



This is a repository copy of *Machining centre performance monitoring with calibrated artefact probing*.

White Rose Research Online URL for this paper:
<https://eprints.whiterose.ac.uk/166193/>

Version: Published Version

Article:

Rooker, T., Stammers, J., Worden, K. orcid.org/0000-0002-1035-238X et al. (3 more authors) (2021) Machining centre performance monitoring with calibrated artefact probing. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 235 (10). pp. 1569-1587. ISSN 0954-4054

<https://doi.org/10.1177/0954405420954728>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:
<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Machining centre performance monitoring with calibrated artefact probing

Tim Rooker^{1,3}, Jon Stammers², Keith Worden³, Graeme Potts⁴, Kevin Kerrigan² and Nikolaos Dervilis³

Proc IMechE Part B:

J Engineering Manufacture

1–19

© IMechE 2020



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0954405420954728

journals.sagepub.com/home/pib

Abstract

Maintaining high levels of geometric accuracy in five-axis machining centres is of critical importance to many industries and applications. Numerous methods for error identification have been developed in both the academic and industrial fields; one commonly-applied technique is artefact probing, which can reveal inherent system errors at minimal cost and does not require high skill levels to perform. The primary focus of popular commercial solutions is on confirming machine capability to produce accurate workpieces, with the potential for short-term trend analysis and fault diagnosis through interpretation of the results by an experienced user. This paper considers expanding the artefact probing method into a performance monitoring system, benefitting both the onsite Maintenance Engineer and visiting specialist Engineer with accessibility of information and more effective means to form insight. A technique for constructing a data-driven tolerance threshold is introduced, describing the normal operating condition and helping protect against unwarranted settings induced by human error. A multifunctional graphical element is developed to present the data trends with tolerance threshold integration to maintain relevant performance context, and an automated event detector to highlight areas of interest or concern. The methods were developed on a simulated, demonstration dataset; then applied without modification to three case studies on data acquired from currently operating industrial machining centres to verify the methods. The data-driven tolerance threshold and event detector methods were shown to be effective at their respective tasks, and the merits of the multifunctional graphical display are presented and discussed.

Keywords

Computer numerical control (CNC), five-axis machining, performance monitoring, artefact probing, Gaussian process, support vector machines

Date received: 9 July 2019; accepted: 18 July 2020

Introduction

To remain competitive in the modern advanced manufacturing sector, it is becoming increasingly important to embrace and implement intelligent systems for process monitoring. The core capability for data and analytics strived for in the Industry 4.0 movement¹ is a key motivator for this change. Information transparency, concerning the application of on-line data collection to facilitate analysis of the physical system, is a particularly attractive topic for research and development, and applying machine learning paradigms to physical engineering processes presents itself as an effective solution to the problem. The resultant output should then be informative in providing technical assistance to the Manufacturing or Maintenance Engineer (ME), fulfilling two of the four design principles set out for Industry 4.0.²

It is not possible to consistently manufacture with absolute accuracy and precision; there will always be some observable, quantifiable degree of error present on a manufactured workpiece, as compared with its idealised specification. However, communication through the Geometric Dimensioning and Tolerancing

¹Industrial Doctorate Centre in Machining Science, University of Sheffield, Sheffield, UK

²Advanced Manufacturing Research Centre, Rotherham, South Yorkshire, UK

³Dynamics Research Group, University of Sheffield, Sheffield, UK

⁴metrology software products Ltd., Alnwick, UK

Corresponding author:

Tim Rooker, Industrial Doctorate Centre in Machining Science/Dynamics Research Group, University of Sheffield, Western Bank, Sheffield S1 3JD, UK.

Email: tjrooker1@sheffield.ac.uk

system³ allows realistic manufacturing of any reasonable engineering design. Variation in quasi-static and dynamic error sources⁴ that affect machining accuracy arise due to local temperature fluctuations, in-process conditions, significant events (such as a tool crash, or calibration activity), errors in the size and form of machining centre components as well as general wear of moving elements throughout normal operation. The extent, effect and complexity of compounding error is unique to any particular machining centre and its life cycle, leading to the possibility that two otherwise identical systems may exhibit significant performance differences in their production output. It follows that the performance of a unique system will tend to drift over time,⁵ as its compound error profile is affected through further operational use. A requirement for repeatable performance in the manufacturing process has led to the development of error identification methods for informing machine calibration cycles, or qualification to operate by assessment of machine capability.

Accordingly, much research in recent years has been dedicated to the accurate identification of the parameters which mathematically define the kinematic error motions, with particular interest in the variable location errors, to inform necessary maintenance actions such as calibration or specific repairs. A prominent example of this is the R-Test,⁶ which is currently recognised as an ISO standard⁷ and has been shown to be effective for multi-axis machine tool calibration.⁸ The original method used three linear displacement sensors to dynamically track the location of a calibrated, spherical artefact through ranges of motion in the rotary axes. An example of recent commercialisation⁹ of the technique employs a non-contact 3-D probe based on eddy currents to the same effect. The procedure results in a detailed error map which can be used to diagnose machine faults (kinematic eccentricity, misaligned axes, backlash etc.) and inform a calibration action. Static variants of the R-Test, where the procedure and equipment setup is replicated but data are collected at discrete intervals, have been applied for complete error map construction of all location errors and the larger class of position-dependent geometric errors.¹⁰

The R-Test is a fast⁹ and reliable method for kinematic error identification in machine tool rotary axes, but it has the drawback of requiring specialist equipment to conduct. A common alternative is to use a touch-trigger probe to inspect an artefact of precisely known dimension, in a procedurally similar manner to the static R-Test. *Calibrated artefact probing* is attractive to many in the industry due to the ubiquitous nature of the touch-trigger probe, used in virtually all modern facilities for pre-machining verification, fault diagnosis and calibration. The main trade-off for this minimised cost is the overall speed of the procedure, as multiple contact points on the artefact must be probed to precisely determine its location, as compared with

the R-Test which requires only one contact per rotary-axis position. An approach involving probing a rectangular artefact across numerous probing patterns¹¹ has been shown to be effective for rotary-axis location error calibration. This work was then extended to construct a complete error map by artefact probing, to much the same effect as the static R-Test publication noted above.¹² Spherical artefacts are often favourable, due to their nominally-identical form when approached from different angles. The artefact is probed numerous times and spherical interpolation applied to determine its true centre point, resulting in an informative error map which can be applied in much the same way as its dynamic counterpart. The *scale and master balls artefact* method¹³ employs touch-trigger probing to locate a collection of precision balls at various axis positions, and has been shown to provide sufficient data to estimate all axis-to-axis location errors in a five-axis machining centre. More recently, the method has been applied for kinematic fault diagnosis.¹⁴

The core focus with artefact probing procedures and commercial software is error identification for machine capability checking or calibration activities. Generally speaking, performance monitoring and fault diagnosis in machine tools are of interest to the community; recent research has considered the mining of general in-process data across networks of machining centres,¹⁵ analysing the energy usage to detect abnormalities¹⁶ or vibration response to assess the health of the axis drives.¹⁷ Monitoring machine performance with artefact probing data is also possible as a secondary application; however, it is often overlooked, and as such, effective systems for interrogating legacy data have not been developed. Such a system could be of great benefit to both the onsite ME and visiting specialist consultants, enabling them to quickly and easily interpret legacy data and extract critical insight on the machine's history, usage and signature. Developing a performance monitoring system like this also presents the opportunity to employ novel predictive modelling techniques, introducing intelligent features and automation to further assist the user.

The paper is structured as follows. Firstly, the experimental procedure for artefact probing is described, as well as the data post-processing and presentation currently employed for performance monitoring purposes within a popular commercial solution.¹⁸ The analytical methods proposed for the monitoring system are then presented; including a tutorial for interrogating the results on a simulated, demonstration dataset, and a visual comparison of the kernel options to select the most appropriate. Then, three case studies are evaluated, applying the methods to real datasets acquired from the manufacturing industry. Finally, concluding remarks and a discussion on the proposed system are presented, with consideration of further research on the topic.

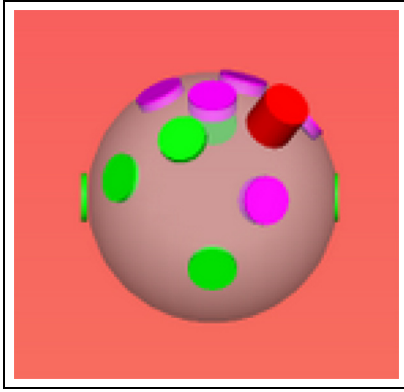


Figure 1. Visualisation of points probed on the sphere surface for a typical indexation.¹⁸ The height of the extrusions indicate the extent of the error identified at each point on the surface.

Experimental procedure

The probing procedure, which this paper aims to develop, involves locating a single, spherical artefact at numerous indexations of the rotary axes. Generally speaking, the Primary-axis of a machine tool is defined as the C-axis, and the Secondary-axis is either the A- or the B-axis, dependent upon the specific configuration. For this paper, a B-C machine tool configuration will be considered. The method has a number of similarities with a recent paper,¹⁹ which proposed a Least-Squares Estimation approach to identify and calibrate rotary-axis location errors. A calibrated, spherical artefact is firstly located at the ‘home’ position, where the rotary axes are indexed at $B=0^\circ$, $C=0^\circ$. This is determined by probing a number of points on the sphere’s surface, as illustrated in Figure 1, and utilising the measurements to determine the position of the sphere centre.

At the home position, it is possible to run a number of checks to quantify the performance of the probe itself. This includes assessing the axial pre-travel (the distance traveled before the probe is triggered), the repeatability of the automatic tool change action or characterising the overall performance. The artefact is then reoriented relative to the probe, incrementally, along specified arcs of rotation to determine the performance of the rotary-axes.

Figure 2(a) shows the typical setup, with the rotary axes indexed at the home position; Figure 3 shows this in a diagrammatic format. The Primary-axis procedure described in this paper locates the sphere at thirteen indexed positions from $C=0^\circ$ to $C=360^\circ$. For the Secondary-axis procedures, there are seven positions indexed from $B=0^\circ$ to $B=90^\circ$ for a positive (+ve) rotation, and from $B=0^\circ$ to $B=-90^\circ$ for a negative (-ve) rotation. Figure 2(b) shows the sphere being probed at in indexation of the Secondary-axis. As the nominal kinematics of the system are known, it is straightforward to calculate the residuals between

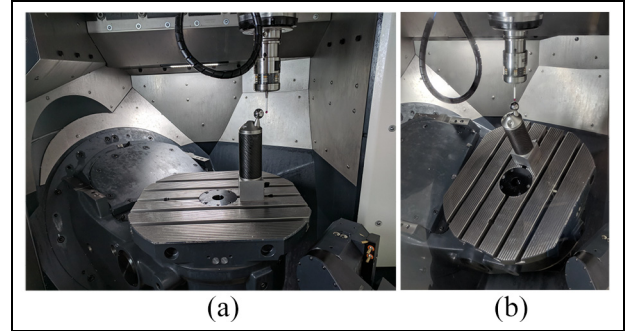


Figure 2. Hardware setup for performing the artefact probing procedure: (a) Probing the sphere at the home position and (b) Probing the sphere at an indexation of the Secondary-axis.

