An Industrial PID Data Repository for Control Loop Performance Monitoring (CPM)

Margret Bauer* Lidia Auret Derik le Roux***
Vered Aharonson***

** Department of Electrical, Electronic and Computer Engineering, University of Pretoria, South Africa, (e-mail: Margret.bauer@up.ac.za; derik.leroux@up.ac.za)}*
***Department of Process Engineering, University of Stellenbosch, South Africa (e-mail: lauret@sun.ac.za)*
****School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa (Tel: +27 82 720 4621; e-mail:vered.aharonson@wits.ac.za).*

Abstract: Control loop performance monitoring methods to detect problems in PID loops are developed and tested using industrial data sets. The data is captured from the process, passed on to the researcher who tries out new detection and diagnosis methods. The data is not generally shared with other researchers working on similar problems. The authors therefore have implemented a data repository to categorise and store the data so that it becomes accessible to all researchers. Existing methods can be compared and enhanced using the data sets. This paper describes the context of CPM as well as the data repository. The repository is set up, hosted and maintained by the South African Council for Automation and Control using a professional web developer.

Keywords: PID, industrial process control, data repository, control loop performance monitoring.

1. INTRODUCTION

The majority of controllers in the process industries are single-input-single-output control loops using a PID control algorithm (Åström and Hägglund, 1995). Although successful and widespread in its applications, studies consistently find that a large proportion of PID control loops are not fulfilling their task to their best potential, that is, they are performing 'poorly' (Vilanova and Visioli, 2012).

Many algorithms and methods have been developed to detect, diagnose control loops that are performing poorly for various reasons (Jelali, 2012), which in this paper will be referred to as control loop performance monitoring (CPM). This paper does not want to give an overview of the methods available but rather refers to the appropriate literature (Huang and Shah, 1999, Ordys et al., 2007, Choudhury et al., 2008, Jelali and Huang, 2010). A fraction of these methods are implemented in industrial software tools. In Bauer et al., (2016), a survey of the current state of the art in control loop performance management reveals that the majority of responding production companies uses a solution to monitor the performance of the controllers. All commercially successful methods have been tested on industrial data from actual processes.

To understand the dynamics for developing performance algorithms it is necessary to examine the involved parties. First, there are production companies in the various industries. In the process industries, the production companies own the plant as well as the plant data. In other industries, such as the aviation industry, equipment manufacturers may own the data.

Technology suppliers provide specialized solutions and consulting services for operation and management of the plant. These providers can be in-house departments or external companies. A third party are academic researchers at universities and other research institutions who develop new algorithms, trying to solve research problems.

All three parties have different motivations. Production companies want to ensure stable and safe plant operation while technology suppliers are looking to sell their solution or services to achieve that. Academics, on the other hand, are measured by publication outputs and increasingly by acquisition of additional monetary funds. Government programs such as the research programs by the European Union provide grants for industry relevant projects.

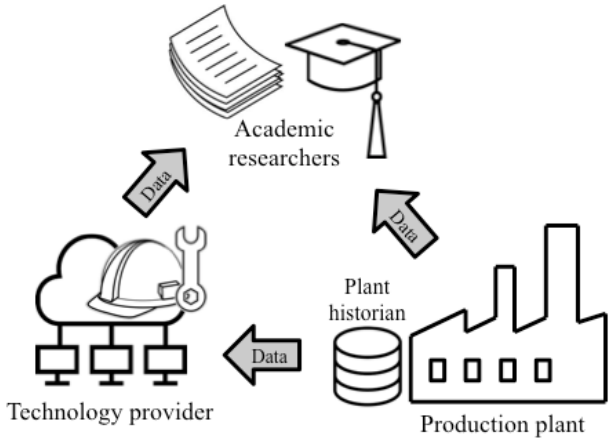

Fig. 1. Interaction between production companies, technology providers and academic institutions.

Presently, the situation of software and algorithm development can be described as follows. When developing CPM algorithms to detect poorly performing loops individual datasets are used to test performance monitoring methods and these are generally not shared with the research community. This is particularly troublesome because the methods are data-driven, that is, require process data. It becomes therefore difficult to compare and evaluate different methods. In the end, it also hinders the development of the best methods that reliably can identify potential problems in a plant.

In recent years the government initiative Industrie 4.0 – referring to the fourth industrial revolution – has driven developments in control and automation. While previous industrial revolutions dealt with mechanization, electrification and digitalization, the fourth industrial revolution connects equipment and processes via communication networks and improves production efficiency through smart planning and optimization. A key requirement for the pervasiveness of Industrie 4.0 is the introduction of standards. For CPM this means that we need recommended and standardized methods to identify process faults.

