Detection of data taking anomalies for the ATLAS experiment

Julio DE CASTRO VARGAS FERNANDES[a], Jose MANOEL DE SEIXAS[a], Giovanna LEHMANN MIOTTO[b]

a: Univ. Federal do Rio de Janeiro (BR), b: CERN

(1) Introduction

In this work, we focus on the problem of online detection of anomalies along the data taking period. Anomalies, in this context, are defined as an unexpected behavior of the TDAQ system that result in a loss of data taking efficiency: the causes for those anomalies may come from the TDAQ itself or from external sources.

Motivation :

- Detect anomalies in the ATLAS[1] Trigger and Data Acquisition[2] system (TDAQ).
- Detect and predict the anomalies in an online environment and warn the responsible people.
- Profile and categorize the kind of anomalies.

Proposed methodology:

- Preprocessing.
- A neural network model[3] is used to decide if the current time series state is within the normal operation or it is an anomaly.
- Decision taking.

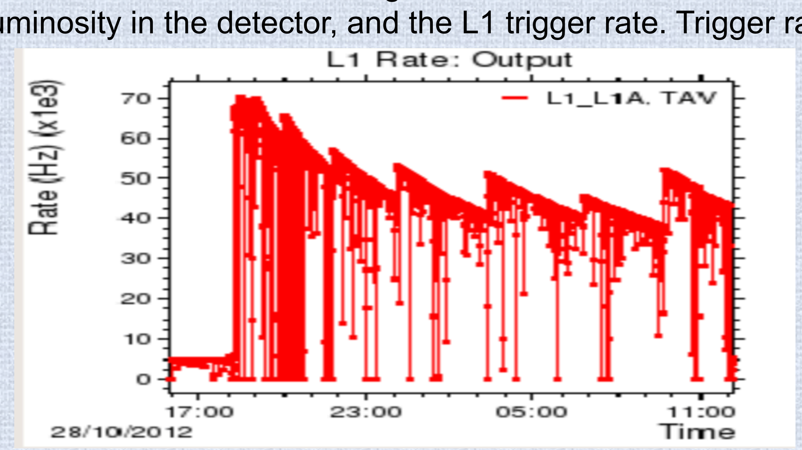
The validity of this approach is demonstrated using a single time series as indicator, the level 1 trigger rate: monitoring data from past physics runs have been used to show that already with a single variable the method is capable of identifying anomalies that had gone unnoticed during data taking.

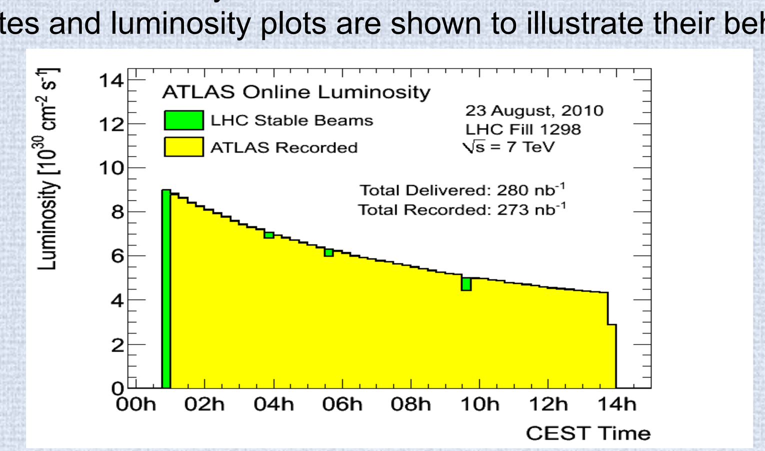
In this work we used a single time series in which all the anomalies were labeled manually and then the neural network was trained to detect them.

(2) Preprocessing

In order to retrieve (and process) the important information we need to remove the fully known components from the time series, since they do not bring any new information. It is important that the processed time series are stationary in the Wide Sense Stationary (WSS).