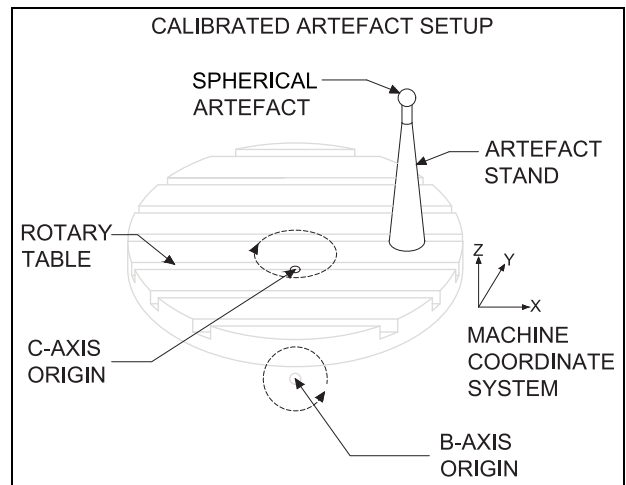


Figure 3. Diagram of typical spherical artefact setup on a B-C configured rotary table.

nominal and actual for each sphere location, to produce the volumetric error map across the full axis index range. The procedure is illustrated in Figure 4(a) and (b).

The intrinsic focus in machine capability is to ensure satisfactory part production in the short-term, implementing a *green light manufacturing* policy where operators can confidently and efficiently run machines without the need for high levels of additional, specialist skills. Proprietary software¹⁸ achieves this by collating the data from a calibrated artefact probing procedure into a summary presentation – referred to as the *benchmark* performance and illustrated in Figure 5 – and detailed reports for deeper analysis and fault diagnosis. Each spoke on the benchmark wheel represents a different test conducted by the software, such as *Rotary Single Primary* (a measure of the Primary-axis performance) or the *Overall probe performance*.

A foundational characteristic with the benchmark performance presentation is its ability to provide an instant appraisal of the system’s health state, which is accessible to personnel of all skill levels. In 1983, *The*

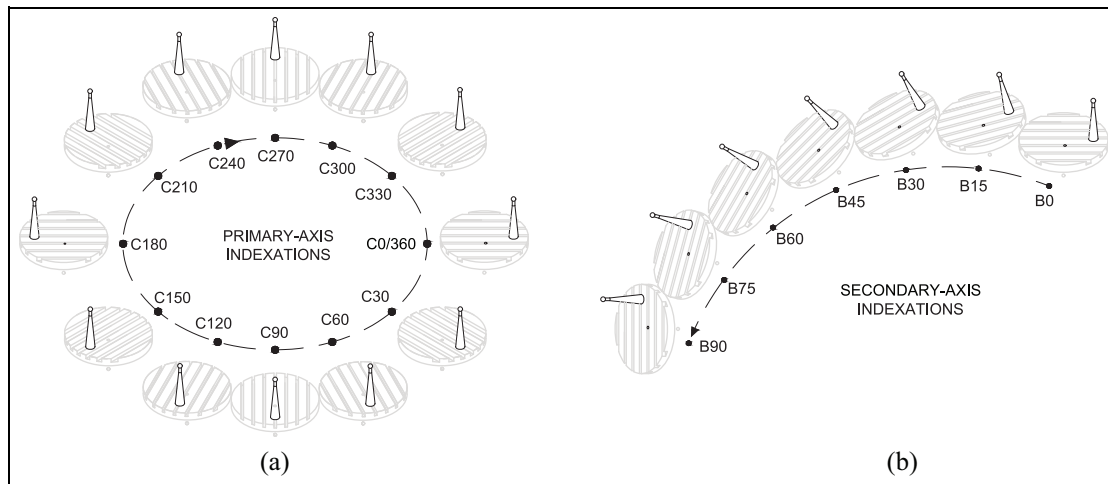


Figure 4. Indexations of the primary- and secondary-axes for the rotary-axis error identification procedures: (a) Typical rotary-axis positions for the Primary-axis procedure and (b) Typical rotary-axis positions for the Secondary-axis (+ve) procedure. The Secondary-axis (-ve) procedure is simply these indexations, but in reverse.

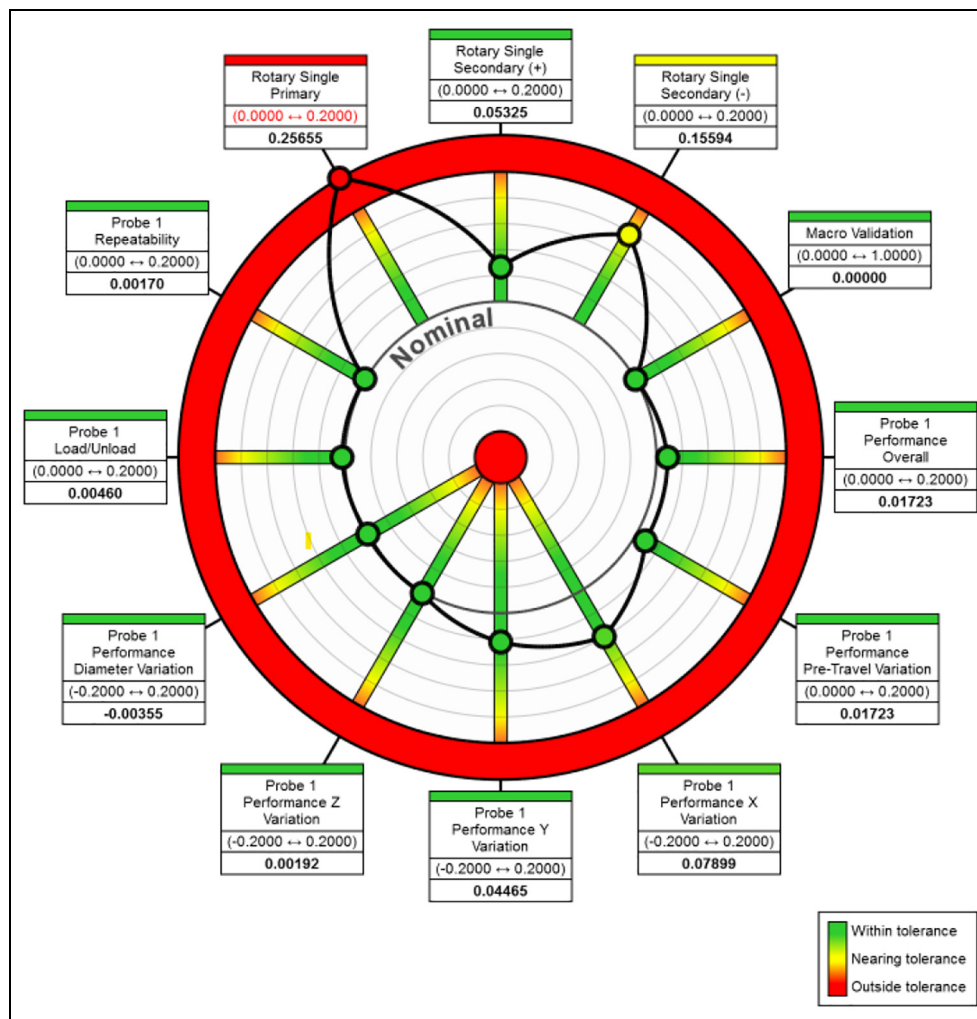


Figure 5. Benchmark performance summary, as produced by proprietary artefact probing software.¹⁸

*Visual Display of Quantitative Information*²⁰ by Edward Tufte was published, concerning the theory and practice of data graphics design. The benchmark performance presentation echoes many of the design practices set out by Tufte. The heavier line-weight afforded to the data elements helps establish contrast in meaning, as compared with the polar axis elements that have lesser graphical importance. Presented in this way, the data elements also serve as a multifunctional graphical element; every machining centre has its own unique *signature*, which is represented by the respective positions of each data element during normal operation. The signature is an important visual element for verifying that a group of machines producing the same part are likely to perform to similar levels. The inclusion of colour can often risk generating a graphical puzzle, in which the viewer is required to consciously decode the information presented before them. The benchmark presentation leverages colour effectively by implementing a red-yellow-green scheme, instantly recognisable due to the ubiquitous traffic light system in modern society.

Analytical methods

Learnable tolerance thresholds

A key decision-making process in capability checking concerns the proper setting of tolerance thresholds, which define the acceptable levels for each feature that characterise the machine's error state. Thresholds are determined by the ME as a result of satisfactory feedback from the quality assurance process, generally depending on the production requirements and the industry for which they are intended. Incorporating normalisation with respect to the tolerance threshold is crucial for assigning meaning to the absolute error measurements obtained by the probing procedure. Reassignment of a machining centre to new production requirements may result in alterations to the tolerance thresholds. There is also the possibility of user error, or even intentional tampering, when setting new tolerance thresholds; this is to be avoided at all costs, and is an important motivator for incorporating learnable tolerance thresholds into any monitoring system with a foundation in geometric accuracy. Moreover, learning the most appropriate tolerance threshold from historical data is highly useful for representing the normal operating condition of any unique machining centre. This representation extends the potential for signature comparison, currently utilised in the benchmark capability checking system, to evaluate historical performance of a group of machines with respect to one-another.

In supervised learning, the core objective of a classifier is to produce a decision boundary separating two or more predetermined classes of data. It is likely that there exist numerous potential solutions for an appropriate decision boundary, so it would be wise to select the one which presents the minimal generalisation

error. A Support Vector Machine (SVM) classifier attempts this by solving for the decision boundary which maximises the margin between the classes of data.²¹ The SVM is a kernel method based on a sparse solution, such that new input predictions depend only on a relevant subset of the training data, known as the *support vectors*. The margin is defined as the smallest distance between the decision boundary and designated support vectors. Conveniently, the identification of model parameters corresponds to a convex optimisation problem, so any local solution can be considered a global optimum.²²

For a linearly-separable case, defined by kernel function, $\phi(\mathbf{x})$, with weights, \mathbf{w} , and bias parameter, b , of the form,

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \quad (1)$$

the optimum decision boundary is determined by applying the constraint,

$$t_n(\mathbf{w}^T \phi(\mathbf{x}_n) + b) \geq 1 \quad (2)$$

where t_n denotes the classification outcome which may be either -1 or 1 , depending on which side of the boundary the data point lies. The classification problem itself is not necessarily linearly separable. To account for this, slack variables, ζ_n , can be incorporated that permit data points to lie on the wrong side of the margin boundary, quantifying the empirical risk associated with those points.²³ The objective function to be minimised is then,

$$E = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \zeta_n \quad (3)$$

with parameter $C > 0$ controlling the trade off between margin and slack variable penalty.

A one-dimensional SVM (1D-SVM) approach is proposed to learn appropriate tolerance thresholds and inform future decisions. The artefact probing procedure considered in this paper is an off-line method, which necessitates a degree of disruption to the production process to conduct. Subsequently, the datasets obtained for use in a performance monitoring context are often relatively small. SVMs are generally better suited to tasks involving smaller datasets as compared with other predictive algorithms – the classic example being the *artificial neural network*, which requires extremely large training sets to properly mitigate the risks of overfitting – due to the fact that only a subset containing the support vectors are required to construct the hyperplane, irrespective of the total size of the training set. In fact, a critical drawback of the SVM is high computational costs for processing large datasets, with a common solution in recent research efforts concerned with extracting reduced training sets that are most likely to contain the support vectors.²⁴

Classes were assigned based on each measurement's relationship with the engineer-determined tolerance

threshold at the time, resulting in two classes; *in-tolerance*, and *out-of-tolerance*. The size of the dataset was increased by sampling each observation numerous times with additional Gaussian noise. This strategy ensured that the generalised model could be properly cross-validated (CV); it is possible that the data may contain only a single instance of one of the classes, causing the standard SVM algorithm²⁵ to fail, even with a Leave-One-Out CV scheme. Additionally, a number of points were introduced in the same manner close to 0.000, simulating the ideal case which will always hold the class *in-tolerance*. A hold-out test set consisting of 10% of the total data was separated, and a grid search CV²⁵ with five folds conducted on the remainder to optimise the trade-off parameter, C . As the application is one-dimensional, there is no requirement to deal with nonlinearity, so a simple linear kernel is appropriate. Appropriate decision boundary construction was assessed by calculating the F1 score,

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

on the hold-out test set, where *precision* is the proportion of relevant instances among the identified *out-of-tolerance* instances, and *recall* is the total amount of *out-of-tolerance* instances that are correctly identified.

