USE OF DBSCAN FOR FULL-AUTOMATIC-DATA-BASED ANOMALY DETECTION METHOD ON TURN-BY-TURN BEAM POSITION MONITORS (TBT-BPMS) IN SuperKEKB

Q. Bruant*, B. Dalena†, V. Gautard, J. Potaczala
Commissariat à l'Énergie Atomique et aux Énergies Alternatives, Paris, France
M. Le Garrec, CNRS/IN2P3-LAPP, Annecy-le-Vieux, France
J. Keintzel, European Organization for Nuclear Research, Geneva, Switzerland
F. Bugiotti, E. Al Bouzidi, T. Tonin, A. Kahla, CentraleSupélec-LISN, Gif-sur-Yvette, France

Abstract

In order to operate a collider at a consistent peak luminosity, it is essential to possess a comprehensive understanding of the complete magnetic lattice of the colliding rings. This requires precise knowledge of the deviation of the actual lattice from the model used as the basis during the design phase. One of the methods available for the purpose of measuring the aforementioned deviation is Turn-by-Turn Beam Position Monitor (TbT-BPM) surveys. The n-BPM method, as developed at CERN, forms the basis of the spectral response of the TbT-BPMs around the rings. This enables the reconstruction of the full effective magnetic lattice. However, the reliability of this measurement depends on the status and precision of each TbT-BPM in the ring. This paper compares two methods for detecting and removing problematic BPMs from magnetic lattice reconstruction scripts, both using DBSCAN. It describes their assumptions and processes, and discusses the results obtained on the High Energy Ring of SuperKEKB.

INTRODUCTION

The SuperKEKB collider is characterized by its very nonlinear optic lattice and its very high current beams circulating both in a Low Energy Ring (LER) and a High Energy Ring (HER)(see Ref. [1]). This configuration happens to cause a high degree of sensitivity of the machine with respect to actual imperfections in the lattice, compared to the model lattice. In order to discern these variations and identify potential sources, it is possible to use a specific type of Beam Position Monitors (BPMs), designated as TbT-BPMs (which are capable of recording the passing beam for each turn), and employ their respective signal to reconstruct the actual optic lattice of SuperKEKB, such as in Ref. [2]. However, due to the challenging environment and modifications to the rings, some of these TbT-BPMs are prone to malfunction, sometimes only for a few measurements. Consequently, it is imperative to detect these defective BPMs and eliminate their signal from the lattice reconstruction algorithm. This study is motivated by the observation that standard quality cuts applied in the TbT-BPM pipeline (SuperKEKB Optics Measurement Analysis or SOMA (see Ref. [3, 4]) based on processes used for LHC, developed by the Optics Measurement and Correction group (OMC) available at Ref. [5], are not able to remove the entirety of the anomalous BPMs from the analysis. This paper presents two versions of an anomaly detection pipeline based on a clustering algorithm called DBSCAN (see Ref. [6]). The distinguishing factor between these pipelines is the computation of the characteristics of TbT-BPM signals. These characteristics constitute the hyperspace inside which the DBSCAN operates in order to differentiate anomalous BPMs from the functional ones.

* quentin.bruant@cern.ch
† barbara.dalena@cern.ch

METHOD

Datasets

In this paper, we use both simulated and experimental HER data from SuperKEKB. The simulated dataset is obtained by single-particle-SAD tracking with the model lattice parameters of June 2024, with $\beta^* = 0.9$ µm, and taking into account Synchrotron Radiation (SR). Gaussian noise is subsequently applied to all tracks and errors (random spikes in tracks, signal removal keeping noise, random zero-mask on part of tracks, ...) are added to randomly selected BPMs. The experimental dataset consists of several measurements performed in June 2024.

Pipeline Description

After the reading of the raw data, both pipelines (Time2Feat and Fixed Features) pre-process the TbT-BPM signals using a standard scaling method

$$\hat{x}_i^t = \frac{x_i^t - \mu_i}{\sigma_i}, \forall (i,t) \in 1, N_{BPM} \otimes 1, N_{turns} \qquad (1)$$

where $\hat{x}_i^t$ is the preprocessed value, x_i^t is the input, μ_i and σ_i are the mean and the standard deviation of the i^{th} BPM signal over the full range of turns, respectively, N_{turn} is the number of turns recorded in the given measurement (i.e. 4096) and $N_{TbT-BPM}$ is the number of TbT-BPMs considered (i.e. 67 out of few hundred BPMs in total).