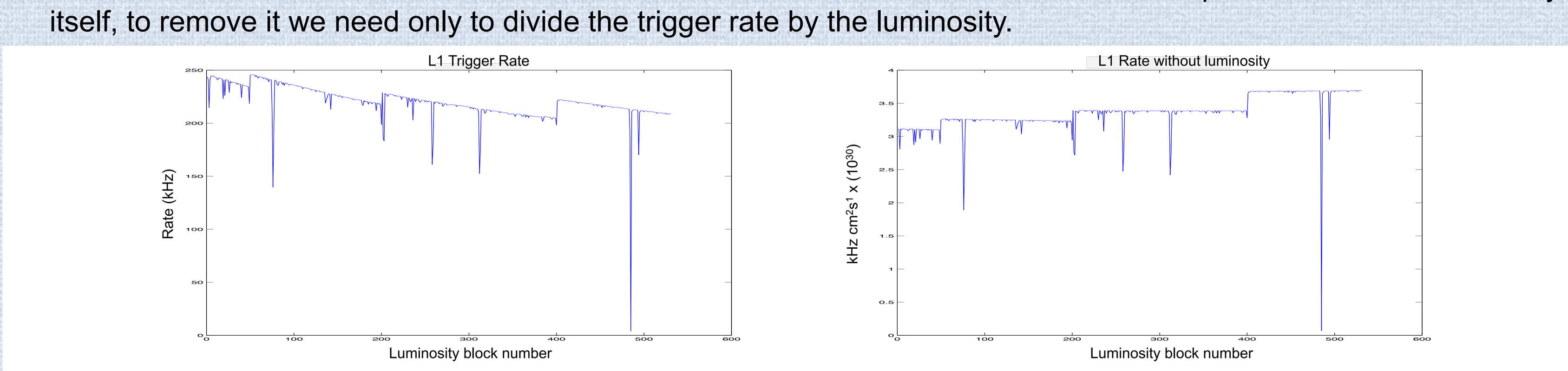
In this work we are monitoring two time series: the LHC collisions luminosity time series, that shows the amount of luminosity in the detector, and the L1 trigger rate. Trigger rates and luminosity plots are shown to illustrate their behavior:



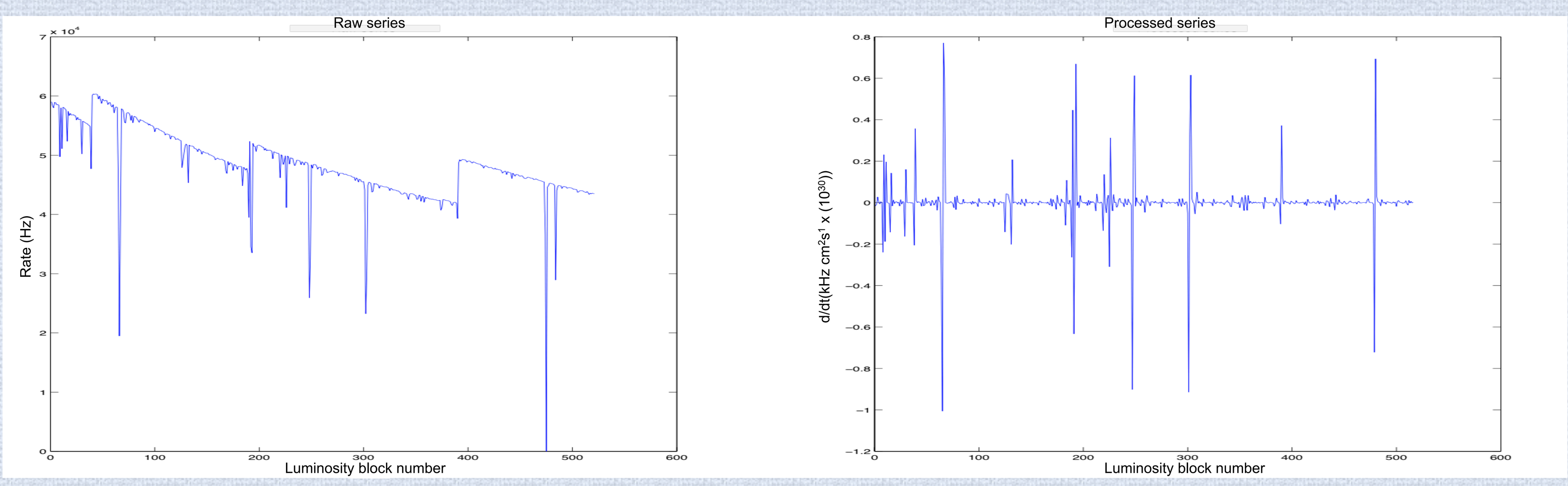


The processing procedure is as follows:

1. Apply the square root function to the rate and the luminosity time series, to reduce the dynamic range of the time series so that the neural networks can treat them more easily.
2. Remove the deterministic trends in the time series, in this case the deterministic trend is presented as the luminosity itself, to remove it we need only to divide the trigger rate by the luminosity.