Legacy trending with incorporated thresholds

In order to effectively present historical machine accuracy data in the temporal domain, it is extremely important to maintain context through inclusion of the tolerance thresholds throughout that period. Referring back to Tufte's principles, it would be visually beneficial to provide this in the form of a multifunctional graphical element. A Gaussian Process (GP) can be loosely interpreted as a generalised, multivariate Gaussian distribution over an infinite-dimensional function space. Rasmussen²⁶ defines the GP as *a collection of random variables, any finite number of which have a joint Gaussian distribution*. Just as a Gaussian distribution can be wholly described by its mean, μ , and variance, σ^2 , so too is the GP specified entirely by its mean $\mu(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$ of the real process $f(\mathbf{x})$,

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \quad (5)$$

where,

$$\mu(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \quad (6)$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))] \quad (7)$$

Prediction with a GP involves sampling from the *posterior* probability distribution, which is obtained by conditioning the joint Gaussian prior distribution on the observations. The solution can be interpreted as restricting the joint prior distribution to contain only those functions which agree with the training data

observations. The covariance function, or kernel,²⁷ is applied elementwise on a selection of training points, arranged in a matrix X , and test points, X_* , to construct the covariance matrices $K = K(X, X)$, $K_* = K(X, X_*)$ and $\mathbf{k}(\mathbf{x}_*) = \mathbf{k}_*$, for a single test point \mathbf{x}_* . In this compact notation, the key equations for a predictive mean, \tilde{f}_* , and variance, $\mathbb{V}[f_*]$, with GP regression are,

$$\tilde{f}_* = \mathbf{k}_*^T (K + \sigma_n^2 I)^{-1} \mathbf{y}, \quad (8)$$

$$\mathbb{V}[f_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T (K + \sigma_n^2 I)^{-1} \mathbf{k}_* \quad (9)$$

where σ_n^2 is a noise parameter, and I is the identity matrix. It is proposed, in this paper, to extend these equations to include a modified version of $\mathbb{V}[f_*]$, incorporating the engineer-determined tolerance thresholds into the GP output as a multifunctional visual element in a *trending GP*. This will be referred to as the Performance Indicating Confidence Interval (PICI), given by,

$$\text{PICI}[f_*] = 1.96 \times \sqrt{\mathbb{V}[f_*]} \times \left(\frac{f_*}{\beta} \right) \quad (10)$$

where β is a vector describing the tolerance thresholds at the corresponding values of f_* , and the expression is multiplied by 1.96 to reflect a 95% confidence bound. This approach provides an efficient means of representing the trend data, which can provide an instant appraisal in a similar manner to the benchmark presentation of the proprietary software.

Event detection

Labelling important events in the data is one of six principles outlined by Tufte for maintaining graphical integrity²⁰; with this, the GP modelling method is further utilised in this paper to construct intelligent *event detectors*, enhancing the quality of information on the trending GPs to assist the ME or visiting specialist. The method is applicable to any three or more variables that may be considered to have some underlying relationship, such as the Primary- and Secondary-axis tests conducted in the probing procedure. Firstly, construct n multivariate input GPs from n variable sources, such that one variable is the prediction target and all others are a multivariate input,

$$f(\mathbf{x}_n) \sim \mathcal{GP}(\mu(\mathbf{X}_m), k(\mathbf{X}_m, \mathbf{X}'_m)) \quad (11)$$

where m indicates a matrix constructed of all vectors in the set where $m \neq n$. After conditioning these GPs on a training set consisting of only non-events, one can make predictions on a test set including both events and non-events, obtaining the residuals between actual and predicted values,

$$\mathbf{r}_{*,n} = \mathbf{y}_{*,n} - f(\mathbf{x}_{*,n}) \quad (12)$$

then find the average residual across all n event detector GPs, \mathbf{r}_* , for each observation in the test set. Events can

now be identified as residuals which exceed the threshold given by the average GP variance, $\mathbb{V}[\mathbf{f}_*]$, at each time step,

$$\text{events} = \begin{cases} 1, & \mathbf{r}_* \geq \tau \times \sqrt{\mathbb{V}[\mathbf{f}_*]} \\ 0, & \mathbf{r}_* < \tau \times \sqrt{\mathbb{V}[\mathbf{f}_*]} \end{cases} \quad (13)$$

where τ is a fixed sensitivity parameter controlling the threshold, and 1, 0 are identified events and non-events, respectively. In any system, faults can be broadly categorised into two main types.²⁸ *Soft* faults refer to those which progressively develop with time, such as general wear and tear on moving components which leads to slowly degrading positional accuracy of the machining centre. *Hard* faults refer to those which occur instantly, such as a collision event which results in an immediate and permanent alteration to the structural loop of kinematic components. The event detector method outlined in this section presents a solution to automatically identify hard faults, and flag the results on the trending GP visualisations. For each of the case studies presented in this paper, the event detector was trained on a subset of the data deemed to represent the normal operating condition, and evaluated on a hold-out test set containing both events and non-events. Ideally, the selected training subset would occur at the very beginning of the data acquisition period. However, in many real-world applications, the proprietary software is initially implemented in response to some problem with the machine. So, it is not uncommon to observe events, or periods of activity not deemed to represent the normal operating condition, at the beginning of a given acquisition period. To account for this, the training subset for each case study was selected as one continuous stretch in the acquisition period which does reflect the normal operating condition. In development and deployment of a production system, consideration must be given to the informed (and preferably, automated) setting of this training period.

Interpreting the graphs: Demo data

This section presents a short tutorial on interpreting the graphical tools produced by the methods described above. For the purpose of this paper, analysis is restricted to five outputs from the artefact probing procedure, though the developed framework can process all items shown in Figure 5. Specifically, these are:

Axis checks

- **Primary-axis** Assessment of the health of the Primary/C-axis. An incremental probing procedure from $B=0^\circ$ $C=0^\circ$ (the home position) to $B=0^\circ$ $C=360^\circ$, rotating only the Primary-axis. Measurements are given as the maximum expected error across axis positions probed, in *mm*.
- **Secondary-axis (+ve)** Assessment of the health of the Secondary/B-axis. An incremental probing procedure from $B=0^\circ$ $C=0^\circ$ to $B=90^\circ$ $C=0^\circ$,

rotating only the Secondary-axis. Measurements are given as the maximum expected error across axis positions probed, in *mm*.

- **Secondary-axis (-ve)** Assessment of the health of the Secondary/B-axis. An incremental probing procedure from $B=0^\circ$ $C=0^\circ$ to $B=-90^\circ$ $C=0^\circ$, rotating only the Secondary-axis. Measurements are given as the maximum expected error across axis positions probed, in *mm*.

Probe checks

- **Overall probe performance** Measure of the probe's overall uncertainty in collecting accurate measurements, in *mm*.
- **Probe pre-travel variation** Measure of the uncertainty attributed to probe pre-travel, or *lobing*, effect, in *mm*.

A synthetic dataset was generated for the purposes of this demonstration, and to select appropriate kernels and hyperparameter optimisation ranges for the GP methods. Data points were simulated over an eighteen-month period, representing the normal operation of a typical machining centre. Certain notable characteristics were introduced into the dataset to illustrate the graphical effects that the proposed methods should have. These are:

- A significant event, or hard fault, occurs in the *Primary-axis* trend, on 2017-12-28.
- A significant event, or hard fault, occurs in the *Secondary-axis (+ve)* and *Secondary-axis (-ve)* trends, on 2018-05-02.
- The tolerance is changed for *Primary-axis*, on 2018-03-14; tolerance is changed for *Secondary-axis (+ve)* and *Secondary-axis (-ve)*, on 2018-05-07.
- There are no recorded *out-of-tolerance* measurements in the *Overall probe performance* and *Probe pre-travel* data.

Figures 6 shows the learnable tolerance threshold method, as applied to the demonstration dataset. On the format; measurements are plotted in a single dimension along the X-axis, with separate symbols for *in-tolerance* and *out-tolerance* class labels. Actual tolerance threshold changes are tagged at their corresponding values, with the date at which each was changed referenced to the right. The data-driven tolerance threshold is represented by the red-green colourmap, with the data-driven tolerance threshold at the decision boundary and two class separations on either side.

There are certain indicators in this data presentation which can provide an instant appraisal for understanding how the machine has performed with respect to the tolerance thresholds, and how the thresholds themselves have been managed. Following the initial setting on 2017-05-10, the tolerance changes on 2018-03-14 and 2018-05-07 are immediately clear, with flags indicating



Figure 6. Demo data: data-driven tolerance thresholds. From top to bottom: *Primary-axis*, *Secondary-axis (+ve)*, *Secondary-axis (-ve)*, *Overall probe performance*, and *Probe pre-travel*. Tolerance change occurrences in the data acquisition period are noted alongside each graphic.

their respective increases from 0.1 to 0.25 mm. The data-driven thresholds for all three axis checks closely match the threshold set by the engineer initially; suggesting that, based on past-usage, the later threshold changes may be unwarranted. For the two Secondary-axis checks, it is noteworthy that one of the *in-tolerance* observed points lies in the *out-of-tolerance* zone designated by the data-driven threshold. Due to the threshold increase on 2018-05-17, this measurement – which is the highest value seen in the dataset – concludes that the machine is capable and fit for production. This change may be completely logical – the result of a new part in production and corresponding new tolerance requirements, for example; it could also represent an unwarranted change by the operator, in order to pass a test which would otherwise fail. The latter could have serious consequences for finished part quality. In either case, the learnable tolerance threshold and the graphical display shown in Figure 6 gives instant access to this information, for either the onsite ME or visiting specialist to interrogate. For the probe checks, all measurements were recorded as *in-tolerance*, and subsequently a data-driven threshold was not constructed. This information, along with the ME-set threshold and distribution of the data, is, again, quickly accessible from the display.

Figure 7 shows the trending GP method with intelligent event detection, as applied to the demonstration dataset. Each subplot presents the time-series

information along the X-axis and measurement recorded by the artefact probing procedure on the Y-axis. Observed measurements are shown in a scatter plot, with separation for the data used to train and test the trending GP. The mean function, μ , and standard deviation, σ (the 2σ confidence interval), are shown. Note that this is only shown above the mean function, as opposed to the typical representation of a confidence interval which would be above and below, to describe the data distribution. The focus, in this case, is on leveraging the confidence interval for visual enhancement, and as such simplifying the plots in this way is more appropriate. Incorporated into the threshold is the PICI, indicating the relationship between measurements and tolerance thresholds at any give time in the series. Red-green colourmapping is applied to give a fast appraisal of performance, with green regions representing good performance, tending to red, which indicate out-of-tolerance measurements. Flags in the axis checks show the event detector outputs.

