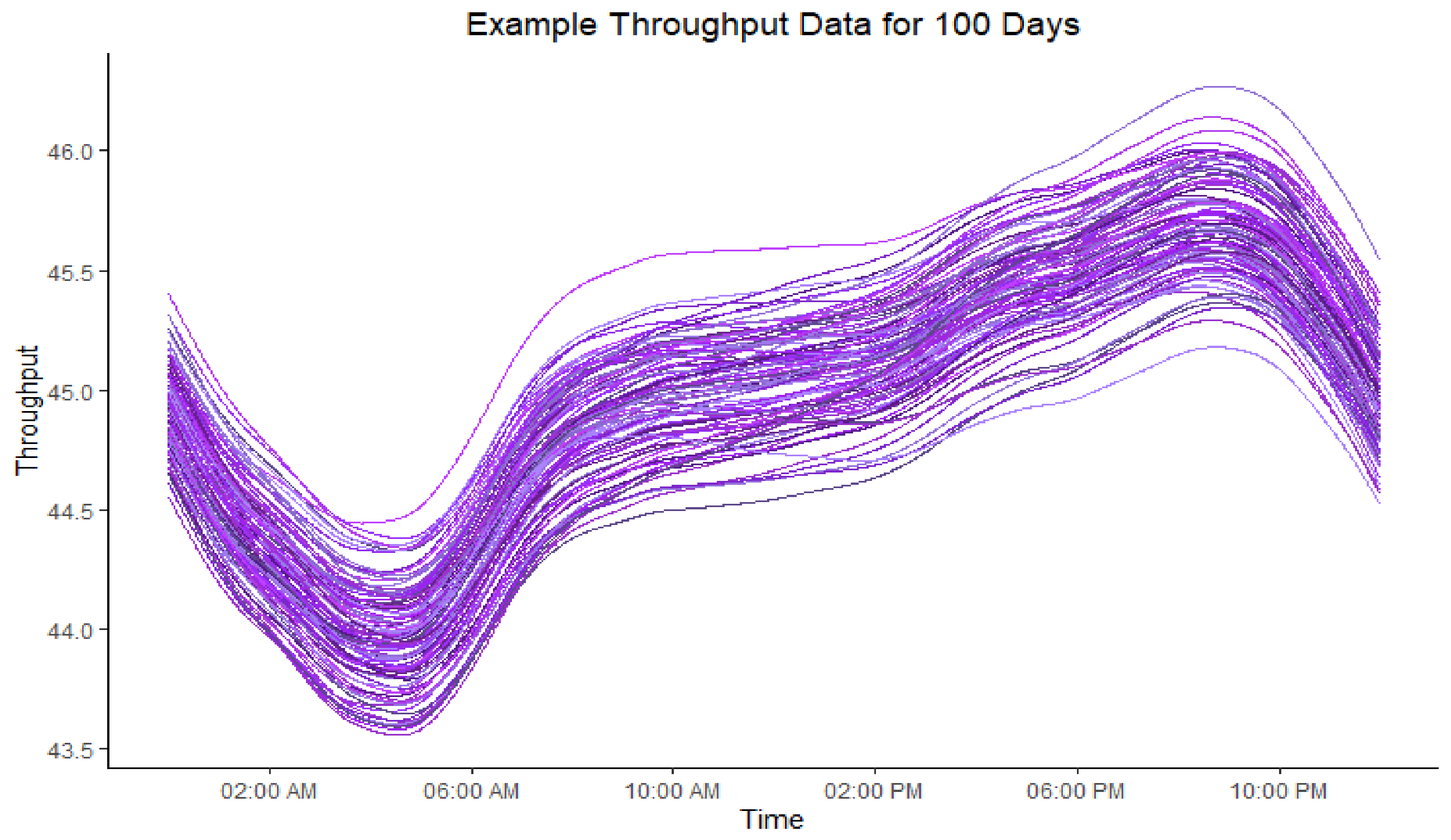

Figure 1: Example of a series of functional observations observed over a single day.

Modelling the Underlying Shape

In order to capture the underlying shape of the data, we assume that the observations are noisy realisations of the solutions to an underlying Differential Equation, $\mathcal{L}$.

This is modelled as

$$X(t) = \sum_{j=1}^{m} c_j u_j(t) + f(t) \tag{1}$$

where $c_j \in \mathbb{R}$, $u_j(t)$ are the m linearly independent solutions to the underlying order m ODE, and $f(t)$ are mean zero Gaussian Processes representing the noise.

Uncovering the Shape Using PDA

Principal Differential Analysis (PDA) fits a linear differential operator

$$\hat{\mathcal{L}} = \hat{\beta}_0(t) + \hat{\beta}_1(t)D + \cdots + \hat{\beta}_{m-1}(t)D^{m-1} + D^m \tag{2}$$

to the observed functional data using a penalised L^2-norm.