3. Remove the stochastic trend, by removing all the unit-roots of the time series (just apply the difference operator i.e. the discrete derivative).
4. Remove the critical errors from the trigger rate (since this type of error is always detected we are not interested in it), these are the moments in which the trigger rate goes to zero.



(3) Preprocessing interpretation

The number of events taking place in the detector follow the given formula : N = L * Cs * Eff, where L is luminosity, Cs is cross section and Eff is the efficiency of the detector (this is made up of several different efficiencies) and the rate of events is defined by : R = dN / dt.

Following the steps of preprocessing, namely dividing by the Luminosity and applying the difference operator, we have : $d(R/L)/dt = Cs*d^2Eff/dt^2$.

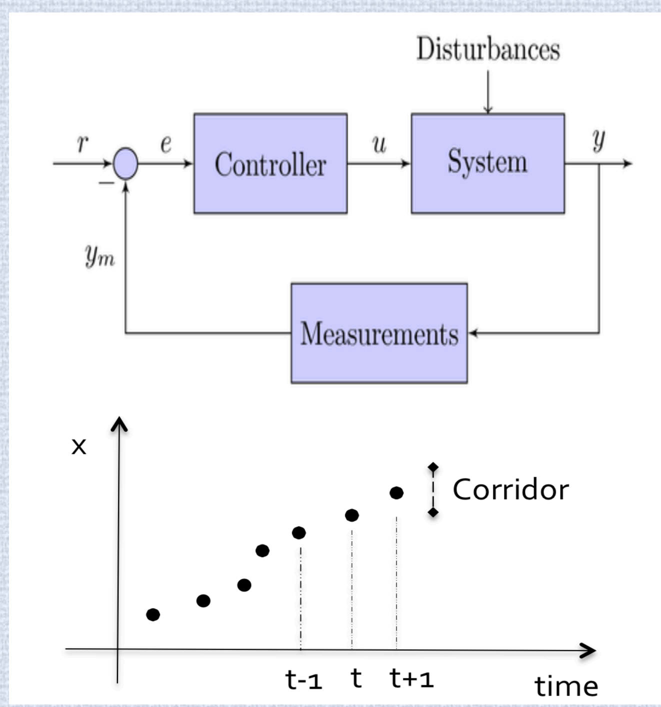
Since the cross-section is constant, the variations that we see in the processed series comes from the efficiency. In a perfect world this would be constant and the result would be 0 (we can see that the baseline is). So apart from the statistics variations in the operation of the detector we also see efficiencies drops and rebounds.

Given this interpretation, some conclusions can be drawn:

- The processed series is composed of two members:
  - The normal DAQ operations (with some statistic fluctuations).
  - The anomalies (errors or problems).
- Problems will appear in the processed series in the form of lost performance, which has two cases:
  - Real anomalies or problems -> The system should detect and warn someone.
  - Deterministic behavior of the detector (rate dropping before prescale, recoveries...) -> The system should detect and ignore.

(4) Data quality approach

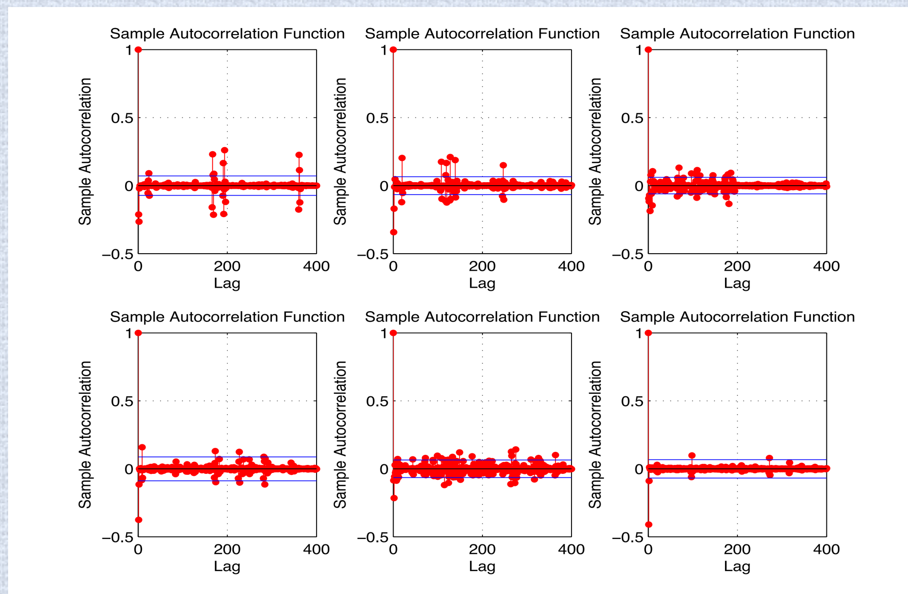
- An anomaly can be seen as a data quality problem in a time series information: with the series modeled, an anomaly can be seen as a model outlier.
- Time series will be monitored in order to evaluate whether an anomaly has occurred or not.
- A validation corridor needs to be constructed in order to define when an outlier occurs. Everything outside this corridor (i.e. outside the model) is an outlier and in the data quality perspective an anomaly.
- In the picture on the right we can see the predicted values by the model and the corridor, that is the uncertainty on the predicted value. If any value is outside this uncertainty area it is then an anomaly.