The graphical representation shown in Figure 7 provides fast access to legacy data, displayed over the collection period. The two significant events, on 2017-12-28 for the *Primary-axis* and 2018-05-02 for the *Secondary-axis*, are clearly visible by spikes in the GP mean function. In both cases, the event detector accurately flags the problem regions. PICI colourmapping shows that there is an exceedance of the tolerance

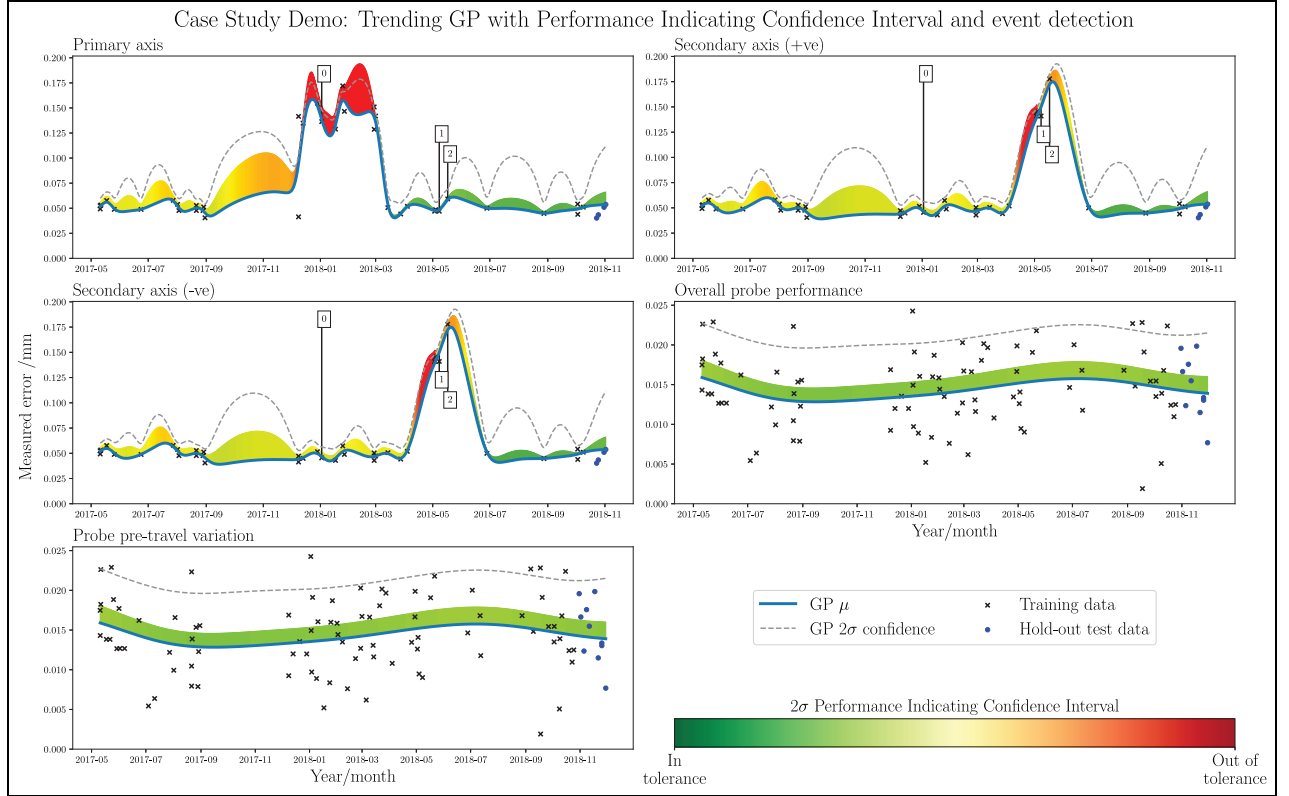


Figure 7. Demo data: trending GP with PICI and event detection. Significant events occur at 2017-12-28 and 2018-05-02. Each subplot shows a different data stream, derived from the extrapolation of the elements in the benchmark summary of Figure 5. From top left: *Primary-axis*, *Secondary-axis (+ve)*, *Secondary-axis (−ve)*, *Overall probe performance*, and *Probe pre-travel*.

Table 1. NMSE results on the GP forecasting demonstration data.

Primary-axis	Secondary-axis (+ve)	Secondary-axis (−ve)	Overall Probe Performance	Probe Pre-travel
0.071	0.075	0.075	0.666	0.666

threshold in both of these regions. The most extreme measurement following the Secondary-axis event on 2018-05-02 shows a tolerance change, as the PICI colour drops from red to yellow, even though the measured error increases. The representation also shows the consistency of data collection, with taller 2σ thresholds corresponding to longer periods of inactivity. The probe checks in the demonstration dataset do not contain any particular soft trend characteristics. The trends are generated by the same Gaussian distribution with a mean of 0.015 mm, and it can be seen that the GP trend reconstructs this well with minimal overfitting. It is notable from these plots that many repeat measurements have been made throughout the collection period, which is a potentially useful piece of information in forming a picture of how the procedure has been applied. The event detector was trained on data covering the first six months of acquisition, and tested on the remaining twelve. Appraising the first three subplots (the *axis* data) in Figure 7, it can be seen that the

detector triggered twice during the first event period (event in the *Primary-axis*), and a further three times during the second event period (event in the *Secondary-axis*).

A test set was held out prior to model training to assess the practicality of predicting future trends with this method. The Normalised Mean Squared Error (NMSE) is applied to evaluate the model performance in predicting near-future events. NMSE values below 1.000 signify the existence of correlation between the test set observations and model prediction, NMSE values above 1.000 suggest poorer performance than simply applying the mean of the data as the predictor.²⁹ The results are presented in Table 1. For the demonstration dataset, the results are reasonable; particularly in the axis checks, where they are very low, indicating a notable correlation between observations and model predictions. The probe checks are less impressive, but still better than simple application of the mean value. However, predicting a future trend in this way relies on

the assumption that proceeding data points will follow a similar pattern; in other words, it is only reliable for predicting soft fault trends. However, hard faults can and do occur throughout the life-cycle of a typical machining centre, which ultimately is the motivation for intelligent event detection highlighted in this paper. For this reason, such an approach to trend analysis will never be appropriate for predicting future states, and can not be implemented for a monitoring system which relies solely on artefact probing data. Given the scope of this paper, however, this does not affect the efficacy of the proposed methods, as the core focus is one of legacy data visualisation for fast appraisal of machine usage.

Kernel selection

The kernel (covariance function) chosen for a GP model can significantly impact the resulting form of the predictor. The Radial Basis Function (RBF), otherwise known as the Squared-Exponential kernel, is a popular kernel selection which is infinitely differentiable and thus very smooth. It is parameterised by a non-negative length-scale parameter, ℓ , and the kernel is given by,

$$k_{RBF}(r) = \exp\left(-\frac{r^2}{2\ell^2}\right) \quad (14)$$

where r denotes the distance $\|\mathbf{x} - \mathbf{x}'\|$. The Matérn kernel is a generalisation of the RBF, with an additional non-negative parameter ν controlling the smoothness of the resultant function; the general form is given by,

$$k_M(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{\ell}\right)^\nu K_\nu\left(\frac{\sqrt{2\nu}r}{\ell}\right) \quad (15)$$

where K_ν is a modified Bessel function.³⁰ The Matérn kernel is most simply expressed when ν is half-integer, $\nu = p + 1/2$, where p is a non-negative integer. Often, the most interesting cases for machine learning are $\nu = 3/2$ and $\nu = 5/2$,²⁶ given by,

$$k_{M(3/2)}(r) = \left(1 + \frac{\sqrt{3\nu}r}{\ell}\right) \exp\left(-\frac{\sqrt{3\nu}r}{\ell}\right) \quad (16)$$

$$k_{M(5/2)}(r) = \left(1 + \frac{\sqrt{5\nu}r}{\ell} + \frac{5r^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5\nu}r}{\ell}\right) \quad (17)$$

Setting $\nu = 1/2$ gives the absolute exponential kernel,

$$k_{M(1/2)}(r) = \exp\left(-\frac{r}{\ell}\right) \quad (18)$$

Finally, the Rational Quadratic (RQ) kernel represents a scale mixture of RBF kernels with different length-scales, given by,

$$k_{RQ}(r) = \left(1 + \frac{r^2}{2\alpha\ell^2}\right)^{-\alpha} \quad (19)$$

where α is a non-negative scale mixture parameter. Figure 8 illustrates the five kernels described in equations (14)–(19), as applied to the trending GP demonstrated in Figure 7. For this application, the best representation is one which fits the training data with minimal inflection between observations, and also provides a smooth confidence interval to harness for a performance indicator. A function that produces high inflection between observations runs the risk of fitting into negative values, which are not permissible for this application where the measured errors are always non-negative. Moreover, it is not possible to know what usage occurred in the period between observations, so the best approximation would be a linear trend between neighbouring observations. There is, of course, a trade-off with function smoothness to be considered here, as perfectly fitting every observation would also be undesirable in terms of producing an informative, easy-to-interpret trend.

Considering the comparison in Figure 8, the RQ kernel presents itself as the most appropriate option for modelling the time-series trend. There is minimal inflection between observations, and the mean function fits smoothly when it passes through more densely-populated regions. The 2σ threshold provides a significant area to house a PICI, whilst having a minimal impact on the Y-axis limits for displaying the mean function itself. The Matérn 5/2 and 3/2 kernels both have a similar issue with inflection, and the issue of the mean function fitting into negative space is noted in both cases. This has a major impact on the PICI in unobserved regions, as the mean function's misrepresentation of the error state is passed onto the PICI. The Matérn 1/2 kernel provides a more realistic fit in the unobserved periods, with a relatively linear relationship between neighbouring points. This characteristic is let down by the confidence intervals, which are extremely large and restrict the quality of information presented by the mean function. The RBF kernel suffers a similar drawback to Matérn 5/2 and 3/2, in that there is considerable inflection in the unobserved periods, dipping below zero for extended periods during the training region and wildly overestimating in the later, unobserved testing region.

Figure 9 presents the same kernel comparison, but as applied to the GP event detector method. Each subplot shows the residual values for the different kernel selections by a blue, solid line, with a 3σ novelty threshold shown by red, dotted line. The two events, at 2017-12-28 and 2018-05-02 are flagged in each of the subplots. The event detectors were trained between 2017-05 and 2018-01, then tested on the remaining data leading up to 2018-11. Logically, the ideal kernel for event detection is one which triggers only during periods of potential events, producing smaller values throughout periods of normal operation which do not exceed the given threshold.