The two methods take different paths. The first method is data-centered. The signal of each TbT-BPM is passed to a library called Time2Feat (see Ref. [7]) which computes several statistical features related to these signals. The most meaningful features are then selected. This selection is based on a process called Principal Feature Analysis (PFA) (see Ref. [8]). The table of all the features computed and their respective values for all the TbT-BPMs is passed to this

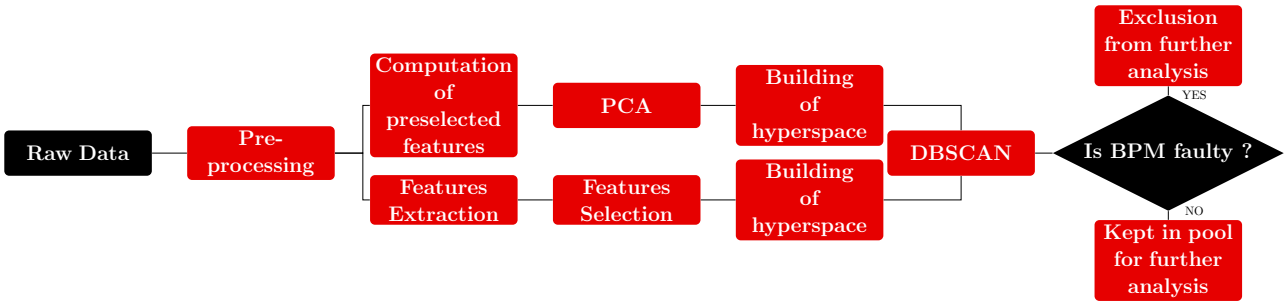

Figure 1: Workflow chart of the two methods in parallel.

process, where a Principal Component Analysis (PCA) is performed to identify the main components of the signal recorded in the features. Following the PCA, a clustering algorithm called K-Means is performed to group the different computed features into the main components obtained via the PCA. Finally, for each group created, the feature exhibiting the closest proximity to the respective main component is selected in order to build the hyperspace. DBSCAN is finally applied to this hyperspace to detect the main concentration of TbT-BPMs, which are assumed to be functioning, and the outliers, which correspond to the anomalous ones. The strength of this approach lies in not requiring neither training nor prior knowledge of the system. It can identify outliers from limited measurements, which is challenging in the context of BPMs, where identifying outliers often requires multiple measurements from each BPM. However, this process also has disadvantages; the computing time needed to evaluate the numerous features of the library (more than 200) scales quickly with the volume of input data. Furthermore, these features are purely statistical, with no physical knowledge of the system.

The second path is the opposite of the first in terms of physics and data-centric characteristics. In this method, the physics-informed features are selected prior to the data 'discovery' by the algorithm. These features are then computed for all the measurements and BPM signals. These are:

- Betatron Tune ν: frequency of the main spectral peak,
- Betatron Amplitude A: the amplitude at ν,
- Noise-to-amplitude ratio (NAR): standard deviation of the SVD residual (SVD cut at 60 % variance) normalized by A,
- RMS spectral amplitude: RMS of FFT bins after removing all strong peaks ($\approx$ 10 % of A),
- Variance of the ICA (see Ref. [9]) residual,
- Harmonic ratio (HR): ratio of the spectral energy in the harmonic band $[1.8\nu, 3.2\nu]$ to the energy at the fundamental

$$HR = \frac{\sum_{1.8\nu<f<3.2\nu} a(f)^2}{a(\nu)^2} \quad (2)$$

PCA is subsequently applied to this table of physics-informed features, and the hyperspace is constructed directly with the resulting principal components. DBSCAN is applied to this hyperspace as described in the first method.

A more graphical representation of these two pipelines is displayed in Fig. 1.

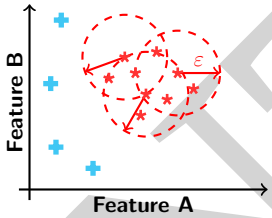

Figure 2: Scheme of DBSCAN process.