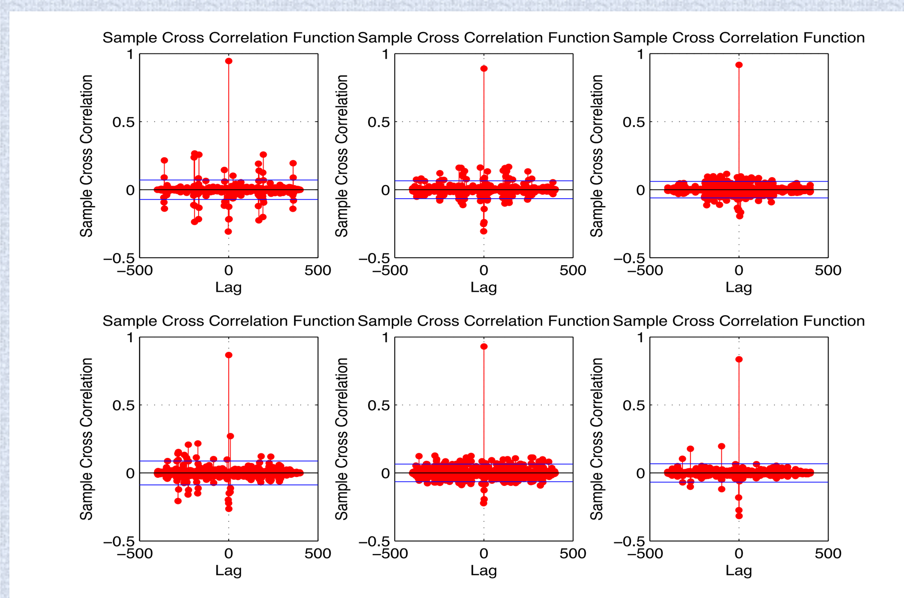


(5) Neural networks

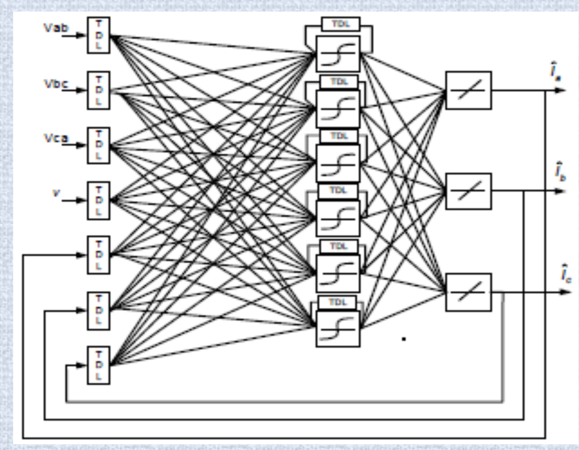
In this work we use neural network models to detect the possible anomalies of the DAQ system. The first thing to do is to decide how many inputs the neural network is going to be fed with. To do that:

- Plot the linear autocorrelation function from the processed time series.
- Plot the nonlinear autocorrelation function from the processed time series.
- From the plots decide how many samples should be used as inputs to the neural network.