This provides an estimate of the true underlying ODE, and allows us to obtain estimates of the functions, $\hat{\mathcal{L}}(f)(t) = \hat{\epsilon}(t)$. These will also be zero mean Gaussian Processes.

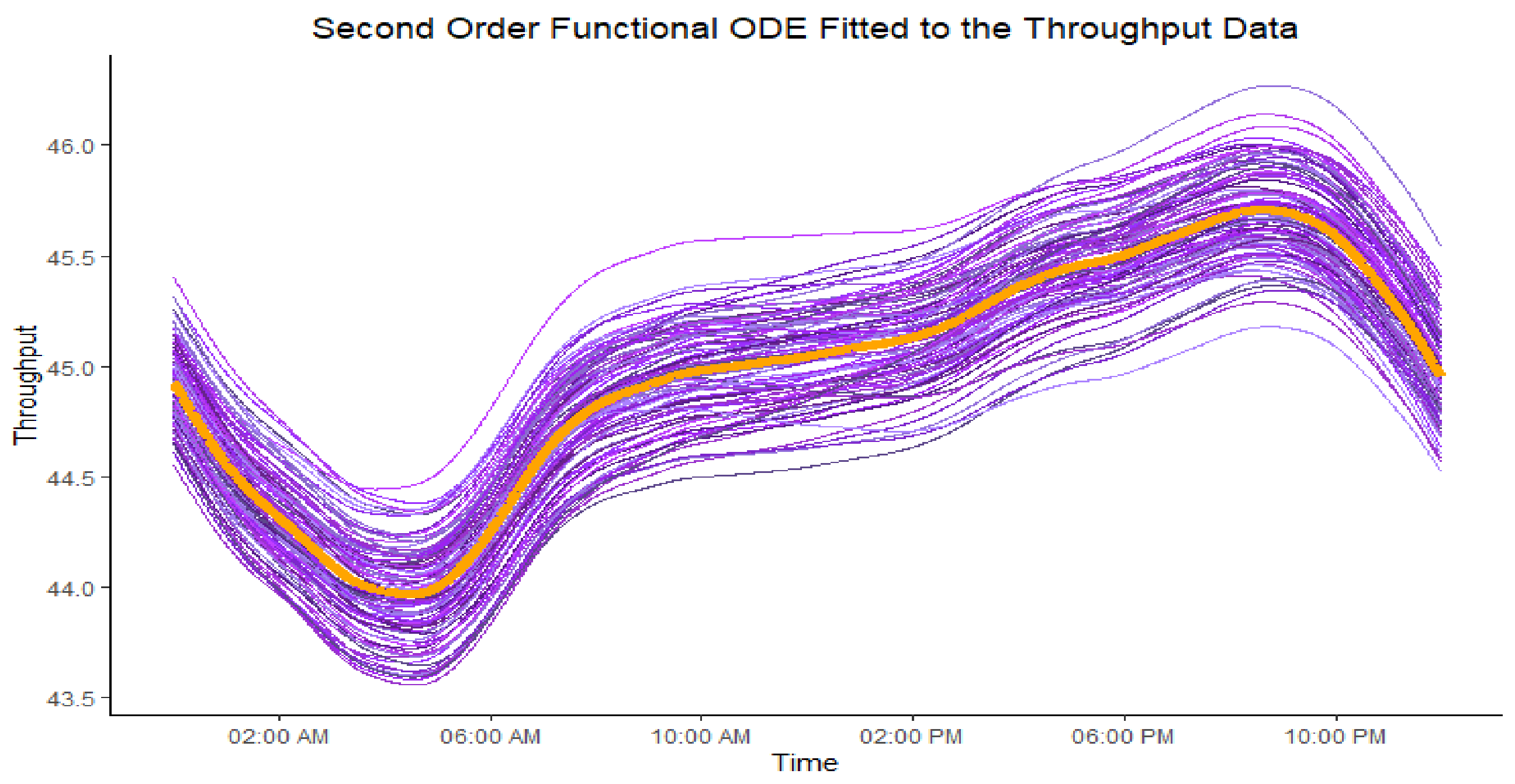

Figure 2: Underlying first order ODE Solution to the throughput data highlighted in orange.

Emergent Phenomena

Emergent phenomena can be thought of as the presence of an additional function, $g(t)$, on some subregion, $\mathcal{S}$, of the observed interval.

This additional function causes a deviation from the shape of the data.