Bearing this in mind, the authors of this paper propose the introduction of a freely accessible database or repository for industrial process data to develop CPM algorithms. The data in this repository is clearly defined and represents common faults in the process industry from the actuator, sensor or process itself. The aim is to make industrial data available to researchers who develop methods for the detection of poorly performing loops.

An example of the impact that a benchmark data base can have on developing comparable research methods and results is that of the Tennessee Eastman fault diagnosis benchmark, developed by Jim Downs and Ernie Vogel of Eastman Chemical Company in Kingsport, Tennessee (Downs and Vogel, 1993). This simulated database consists of normal and fault data of a chemical engineering process, and was made accessible by the Matlab Simulink implementation by Larry Ricker[1]. The impact on fault diagnosis research has been large and the research article has been cited by over 1000 papers using the benchmark as a comparative platform.

The availability of a standard benchmark has promoted systematic research effort: not only can the performance of different methods be fairly assessed on the same data but the familiarity of the benchmark allows researchers to focus on the methods, and not the data collection.

Even though the Tennessee Eastman benchmark is popular with fault diagnosis research, some limitations are present in this dataset, discussed here. Since the origin of the data is a simulation, the true complexity of real online data – especially in terms of noise, outliers and missing data – is not sufficiently captured. The excitation of the data during normal operating conditions is also simplistic, which would result in developed algorithms being subject to false alarms.

[1]https://depts.washington.edu/control/LARRY/TE/download.html

Another example of a benchmark data base is the UC Irvince Machine Learning Repository [Lichman, 2013]. This repository contains 416 datasets, each dataset with detailed metadata, including relevant papers and papers that cite the dataset. From Google Scholar, this repository and its website have been cited more than 5 000 times since 2007.

In this paper, we describe the underlying prerequisites, approaches and implementation issues for such a data repository. To the knowledge of the authors, no such repository accessible to everyone currently exists. Section 2 describes the nature of industrial process data and the issues that can occur when analysing the data. Section 3 outlines the key features of such a data repository including the file and data structures. When discussing the idea of a data repository with control engineering colleagues many suggestions to further resources that can be published on such a platform were proposed. Section 4 therefore gives the limitations of the first version of the repository and proposes possible extensions to the platform. Section 5 concludes the initiative of setting up such a database.

2. INDUSTRIAL PROCESS DATA

With the development of IT solutions, communication and computational capabilities, industrial process data has been under intense scrutiny. Buzz words such as ‘big data’, ‘data analytics’ and ‘data scientists’ have infiltrated automation vendors and production companies alike (Lee et al., 2014). The hype has to be taken with some caution because process data has been around since the widespread introduction of distributed control systems (DCS) in the 1980s and not much has changed in the use and storage of these systems.

The motivation to use process data analytics to improve process performance is simple: data analytics are non-invasive and comparatively cheap to use. Many problems occurring during operation manifest themselves in the time trend data of the process. For example, a malfunctioning, sticky flow valve controlling the stream in a heat exchanger often introduces nonlinear, discernable oscillations in the controlled temperature. Experts can look at process data and pick out – with some accuracy – specific problems. The aim of CPM is to automatically detect, identify and categorize common problems, which affect the operation of the plant.

This section describes the data formatting and origin in the process industries. Also, the most common mistakes that can occur when analysing plant data are discussed. The result of this experience is to focus on relevant and clearly distinguishable time trends, which are described towards the end of this section.

2.1 Data historians and format

Modern industrial plants capture the time series data generated by the controllers on the DCS, which in turn passes the data on to a plant historian, also known as data historian or operational historian. An overview of the system architecture is shown in Fig. 2.

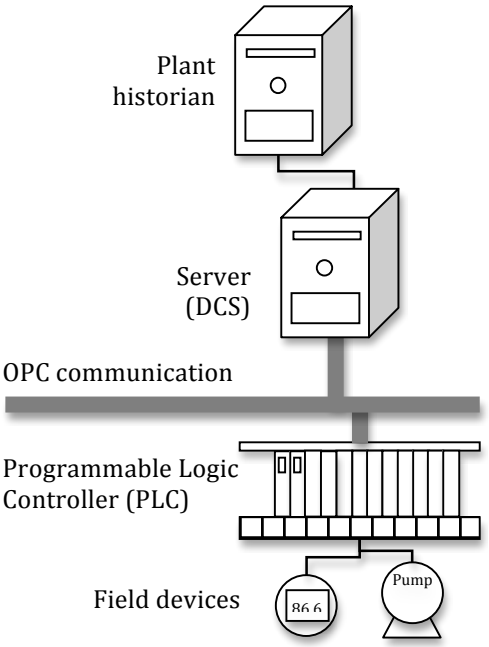

Fig. 2. System architecture from sensor to plant historian.