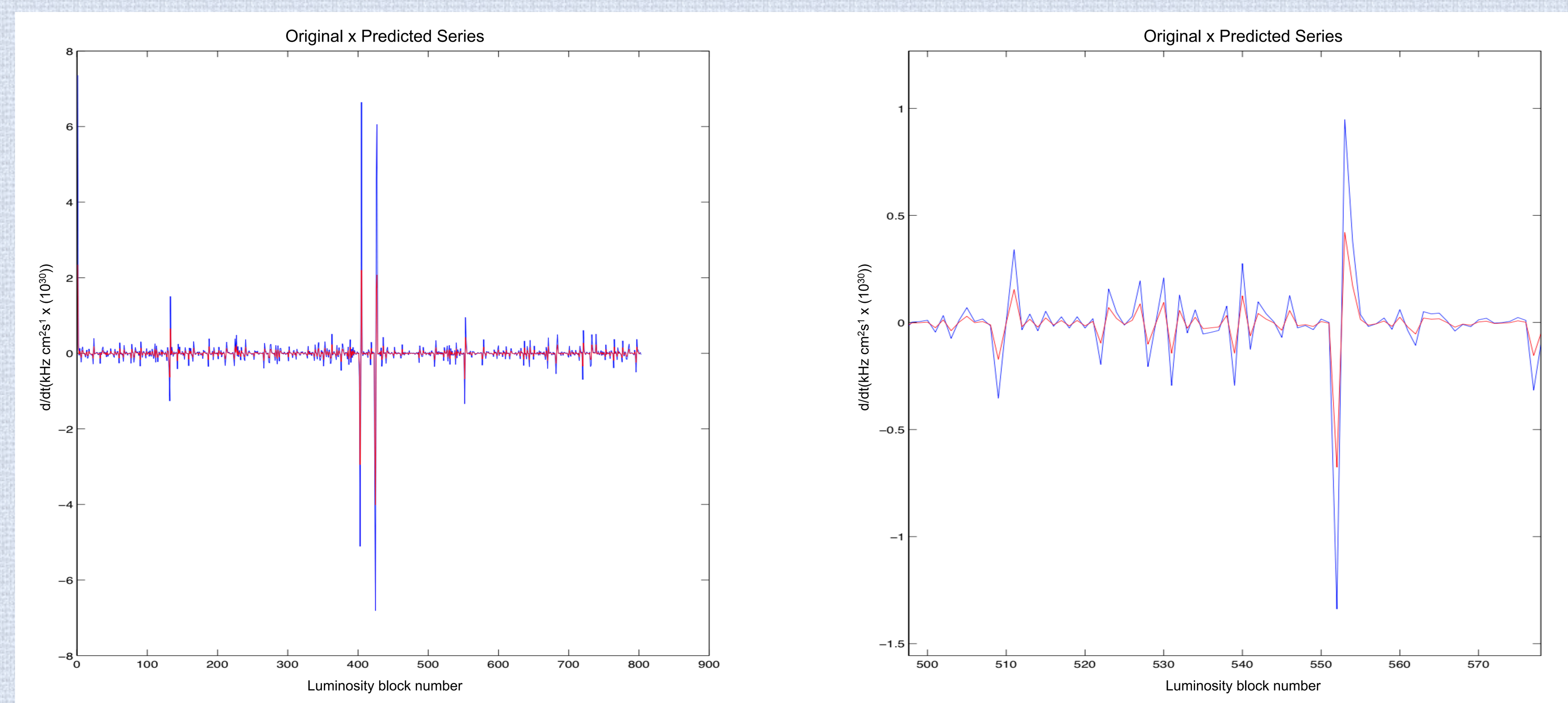
In this work several models of neural networks have been tested and the one that showed the best results are the NARX model.
The architecture of the neural network is presented below:

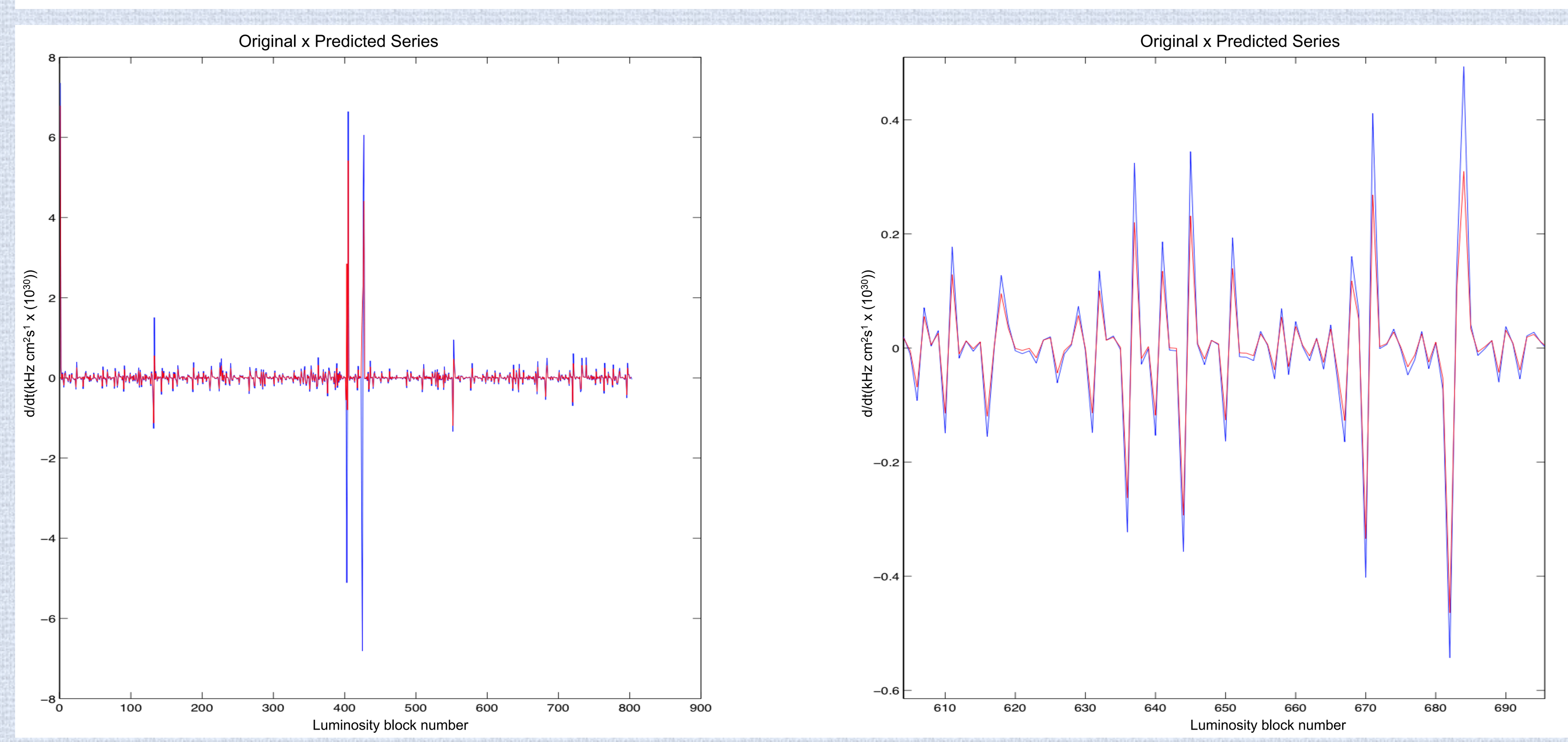


- The NARX neural network predicts the next value of the time series based on the past values and the current value of the time series.
- The NARX network was chosen because it showed: the best performance values (best accuracy for the predict values), the best speed of reaction in the online system and the best dynamic range of all the tested model.
- It is especially suited for online systems.

Below we can see the results for two NARX neural networks, the difference between them is the training criteria, the first one was trained to have a bigger dynamic range and the second one to be more precise. The blue time series is the original and the red time series is the predicted time series.

- On the left the whole time series is shown. Showing the overall prediction.
- On the right there is a zoom of the time series to show the accuracy of the networks.





(6) Conclusions

The neural networks models have proven that they can correctly identify when an anomaly occurs, purely based on trigger data. Thus we can immediately build a system that detects anomalies and predicts their happenings.

In future works we will focus on if we can detect where these anomalies are coming from by monitoring more time series concurrently in an online manner. To do so we will first have to create an anomaly profile, by looking through all the past runs and identifying all the rate drops that do not have a known cause.

(7) References

1. The ATLAS Experiment at the CERN Large Hadron Collider, J. Instrum. 3, 2008, The ATLAS Collaboration.
2. ATLAS high-level trigger, data-acquisition and controls: Technical Design Report, 2002, The ATLAS Collaboration.
3. Neural Networks: A Comprehensive Foundation (2nd Edition), 1998 ,Simon Haykin, McMaster University, Canada.