Observing the subplots in Figure 8, it appears that the GP residuals generally peak as intended when

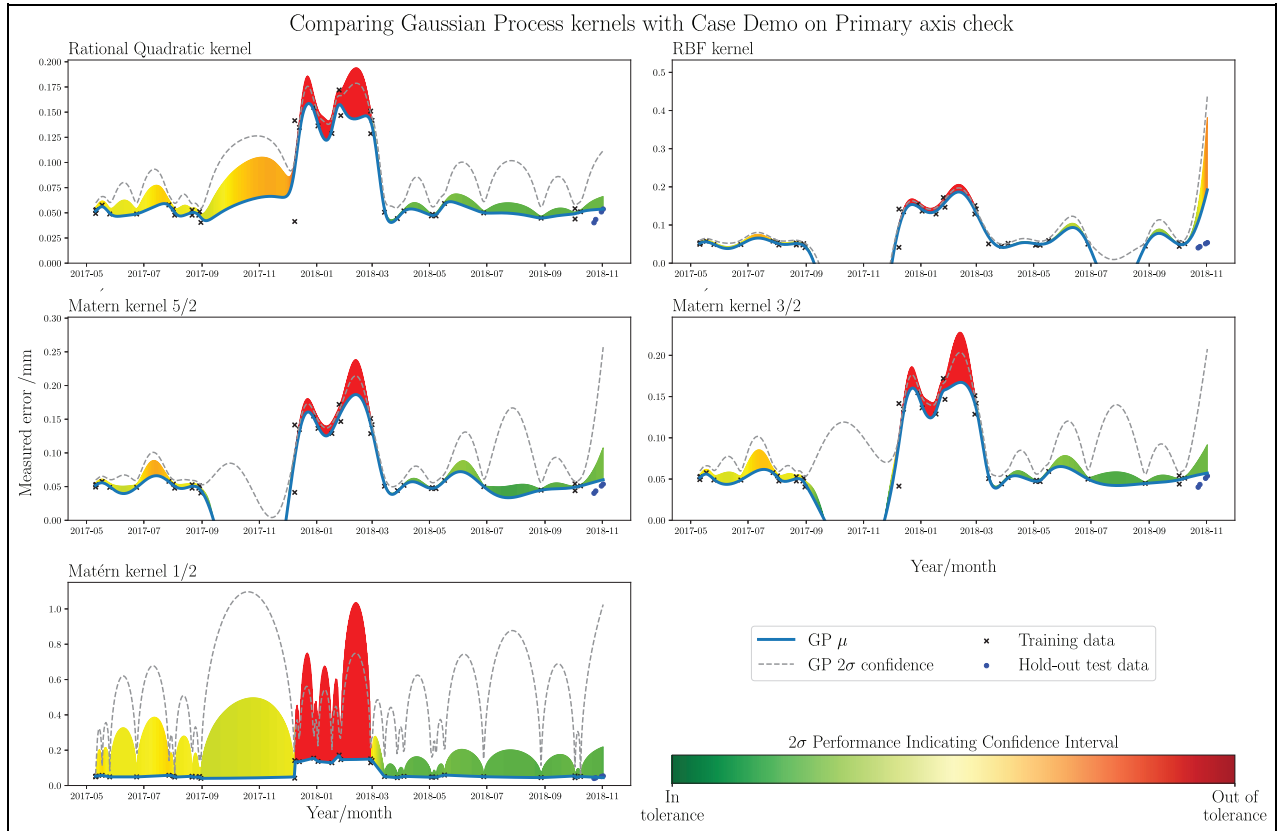


Figure 8. Comparing five different kernels for the trending GP, on the *Primary-axis* procedure. Kernels from top left: Rational Quadratic, Radial Basis Function, Matérn 5/2, Matérn 3/2 and Matérn 1/2.

events occur for all kernel options. The weakest seems to be the RBF kernel, which activates less accurately for event 1 and misfires at the end of the acquisition period. RQ, and the three Matérn kernels, all produce similar trends. Considering the threshold values obtained through the GP variances, the RQ kernel is the only option to produce a reasonably smooth threshold, whereas the RBF and Matérn options result in erratic thresholds. A smoother, more predicable decision boundary is preferable for this application, as large, unexpected jumps may result in false positive event detection, such as that which occurs in the RBF/Matérn systems around 2018-09. For these reasons, the RQ kernel was selected as most appropriate for the event detector. The setting of the novelty threshold could be varied based on requirements, and the experience of the ME responsible. However, for this assessment, a conservative 3σ rule was deemed appropriate, effectively triggering for both events in the simulated example, with minimal extra triggers once the event was under way. This novelty threshold setting was fixed and carried through to the case studies in the following section.

Case studies

This section now presents three case studies on engineering datasets, acquired directly from anonymised

industrial partners. The processing and display of results for each method are identical to those which are introduced in the previous section, *Interpreting the graphs: Demo data*.

Case study I

Figure 10 presents the threshold checking tool for case study 1. It is clear that the threshold has been changed numerous times throughout the acquisition period, which may indicate a mixture of parts for production with varying tolerance requirements. In the axis checks, however, many of the tests have returned *out-of-tolerance* class labels. This suggests that the machine has been consistently not conforming to the tolerance requirements, reflecting a poor monitoring strategy which may ultimately lead to issues with finished part quality. The threshold-checking presentation makes this inference immediate and easily accessible.

Table 2 presents the F1 scores for the data-driven threshold classifiers. The data is linearly separable in three of the five cases, returning F1 scores of 100% in the test set. In the other two there is some crossover, so the soft margin parameter permits some misclassification, however the F1 scores are still very good, at 99% for the *Primary-axis* and 89% for the *Secondary-axis* (*-ve*). In all but *Primary-axis*, a tolerance threshold is optimised which logically separates the observations.

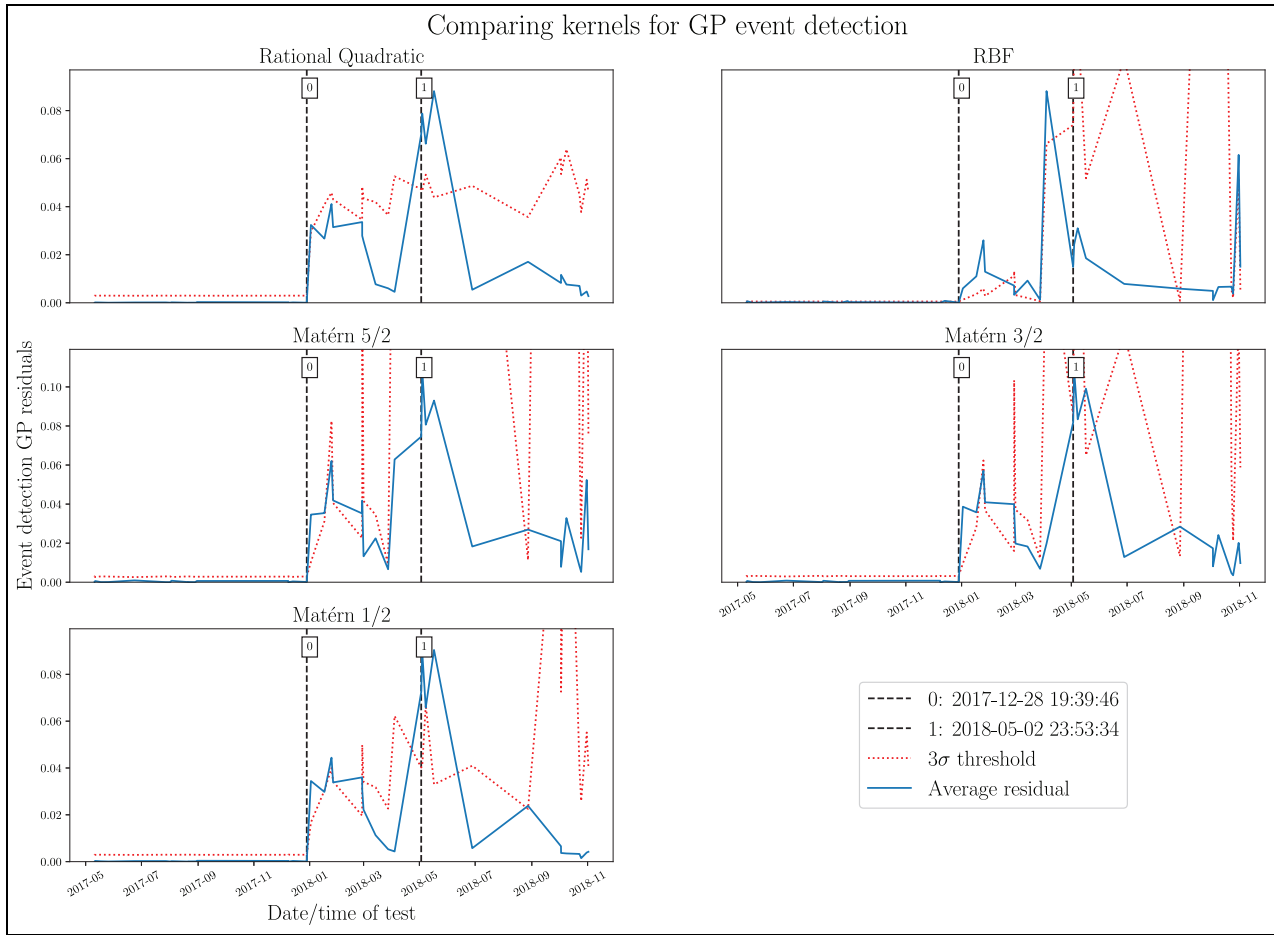


Figure 9. Comparing five different kernels a GP event detector, with simulated events at 2017-12-28 and 2018-05-02. Kernels from top left: Rational Quadratic, Radial Basis Function, Matérn 5/2, Matérn 3/2 and Matérn 1/2. The 3σ decision threshold is shown by dotted lines, alongside the target events for the assessment by vertical dashed lines.

The threshold for the probe checks very closely resembles the final ME setting for *Probe pre-travel*, and gives a similar but slightly more conservative setting for *Secondary-axis (+ve)* and *(-ve)*. In the *Primary-axis*, there is only one *in-tolerance* observation available to construct the threshold which is mixed deep within the cluster of the other class. As a result, the calculation is biased towards the simulated ideal data points, and the threshold is constructed at a very conservative (low) value. Although this isn't likely to be an appropriate tolerance setting, it is a logical conclusion for the algorithm based on the historical data, and is a further red flag pointing towards mis-management of the tolerancing in this case study.

Figure 11 shows the output of the trend analysis tool as applied to case study 1. The PICI integration paints an immediate picture of the error states with respect to tolerance thresholds throughout the acquisition period, particularly when viewed alongside Figure 10. In *Primary-axis*, it is clear that the machine has been running consistently out-of-tolerance. Closer inspection of the 2σ confidence line indicates that the

tolerance was exceeded to a greater extent at the beginning of the acquisition period, pointing to the possibility that the increase may have occurred in order to force the test to pass. Tolerance changes are more clearly visible in the probe checks, with a clear change occurring around 2016-08 and another in 2017-04 as made apparent by harsh changes in the PICI. Table 3 presents the NMSE results for time-series GPs as future trend predictors. The problematic nature of applying a trending predictor to a system with potential for hard faults is particularly apparent in the *Secondary-axis* checks, where the NMSE scores are 6.93 and 1.59 for checks in the positive and negative directions, respectively. Pre-identified target events for case study 1 occur at 2016-07-25, 2016-08-26 and 2017-04-18. The event detector was trained on five months worth of data collected between 2016-11 and 2017-04, where the system is operating consistently in normal condition. The automated system flags the target events well, with two triggers around the first two events and a single trigger for the final event.