The plant historian stores the data in an SQL or similar database (SQL+ and Hadoop are some examples). The data can be displayed, analyzed and filtered by time selection as well as tag criteria more or less comfortably. Examples of common plant historians are the PI historian by OSIsoft, IP21 by Aspentech or 800xA history by ABB. These historians store the data with a timestamp and most commonly provide an export functionality that allows exporting the data in .csv format.

2.2 Pitfalls of data-based performance analysis

There are certain mistakes that are easy to make when analyzing plant data – for beginners as well as experts. This section describes the authors' experiences when dealing with the analysis of process data for industrial production plants. Often, this occurs when academics are approached by industry to assist with the detection and development of CPM algorithms. The authors believe that this experience is made by many control practitioners in similar situations.

Too much data The authors have experienced the following situation. A control engineer at an industrial production site approaches a researcher. The control engineer often has very little time at their hand because she is responsible a large number of control loops. She would be thrilled to get help on the most poorly performing loops in order to stabilize the process and improve efficiency. At a first meeting the control engineer hands over a complete data set of the plant, often with the comment: "How much data do you want?", meaning, are two months enough or would you like to have the whole two years.

This assumption is that the database can be entered into one very smart algorithm, which extracts the most commonly known problem. In fact, some commercial solutions promise exactly that. However, due to the nature of the different dynamic behaviour of production sites, the authors argue that with current algorithms it is at the present time impossible to get meaningful insights into the plant by looking at the entire data repository. It is crucial to know what to look for. Expert knowledge of an experienced engineer relating to the process and the measurements in question is currently always required to get meaningful results. The purpose of the data repository described in this paper is to work against the flood of data and provide data that has been processed and reviewed by experts.

False positives Just how difficulty it is to reliably detect process upsets becomes evident when examining real time trends in all its variations and exceptions. The assumption often made is that process data is always stationary and shows non-stationary characteristics can be seen from the sample data trends in Fig. 3. These time trends shown are all examples of normal process operations and are clearly fluctuating and in the case of the plot on the left shows a regular pattern. The trends exemplify that the time trend data can take on any form. It is important to test methods have to be tested against any variation to identify false positive faults. The data repository contains examples of these trends that are observed during normal operation but appear to be disturbances.

Time window selection Another difficulty is the selection of the start and end time of a disturbance. If a disturbance persists for longer it may be feasible to extend the period but for most methods it is helpful to exclude non-stationary data, that is, periods where the variable is fluctuating with low frequencies. This is because other plant upsets will temper with the analysis of most methods. Mostly it is a good idea to wait until the disturbance has established itself before capturing the data. Furthermore, some troublesome disturbances occur for a short period of time only. Data analysis algorithms often do not work on such a short selection so the data cannot be used for analysis.

2.3 Data categorization

There is an abundance of data generated in modern process plants. It is easy to get lost in the shear amount of data. Control engineers and ultimately CPM algorithms looking at the data need to focus on clearly discernible problems that can be detected in the time trends.

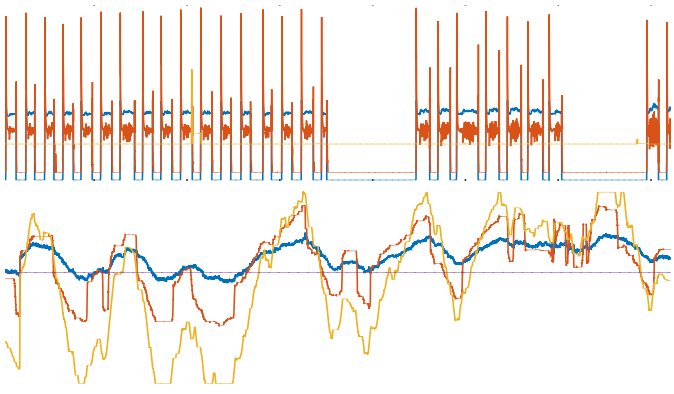

Fig. 3. Two examples of normal operating process data.

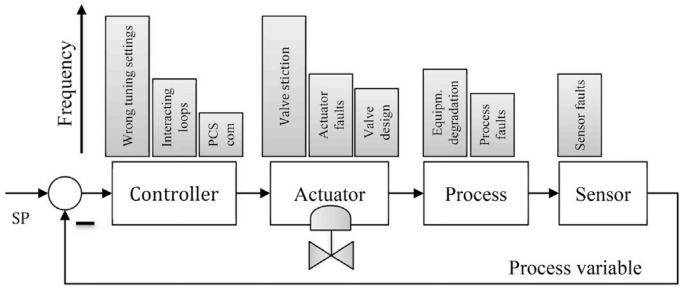

Fig. 4. Feedback control loop with fault categories from Bauer et al. (2016).