ε Tuning

To ensure optimal functionality within the context of our system, it is imperative to calibrate the DBSCAN hyperparameter, denoted by ε, to align with our specific operational environment. This hyperparameter sets the Euclidean distance threshold in hyperspace for a BPM to no longer be part of the main distribution, as illustrated in Fig. 2. To achieve this optimization, a simulation of the SuperKEKB HER is performed, using a specific excitation amplitude that is comparable to the amplitude that is anticipated to be observed in the actual experimental data. In the Fixed-features method, slight variations in the arrangement of BPMs in hyperspace can shift the optimum epsilon. This phenomenon is accentuated in the Time2Feat method due to possible modifications in the selected features, generating a different hyperspace for DBSCAN to operate on.

Following multiple attempts to implement the Time2Feat method with a perfect simulation, the detection of at least one BPM of the Interaction Region (IR-BPM) was quasi-constant. This may be due to the IR-BPM's precise location within the ring, where magnetic focusing is maximum, and several optical functions behave differently than elsewhere, resulting in significant disparities in beam dynamics. To account for this possibility, the same tuning was performed, excluding the four IR-BPMs. The results of this tuning are displayed on Fig. 3 for both methods on the simulated dataset with two beam excitations. In view of the results of the tuning process, it has been determined that two distinct values of the ε parameter must be considered for both methods, in order to encompass the various excitation cases that may arise when testing on real experimental data. The Time2Feat model is to be tested using the following values of ε: 7.50 and 8.25. Conversely, the Fixed-Features model

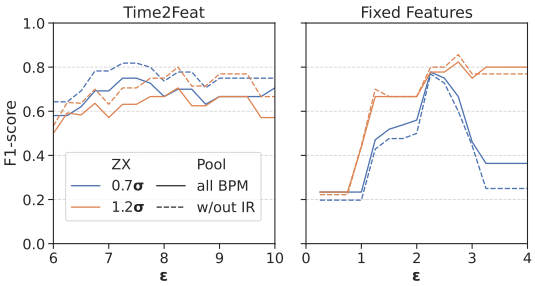

Figure 3: Evolution of the F1-score w.r.t ε for two different level of beam excitation and for both methods and both considering or removing the IR-BPMs in the simulated HER.

is to be tested with two alternative values of the ε parameter, namely 2.25 and 2.75. These values of ε correspond to optimal values for the F1 score when the beam is excited to 0.7 and 1.2 times the beam size, respectively.

RESULTS

The results of the tests conducted using the aforementioned process are presented in Table 1 and 2 . The findings presented herein are derived from experimental observations conducted in June of 2024. We consider 8 measurements during which the beam has been excited by a vertical kicker with a tension ranging from 0 to 1.3 kV. The result of the detection presented here is the union of the detections for each measurement. This detection is confronted with a list of probable anomalous BPMs provided by the SuperKEKB operating team.

The list of detected BPMs demonstrates a high degree of stability when transitioning from low to high excitation cases. This phenomenon is particularly evident in the Time2Feat method, where the detected BPMs are identical. Furthermore, the Fixed-Features method tends to detect more BPMs, at the price of lowering Precision, while maintaining the same Recall as Time2Feat. This resulted in a lower global F1-score. Furthermore, it is worth noticing that a number of BPMs not included in the provided list have been repeatedly identified as anomalous by at least one of the two methodologies. For instance, the BPM labeled MQEAE25 is present for both methodologies. Conversely, the BPM labeled MQD3E31 is consistently detected by the Time2Feat method, but not by the other. Following a visual inspection of the respective time series in both transverse planes (see Fig. 4), it was determined that the BPMs in question were defective. MQEAE25 exhibits a pronounced global amplitude, surpassing that of the other BPMs, accompanied by random spikes in the signal, suggesting a BPM with high conversion electronics that is potentially blind. In contrast, MQD3E31 displays a substantial offset, exceeding the limits observed otherwise in SuperKEKB. The metrics have been updated to include the two BPMs added to the Anomalous-BPMs list, and these can be found between parentheses. It can be observed that with the incorporation of these BPMs into the provided list, the metrics (F1-score, in particular) values appear to be similar to the values observed in the simulations during the tuning process. This behavior provides substantial evidence that these BPMs are defective and can be consistently added to the list of anomalous BPMs, which is therefore incomplete.