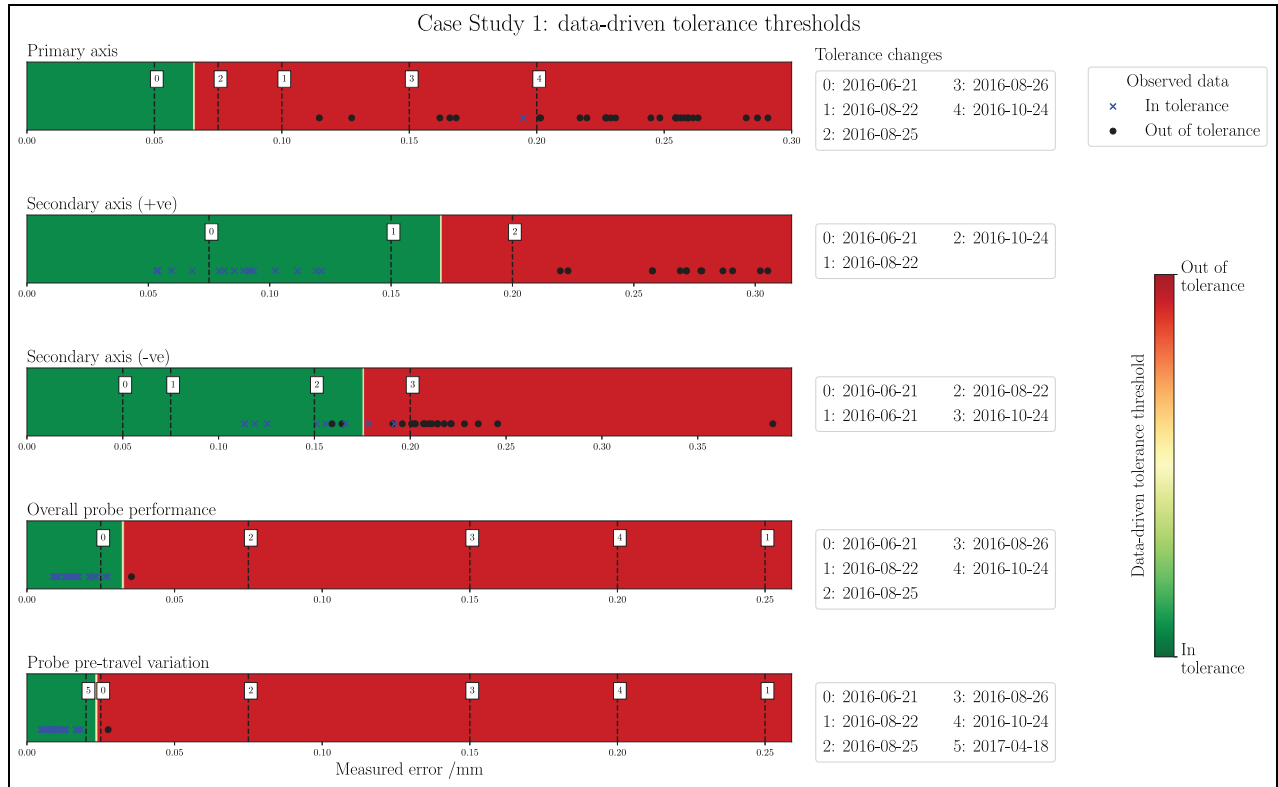


Figure 10. Case study 1: data-driven tolerance thresholds. From top to bottom: *Primary-axis*, *Secondary-axis (+ve)*, *Secondary-axis (-ve)*, *Overall probe performance*, and *Probe pre-travel*. Tolerance change occurrences in the data acquisition period are noted alongside each graphic.

Table 2. F1 scores on a 10% hold-out test set, for the 1D-SVM data-driven tolerance threshold.

Case Study no.	Primary-axis	Secondary axis (+ve)	Secondary axis (-ve)	Overall probe performance	Probe pre-travel
1	0.99	1.00	0.89	1.00	1.00
2	1.00	1.00	1.00	0.93	0.91
3	0.95	1.00	0.95	N/A	0.99

Table 3. NMSE scores on a 10% hold-out test set, for the time-series GP.

Case Study no.	Primary-axis	Secondary axis (+ve)	Secondary axis (-ve)	Overall probe performance	Probe pre-travel
1	0.05	6.93	1.59	0.06	0.28
2	0.05	0.03	0.21	0.25	1.10
3	0.21	0.36	0.14	0.03	0.05

Case study 2

Figure 12 presents the threshold checker for case study 2. All three axis checks are linearly separable between the two classes, which is immediately clear in the Figure and also by the F1 scores in Table 2. This separation indicates a methodical approach to setting the tolerance thresholds, and that the machine is likely to be under relatively constant tolerance conditions throughout the operating period. In all cases, a set of sensible tolerance

thresholds were learned which closely reflect most of the final values settled on by the ME. The exception to this is in *Secondary-axis (+ve)*, which sees a peculiar increase to 0.5mm near the end of the acquisition period. It is unlikely that this change was warranted, particularly considering that the *Secondary-axis (-ve)* check remained at a logically more appropriate 0.25mm . Presenting the data in this way makes this unwarranted change extremely easy to spot and correct,

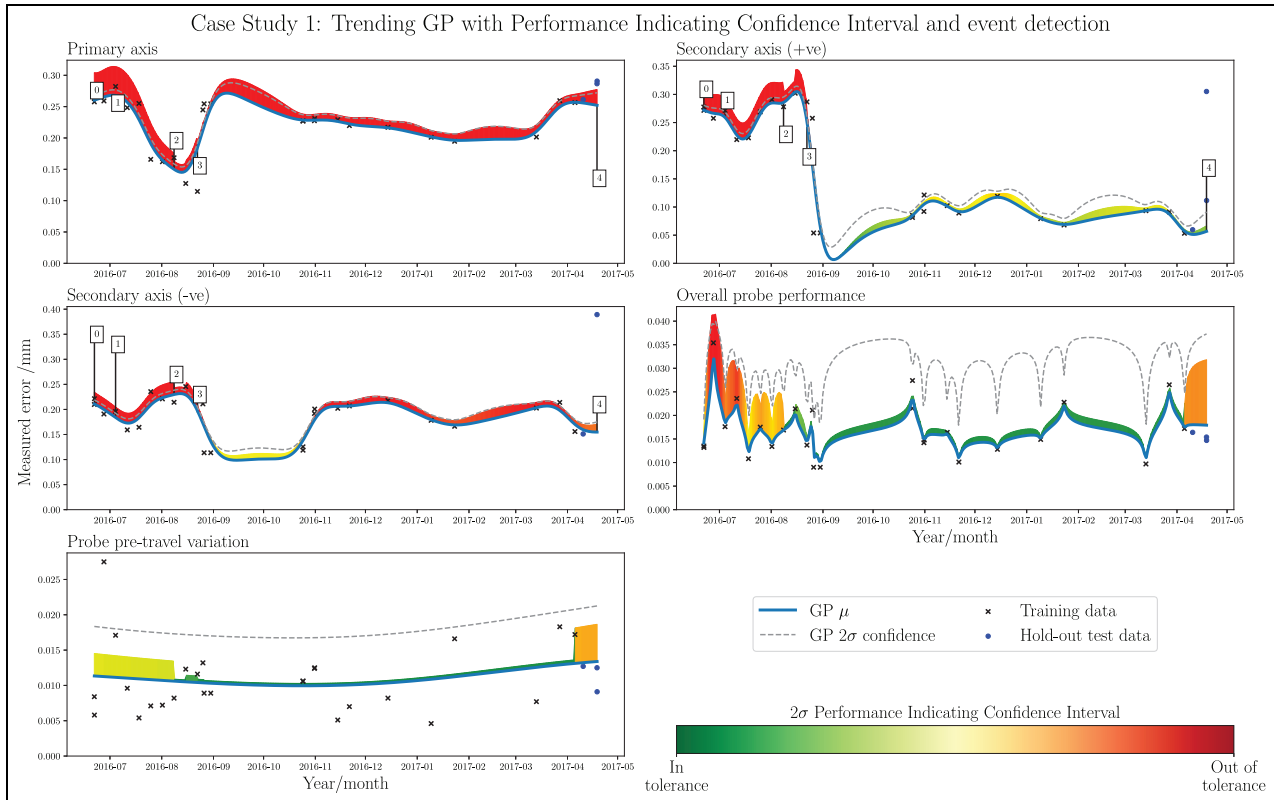


Figure 11. Case study 1: trending GP with PICI and event detection. Each subplot shows a different data stream, derived from the extrapolation of the elements in the benchmark summary of Figure 5. From top left: *Primary-axis*, *Secondary-axis (+ve)*, *Secondary-axis (-ve)*, *Overall probe performance*, and *Probe pre-travel*.



Figure 12. Case study 2: data-driven tolerance thresholds. From top to bottom: *Primary-axis*, *Secondary-axis (+ve)*, *Secondary-axis (-ve)*, *Overall probe performance*, and *Probe pre-travel*. Tolerance change occurrences in the data acquisition period are noted alongside each graphic.

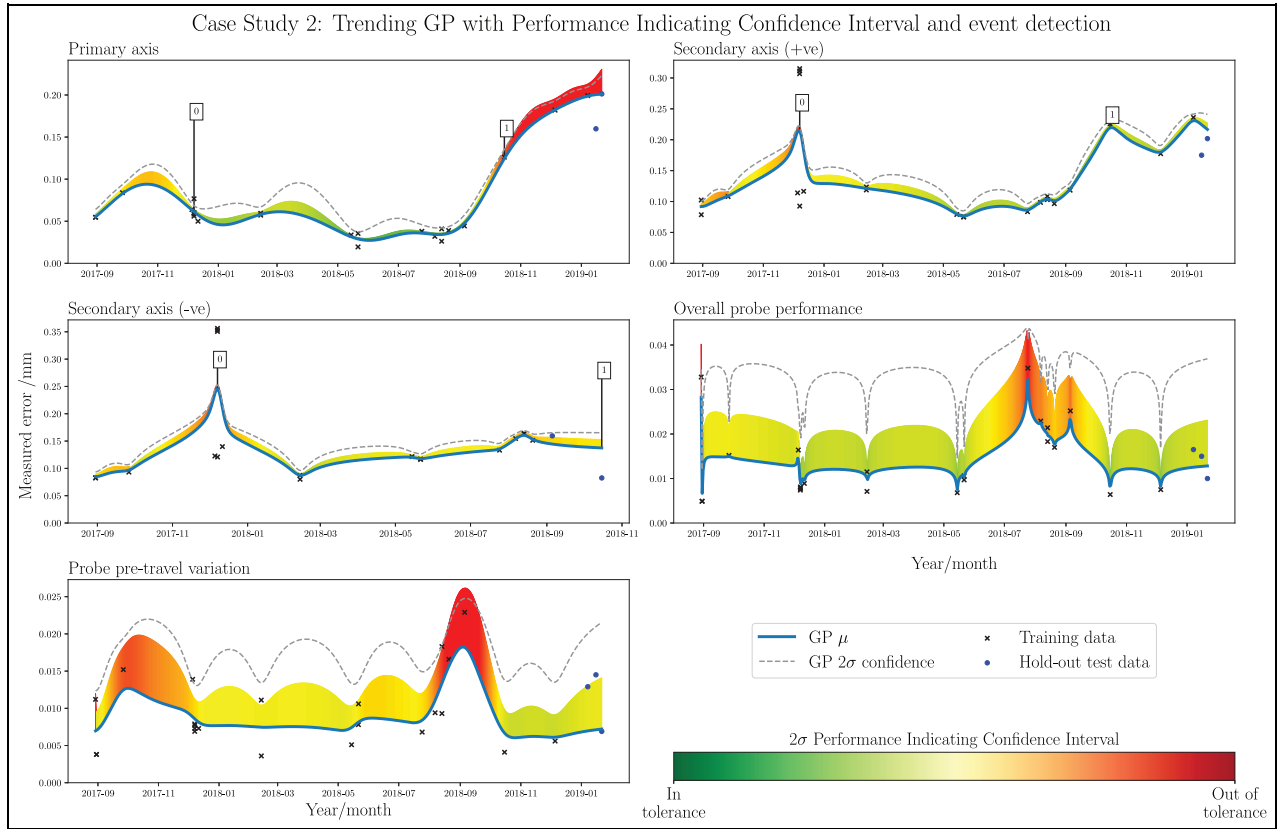


Figure 13. Case study 2: trending GP with PICI and event detection. Each subplot shows a different data stream, derived from the extrapolation of the elements in the benchmark summary of Figure 5. From top left: *Primary-axis*, *Secondary-axis (+ve)*, *Secondary-axis (-ve)*, *Overall probe performance*, and *Probe pre-travel*.

before it has the chance to negatively affect finished part quality. Identification of this change is also a core justification for implementing a data-driven tolerance threshold, establishing a two-tiered warning system for ME-determined and model-determined normal operation. F1 scores indicate a linear separation of classes in the three axis checks, and scores are above 90% for the two probe checks.