We model the change in the data on some subregion $\mathcal{S}$ as

$$\mathcal{L}(X)(t) = \begin{cases} \epsilon(t) & t \in \mathcal{T} \setminus S \\ \mathcal{L}(g)(t) + \epsilon(t) & t \in S. \end{cases} \tag{3}$$

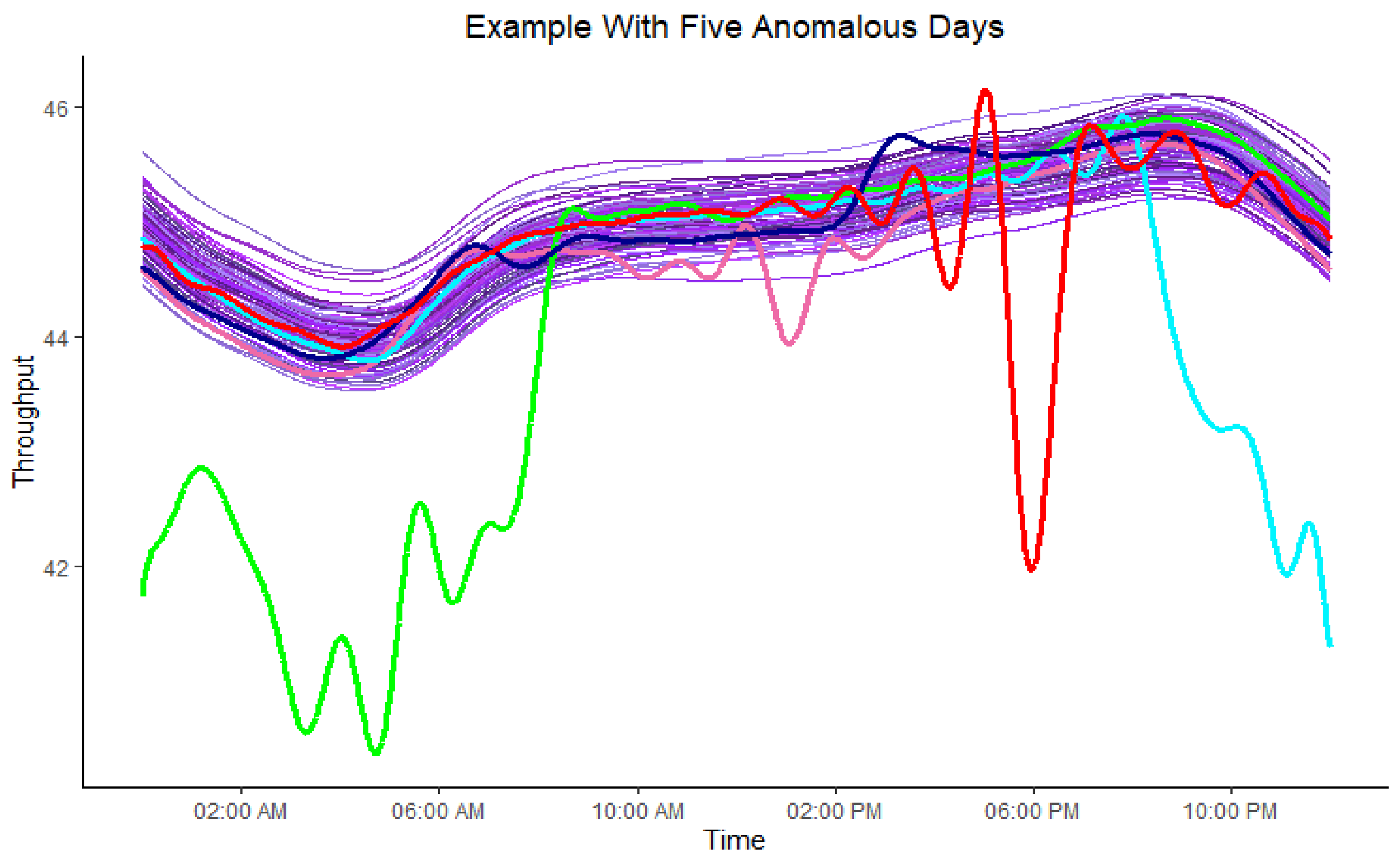

Figure 3: Example with five anomalous days highlighted.

Non-Parametric CUSUM Test

Anomaly detection is performed on the estimated residual functions, $\hat{\epsilon}(t)$.

A Non-Parametric CUSUM test is proposed to detect the presence of $g(t)$.

This uses a test function, $S(k)$, and test statistic

$$\Delta(\tau) = \sum_{k=2}^{\tau} S(k).$$

Given a threshold, γ, this leads to the stopping time

$$\inf \left\{ \tau : \sum_{k=2}^{\tau} S(k) \geq \gamma \right\}. \tag{4}$$

For a new observation, $Y(t)$, the test function used is:

$$S(k) = (\hat{\epsilon}(k) - \hat{\epsilon}(k-1))^2 \qquad 2 \leq k \leq T. \tag{5}$$

We choose γ so that the probability of a false alarm can be controlled.