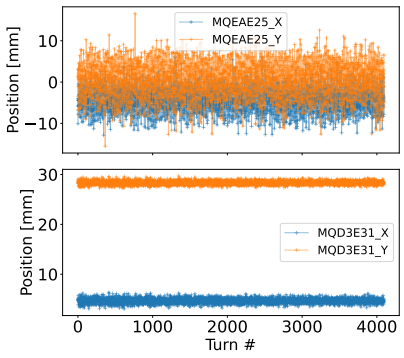

Figure 4: Plot of the signal of MQEAE25 and MQD3E31 in one of the tracks used for testing.

Table 1: Results for each method for ε optimized for low excitation. The bold names correspond to BPMs appearing in the anomalous-BPMs list provided by SuperKEKB team. The bottom line correspond to the value of the Precision/Recall/F1-score. The values in parentheses correspond to the case where MQEAE25 and MQD3E31 are added to the anomalous-BPMs list provided.

Time2feat	**Fixed-features**
MQD3E8, **MQEAE20**, **MQD3E18**, **MQEAE35**, **MQR2ORE**, MQEAE25, +4	**MQD3E8**, **MQEAE20**, **MQD3E18**, **MQEAE35**, **MQR2ORE**, MQEAE25,+16
0.50 (0.70) / 0.83 (0.88) / 0.63 (0.78)	0.23 (0.32) / 0.83 (0.88) / 0.36 (0.47)

Table 2: Results for each method for ε optimized for high excitation.

Time2feat	**Fixed-features**
MQD3E8, **MQEAE20**, **MQD3E18**, MQEAE25, **MQEAE35**, MQD3E31, +3	**MQD3E8**, **MQEAE20**, **MQD3E18**, MQEAE25, **MQEAE35**, **MQR2ORE**, +7
0.56 (0.78) / 0.83 (0.88) / 0.67 (0.82)	0.38 (0.46) / 0.83 (0.75) / 0.53 (0.57)

CONCLUSION

In this paper, two DBSCAN-based methods were compared into detecting anomalous BPMs in the SuperKEKB HER. The findings reveal that both methods demonstrate a F1-score in excess of 50 % for the high excitation tuning. Although the Fixed-Feature approach demonstrates greater efficiency, the Time2Feat method displays superior overall performance, particularly with regard to precision. Furthermore, several anomalous BPMs were identified that were not included in the procured list. These results necessitate the extension of this study to both rings in SuperKEKB and with larger data volumes, in order to identify other anomalous BPMs and refine the knowledge of their optics functions.

REFERENCES

[1] Y. Ohnishi *et al.*, "Accelerator design at superkekb", *Prog. Theor. Exp. Phys.*, vol. 2013, no. 3, 03A011, Mar. 2013. doi:10.1093/ptep/pts083

[2] A. Langner and R. Tomás, "Optics measurement algorithms and error analysis for the proton energy frontier", *Phys. Rev. ST Accel. Beams*, vol. 18, no. 3, p. 031002, Mar. 2015. doi:10.1103/PhysRevSTAB.18.031002

[3] J. Keintzel *et al.*, "SuperKEKB Optics Measurements Using Turn-by-Turn Beam Position Data", in *Proc. IPAC'21*, Campinas, Brazil, May 2021, pp. 1352–1355. doi:10.18429/JACoW-IPAC2021-TUPAB009

[4] J. Keintzel, "SuperKEKB Optics Measurement Analysis (SOMA)", https://github.com/JacquelineKeintzel/SOMA,

[5] OMC-Team *et al.*, Omc3, 2022. doi:10.5281/zenodo.5705625

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", in *KDD*, pp. 226–231, 1996. http://dblp.uni-trier.de/db/conf/kdd/kdd96.html#EsterKSX96

[7] A. Bonifati, F. D. Buono, F. Guerra, and D. Tiano, "Time2feat: learning interpretable representations for multivariate time series clustering", *Proc. VLDB Endow.*, vol. 16, no. 2, pp. 193–201, Oct. 2022. doi:10.14778/3565816.3565822

[8] Y. Lu, I. Cohen, X. S. Zhou, and Q. Tian, "Feature selection using principal feature analysis", in *Proc. Int. Conf. Multimedia*, pp. 301–304, Sep. 2007. doi:10.1145/1291233.1291297

[9] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications", *Neural Networks*, vol. 13, no. 4, pp. 411–430, Jun. 2000. doi:10.1016/S0893-6080(00)00026-5