Figure 13 presents the trending GP tool for case study 2. The probe checks provide a good example of the PICI visualisation, giving a clear indication of the historical performance and out-of-tolerance instances around 2018-08/09. NMSE results indicate decent soft trend predictive performance, with only *Probe pre-travel* producing a high value above 1.000. It is evident that the accuracy of the *Primary-axis* deteriorates towards the end of the acquisition period, which accounts for the out-of-tolerance points observed in Figure 12. There is clearly a hard fault which occurs in 2017-12, causing a spike in the trend; however, as corrective action was immediately applied, the full magnitude of the error is not reflected. Although the full extent of the tolerance exceedance is not visualised in the trend, this is actually a better representation of the machine's general history, as the hard fault occurs for an insignificant duration in the overall chronology. The graphical presentation does, however, provide a visual

prompt of this region as an area of interest for the interrogating ME or visiting specialist. The event detector - trained on the stable period of normal operating condition between 2018-02 and 2018-10 - exactly identifies a hard fault event on 2017-12-07 in the *Secondary-axis*, and also identifies a soft fault event in both *Primary-axis* and *Secondary-axis (+ve)* later in the time series, identifying a degradation from 2018-10-15. It must be noted that the method relies on consistent data collection across all of the input variables; the final four collections after 2018-10-15 omitted *Secondary-axis (-ve)*, as such it was necessary to omit these points when computing the event detector. As the last data point collected for *Secondary-axis (-ve)* did indicate the presence of a fault, the system was able to pick it up. However, it is a noteworthy restriction of the method, and a commercially-deployed system should consider robustness to inconsistencies in data pre-processing.

Case study 3

Figure 14 presents the threshold checking method as applied to case study 3. Immediately apparent is the lack of a data-driven threshold for *Overall probe performance* and failed construction of a data-driven threshold in *Probe pre-travel*. In the case of *Probe performance*, this indicates that all measurements were

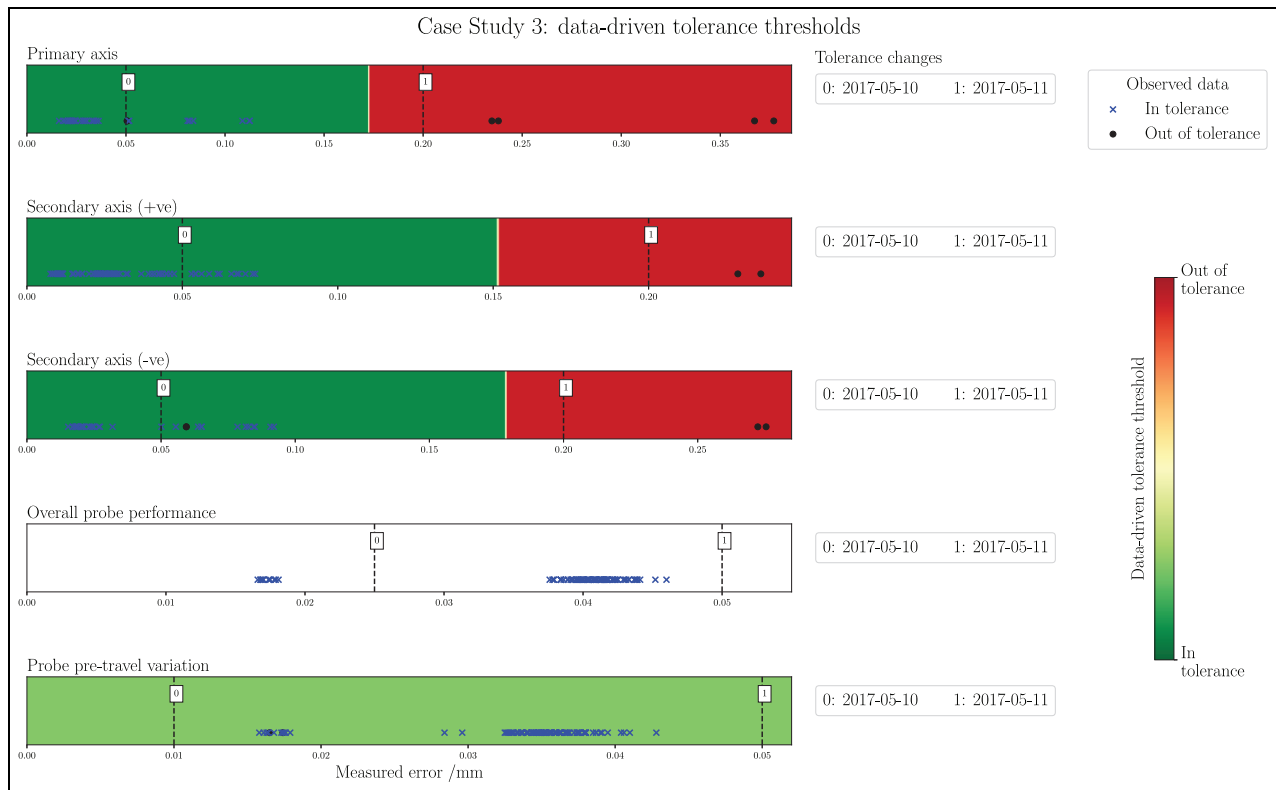


Figure 14. Case study 3: data-driven tolerance thresholds. From top to bottom: *Primary-axis*, *Secondary-axis (+ve)*, *Secondary-axis (-ve)*, *Overall probe performance*, and *Probe pre-travel*. Tolerance change occurrences in the data acquisition period are noted alongside each graphic.

considered by the ME as in-tolerance. This is generally good news; however, it is also clear that many of the points are situated quite closely to the first tolerance change (2017-05-11); at $50\mu\text{m}$, this is a loose tolerance for a modern probe, which should generally be capable of performing to $20\mu\text{m}$ or less. Ultimately, it is the ME who must decide on an acceptable limit for each check, but it is food-for-thought at least which is highlighted by this visualisation method. A data-driven threshold is not assigned to *Probe pre-travel* check, as there is an out-of-tolerance point mixed in; viewing Figure 14 in conjunction with the trending tool in Figure 15, it can be seen that this out-of-tolerance point occurs at the very beginning of the acquisition period and is quickly corrected-for. This is a common practice in first establishing use of the software; however, the $50\mu\text{m}$ setting raises the same potential concerns as discussed for the *Overall probe performance* check. The tolerance change dates are echoed in the axis checks, which essentially show a single tolerance setting was established on the second day of acquisition and maintained throughout the full period. Again, this points to consistent usage for production and methodical setting of the tolerance thresholds. In each of the axis checks, a tighter tolerance is learned than the ME had set. F1 scores again confirm the construction of the threshold, with linear separability in *Secondary-axis (+ve)* and a small

degree of misclassification in *Primary-axis* and *Secondary-axis (-ve)*.

The trending GPs in Figure 15 indicate the presence of hard faults around 2017-08-01 and 2017-08-28, which are also clearly visible as the out-of-tolerance points in Figure 14. The event detector, trained on the stable period between 2017-12 and 2018-05, accurately picks out the events, with a second trigger also occurring close to the second. The probe checks with PICI illustrate both the general trend and measurements with respect to tolerance threshold nicely. It is instantly clear that the recorded measurements were close to the specified tolerance throughout most of the acquisition period. The NMSE results in Table 3 indicate favourable performance for predicting future observations with these trends, due to the highly stable behaviour that is observed from 2018-01 onwards. One point to note is that the trend in the *Primary-axis* does dip below zero for a short time, following correction of the hard fault detected on 2017-08-28. Although the selected kernel and hyperparameter settings drastically reduce the likelihood of this happening, the occurrence in this case study shows that it has not been entirely eradicated. In a full deployment system, it may be worthwhile to include an extra step following the trending GP construction, in which any below-zero regions are replaced by more logical above-zero values; for example, by a

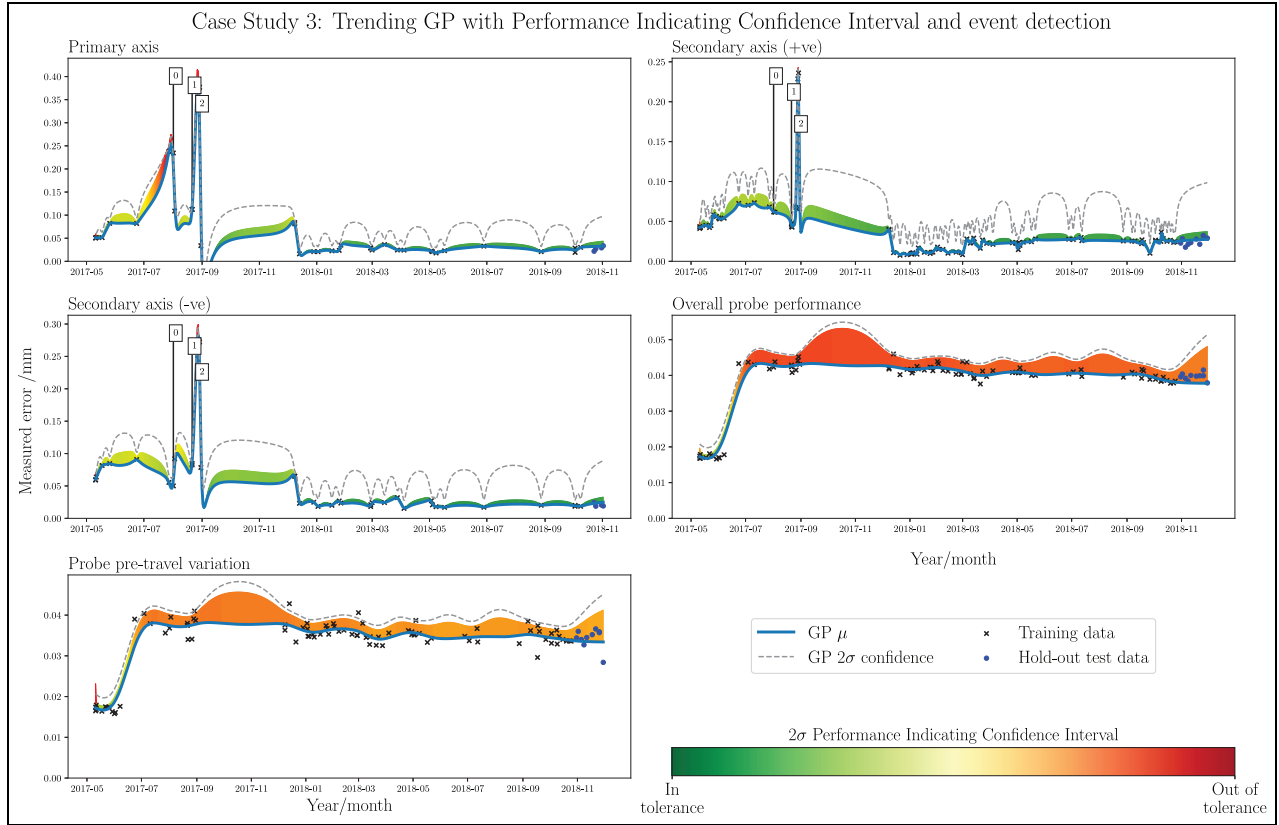


Figure 15. Case study 3: time series trend analysis with PICI and event detection. Each subplot shows a different data stream, derived from the extrapolation of the elements in the benchmark summary of Figure 5. From top left: *Primary-axis*, *Secondary-axis (+ve)*, *Secondary-axis (-ve)*, *Overall probe performance*, and *Probe pre-travel*.

linear interpolation between the two neighbouring datapoints.