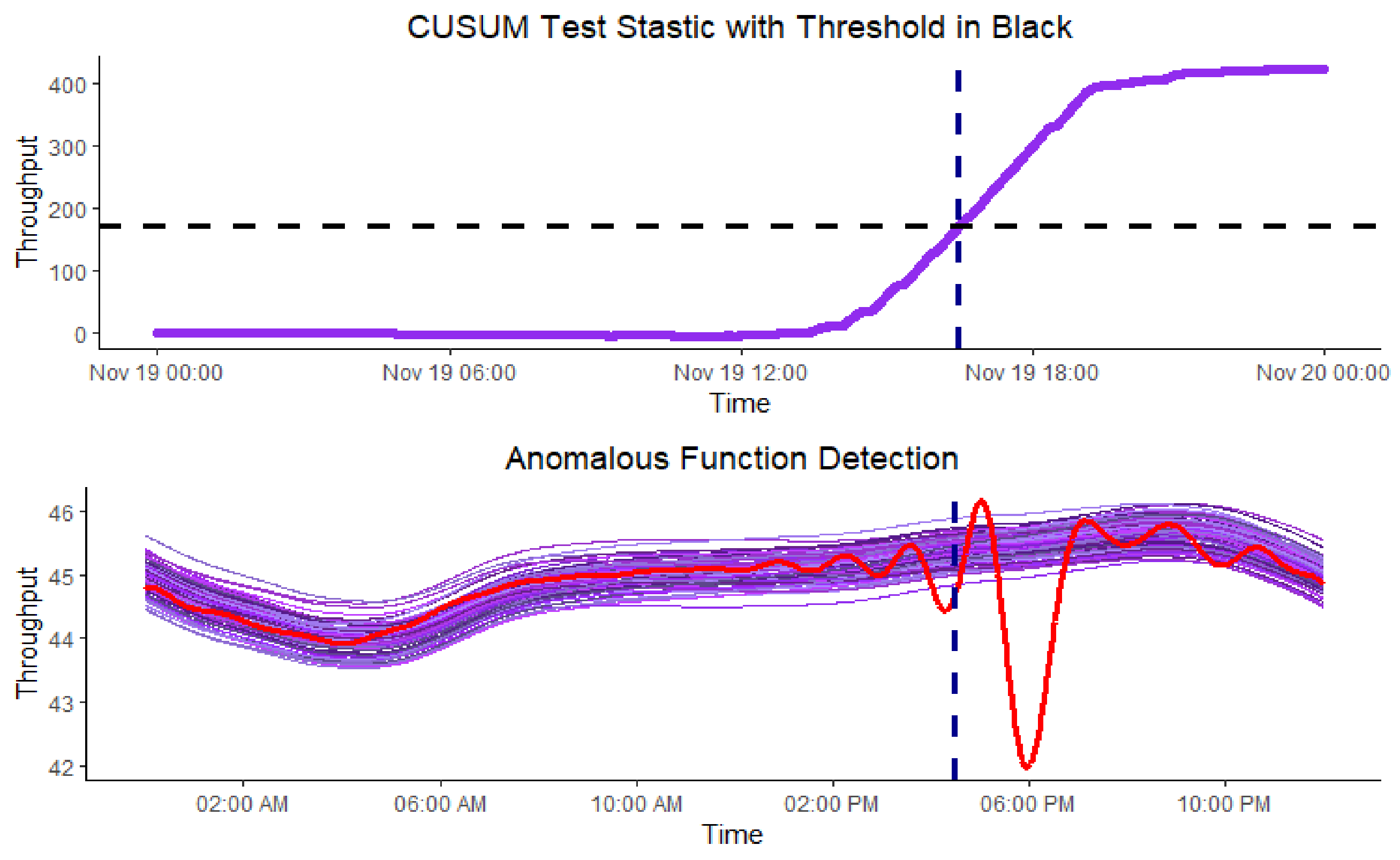

Figure 4: Example of Test Statistic and the associated anomaly detection compared to the non-anomalous underlying data.

Setting the Threshold Using the FCLT

A threshold, γ, can be set using the Functional Central Limit Theorem.

This states that the scaled sum of a centered, unit variance, stochastic process $Z(1), \ldots, Z(T)$ converges to $B(1)$, a standard Brownian Motion.

$$\sum_{k=2}^{T} Z(k) \sim \sqrt{T}\,\mathrm{B}(1) \quad \text{as} \quad T \to \infty$$

We can use a quantile of this to control the probability of a false alarm.

References

Ramsay, J. (2005). Functional Data Analysis. Springer, Dordrecht, 2nd ed.