Case studies: Comparison

The standardised analysis and visualisation techniques presented in the previous section give rise to the possibility machine usage comparisons between the three case studies. Comparing the tolerance threshold changes in Figures 10, 12 and 14, it is evident that the three machines have been managed in very different ways. In case study 1, there are numerous changes to the tolerance threshold throughout the acquisition period, and it is also clear that many of the measurements obtained fell into the *out-of-tolerance* class. This is a particular contrast to case study 3, which indicates no change to the tolerance after the initial setting period and a majority of measurements firmly in the *in-tolerance* class. This suggests a more consistent management approach in case study 3, with much better adherence to the tolerance thresholds that were initially set by the onsite ME.

Comparing the trends in Figures 11, 13 and 15 reveals similar characteristics. It is observed that the PICIs for the axis checks in case study 1 regularly exceed the tolerance thresholds throughout the acquisition period; this a direct contrast to the axis checks in case study 3, where there is clearly a short period

involving a hard fault near the beginning of the period, followed by strict conformance to the tolerances thereafter. Case study 2 generally indicates conformance to the tolerances in the axis checks, with a similar hard fault visible early in the acquisition period. However, there is evidence of a soft fault developing towards the end of the period in case study 2 axis checks, which is a notably different characteristic to the other case studies and easily recognisable in the trend visualisation.

An interesting observation is found in comparing the probe checks, where there is significantly lower variance in the trends for case study 3 as compared with case studies 1 and 2; this likely indicates that the probing procedure is more repeatable across different instances of the procedure conducted at different times. The proprietary software currently contains a check to assess the repeatability of the probe at the time of the test, however interpretation of the trending GPs proposed in this paper allow straightforward extension of this for the purpose of longer-term performance monitoring.

Discussion

Historical machine indicators - such as the prevalence of hard or soft faults, long-term probe repeatability of tolerance management practices - can be very useful for building up a picture of general machine usage, and are available through the appropriate processing of a relatively

limited data source obtained from historical artefact probing data. Moreover, the unique compositions of indicators observed in different machine tools support the notion of signature comparison; just as the benchmark reports can be used to compare differing signatures of machine tools within a population, so too can the methods proposed in this paper be used for comparing signatures that represent machine tool usage.

Similar to the signature, any given machine also has a normal operating condition, which may differ in a small or large way to another machine considered its counterpart. The objective with producing a data-driven tolerance threshold, as opposed to relying solely on the maintenance engineer determination, is that it is possible to set the most appropriate tolerance which represents the normal operating condition of the unique machining centre. Learning tolerance thresholds in this way provides the ME with an additional empirical comparator for assessing the performance of a population of machines, based on historical data collected throughout their respective usages.

A key contribution of this paper is in the organisation and presentation of the historical data, significantly reducing the difficulty for an onsite ME or visiting specialist engineer to gain a fast appraisal of a given machine tool's usage signature. The display methodologies were produced with consideration to Tufte's design principles, which, although somewhat subjective, should support effective communication of the data with a reasoned application of visual aesthetics. The Performance Indicating Confidence Interval is an example of this, enhancing the trending Gaussian Process into a multifunctional element which communicates both the absolute error trend, and provides the context of performance by integrating the variable tolerance thresholds throughout the acquisition period.

In a full deployment situation, the benefit to the user would be significantly increased through the use of an interactive dashboard-style interface. Displaying the data-driven tolerance and trending Gaussian Process tools side-by-side and having a linked hover-over function – such that, when an observation is hovered over on one method, it is highlighted in the other – would be a useful feature, providing a richer user experience and making the graphs easier to interpret. The hover-over function should also provide quick access to the relevant detailed reports, making in-depth analysis in areas of interest easier to conduct.

It is noted that the smoothness of the predictive mean function – attributed to kernel and kernel hyperparameter selection – affects the resultant form of the trending Gaussian Process. With a smoother mean function, quickly-corrected hard faults are not as readily represented in the trend; this is more suitable for visualising the general historical performance, and the addition of the event detection method should help bring attention to these quick corrections of hard faults, but in a more appropriate manner. Normalised Mean Squared Error results on a hold-out test set at the end

of the acquisition period evaluated the potential for using the trending Gaussian Process as a forecasting method. It was initially expected that this would not be an effective application for forecasting, due to the likelihood of hard faults interrupting any predictable, soft trends. As the core objectives with the trending Gaussian Process are trend visualisation and enrichment of historical data, this drawback does not affect the contribution of this paper. In order to reliably forecast future performance in a system like this, peripheral data streams could be utilised to identify the occurrence of hard faults, or progression of soft faults, adjusting the prediction accordingly. This functionality is out of the scope of this paper, but would be interesting future work.

The automated event detection system was shown to be effective at identifying target events, in both the simulation dataset and the three real-world case studies. In a fully deployed system, consideration must be given to defining a suitable subset for training, as the initial data acquisition period for real-world systems is not guaranteed to be free from events and representative of the normal operating condition. The datasets generated for this application are not likely to be large, so it would be appropriate to allow an experienced user the ability to retrospectively reset the training period, should the requirement arise. The event detector flags in a fully-deployed system would, again, benefit from an interactive interface, with the option to manually input diagnostic notes. Automating this diagnostic element to populate the notes without user input would be valuable further research.

Concluding remarks

The methods set out in this paper develop the information acquired from a common machine capability checking system, into a performance monitoring system for the benefit of the onsite maintenance- and visiting specialist-engineers. Performance indicators such as the incidence of hard and soft faults, management of tolerance thresholds and repeatability of measurements are identified within the datasets, and brought to the attention of the engineer through automated analysis and insightful data visualisation. The methods process only the summary statistics obtained via a calibrated artefact probing procedure; to develop the work further, this could be expanded to consider all of the raw artefact probing data, as well as information obtained from external sources as part of a wider manufacturing execution system. This would allow a much richer level of inference to be attained, including more extensive automated elements to further enhance the benefit to the user.

Acknowledgements

The authors would like to gratefully acknowledge metrology software products ltd. and the Engineering and Physical Sciences Research Council (EPSRC) grant EP/I01800X/1 for supporting this research. KW would like to additionally acknowledge support from an EPSRC Established Career Fellowship EP/R003645/1.



Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The research project was conducted in collaboration with industrial partner metrology software products ltd. (msp). In accordance with this partnership, the primary application of the work is in developing an msp procedure for performance monitoring; however, the methods presented are generic, and can be applied to any other techniques that utilise artefact probing for similar purposes.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by EPSRC grants EP/I01800X/1 and EP/R003645/1, and metrology software products ltd.

ORCID iDs

Tim Rooker  <https://orcid.org/0000-0001-9558-144X>
Jon Stammers  <https://orcid.org/0000-0001-9950-5225>

References

- Reinhard G, Jesper V and Stefan S. Industry 4.0: building the digital enterprise. *2016 Global Industry 40 Survey* 2016; 1.
- Hermann M, Pentek T and Otto B. Design principles for industrie 4.0 scenarios. In: *Proceedings of the annual Hawaii international conference on system sciences*, Koloa, HI, USA, 5–8 January 2016, pp.3928–3937. Washington, DC: IEEE Computer Society.
- International Organization for Standardization. *ISO 16792: technical product documentation - Digital product definition data practices*, 2015.
- Schwenke H, Knapp W, Haitjema H, et al. Geometric error measurement and compensation of machines-An update. *CIRP Ann Manuf Technol* 2008; 57(2): 660–675.
- International Organization for Standardization. *ISO 230-9 Test Code for machine tools, Part 9: estimation of measurement uncertainty for machine tool tests*, 2005.
- Weikert S and Knapp W. R-test, a new device for accuracy measurements on five axis machine tools. *CIRP Ann Manuf Technol* 2004; 53(1): 429–432.
- International Organization for Standardization. *ISO 230-7 Test code for machine tools - Part 7: geometric accuracy of axes of rotation*, 2015.
- Bringmann B and Knapp W. Model-based 'Chase-the-Ball' calibration of a 5-axes machining center. *CIRP Ann Manuf Technol* 2006; 55(1): 531–534.
- IBS Precision Engineering. Rotary analyzer, <http://www.ibspe.com>
- Ibaraki S, Oyama C and Otsubo H. Construction of an error map of rotary axes on a five-axis machining center by static R-test. *Int J Mach Tools Manuf* 2011; 51(3): 190–200.
- Ibaraki S, Iritani T and Matsushita T. Calibration of location errors of rotary axes on five-axis machine tools by on-the-machine measurement using a touch-trigger probe. *Int J Mach Tools Manuf* 2012; 58: 44–53.
- Ibaraki S, Iritani T and Matsushita T. Error map construction for rotary axes on five-axis machine tools by on-the-machine measurement using a touch-trigger probe. *Int J Mach Tools Manuf* 2013; 68: 21–29.
- Mayer JRR. Five-axis machine tool calibration by probing a scale enriched reconfigurable uncalibrated master balls artefact. *CIRP Ann Manuf Technol* 2012; 61(1): 515–518.
- Xing K, Rimpault X, Mayer JR, et al. Five-axis machine tool fault monitoring using volumetric errors fractal analysis. *CIRP Ann Manuf Technol* 2019; 68(1): 555–558.
- Wang W, Li H, Huang P, et al. Data acquisition and data mining in the manufacturing process of computer numerical control machine tools. *Proc Inst Mech Eng B J Eng Manuf* 2018; 232(13): 2398–2408.
- Chen X, Song Z, Li H, et al. Research on fault early warning and the diagnosis of machine tools based on energy fault tree analysis. *Proc Inst Mech Eng B J Eng Manuf* 2019; 233(11): 2147–2159.
- Jia P, Rong Y and Huang Y. Condition monitoring of the feed drive system of a machine tool based on long-term operational modal analysis. *Int J Mach Tools Manuf* 2019; 146: 103454.
- Hammond P and Brown T. NC-Checker - metrology software products ltd., 2010. <http://metsoftpro.com/nc-checker/>
- Chen YT, More P and Liu CS. Identification and verification of location errors of rotary axes on five-axis machine tools by using a touch-trigger probe and a sphere. *Int J Adv Manuf Technol* 2019; 100(9–12): 2653–2667.
- Tufte E. *The visual display of quantitative information - Graphical Excellence*. Cheshire, CT: Graphics Press, 1983.
- Boser BE, Guyon IM and Vapnik VN. Training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual ACM workshop on computational learning theory*, Pittsburgh, PA, USA, 27–29 July 1992. New York: Association for Computing Machinery.
- Bishop CM. *Pattern recognition and machine learning*. New York: Springer-Verlag, 2006.
- Wan V and Renals S. Speaker verification using sequence discriminant support vector machines. *IEEE Trans Speech Audio Process* 2005; 13(2): 203–210.
- Nalepa J and Kawulok M. Selecting training sets for support vector machines: a review. *Artif Intell Rev* 2019; 52(2): 857–900.
- Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12(2011): 2825–2830.
- Rasmussen CE and Williams CKI. *Gaussian processes for machine learning*. Cambridge, MA: MIT Press, 2006.
- Shawe-Taylor J and Cristianini N. *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press, 2004.
- Martin K. A review by discussion of condition monitoring and fault diagnosis in machine tools. *Int J Mach Tools Manuf* 1994; 34(4): 527–551.
- Cross E. *On structural health monitoring in changing environmental and operational conditions*. PhD Thesis, University of Sheffield, 2012.
- Howlett J, Abramowitz M and Stegun IA. Handbook of mathematical functions. *Math Gaz* 1966; 50(373): 358.