These problems have been identified by plant personnel and been recorded in the literature. Several overview articles have been published in recent years giving an overview of problems and methods alike. The fault categories and examples in the data repository described in this article are aligned with the findings in Bauer et al. (2016). Firstly, the faults can be sorted according to where in the control loop the fault originates. Fig. 4 is taken form the same article and shows a SISO feedback control loop with the key elements: controller, actuator, process and sensor. Something can go wrong with all of these elements and the most common faults are indicated on top of the element in Fig. 4. The height of the box indicates the frequency with which the faults are observed. According to the survey of Bauer et al. (2016) the most common problems are wrong tuning settings and valve stiction, followed by equipment degradation and sensor faults.

The faults listed in Fig. 4 often lead to deviation from the desired setpoint and manifest themselves in characteristic time trends. An expert can often identify valve stiction from the characteristic nonlinear oscillation. There are more than a dozen algorithms to detect valve stiction (Jelali, 2012). Other examples of data from which control engineering practitioners can identify faults are saturation where the controlled variable does not exceed a certain threshold because of physical constraints, sluggish or slow oscillation which are often due to poor tuning settings and quantisation issues in the sensor where the process variable is measured in intervals. Sinusoidal oscillations are common, too, and can a number of causes such as poor tuning, process upsets or external disturbances. Control loop monitoring methods are predominantly stochastic or databased and for these methods the correct data examples for detection is critical.

3 INDUSTRIAL DATA REPOSITORY

This website is a resource for industrial process data of PID loops. The data is contributed by industrial control practitioners as well as by academic researchers who work closely with industry. The aim of this repository is to provide a test environment for data that cannot be simulated as industrial data contains features that cannot be simulated.

The purpose of the repository is to (i) develop robust control loop performance monitoring methods and (ii) compare existing methods to find the best method for different data trends and situations.

This section describes the implementation platform as well as the file structure and the format the data is stored in. All data published on the platform has been screened and verified and this process is also described in this section.

3.1 Implementation platform

The data repository is hosted and maintained by the South African Council for Automation and Control (SACAC), which is the National Member Organisation of IFAC in South Africa. Like IFAC SACAC is a non-profit organization but supports the industry and in this instance the implementation, upkeep and maintenance of the repository. SACAC's website URL is:

http://sacac.org.za/

The industrial data repository resides under the tab 'Resources'. The SACAC website uses Wordpress as its Content Management System (CMS) and is maintained by Sven Uhlig from Studio UHU. Although all the repository files are stored in a Dropbox account, the 'Dropr' plugin for Wordpress provides access from the SACAC website to the files. Queries and contributions can be sent to the following email address: piddata@sacac.org.za.

3.2 File organization

The data is organised into the fault criteria described in Sec. 2.3. Thus, the ten fault categories are as follows:

1. Tuning settings e.g. aggressive or sluggish
2. Valve stiction
3. Saturation due to e.g. incorrect valve design
4. Actuator faults – other
5. Sensor faults
6. Process faults e.g. fouling
7. Interacting controllers
8. Communication problems
9. Other
10. Unknown

The underlined words are then taken up in the name of the file trend so that the name is a description of the trend. All data is stored in .csv format. The filename contains further information about the data set, namely the type of measurement (flow, temperature, pressure, level, other), the industry, the contributors surname and the year. For example, saturation-F-chemical-smith-2018.csv describes a time trend of a saturated flow controller in the chemical industry, the contributors surname is Smith and the data was captured in 2018. If there are several datasets with the same criterion these are numbered after the year.

It is particularly important to note that the authors and project participants have screened all the data to be published and have verified the root causes. Data trends that are 'suspected' fault categories are labelled accordingly. The data format in the .csv file is described in the next section. In addition to the .csv file there is a .txt file with the same filename that describes the data in more detail as follows:

```
File: quantisation-F-minerals-bauer-2017.csv
Type of measurement: Flow
Industry: Minerals processing
Data length: 1000
Sampling rate: 10 seconds
Company: Anonymous
Normalised: Yes
Contributor: Margret Bauer, University of the
Witwatersrand
Year of origin: 2017
First appearance in publication: none
Description: This is an example of quantisation in
a flow measurement. The data is normalised using a
larger data time period.
```

Tab. 1 lists the fields of the .txt file and the content with examples.

3.2 Data format of .csv file

The individual data sets are from single-input single-output PID control loops. These loops have three measurements which are captured as time series: process variable, desired setpoint and the controller output. These three measurements are captured as time trends which results in the first recorded measurement, the time stamp. Thus, each data set is in the form of a .csv file. For each file there are four columns, as listed in Tab. 2.

Table 1. Information and categories of data set.

Data information	**Values and examples**
Cause of poor performance	Any of the categories listed above
Type of measurement	Flow, Temperature, Pressure, Level or Other
Industry	Chemical, oil & gas including petrochemicals, minerals processing, paper, power generation, metals processing, manufacturing, food and beverages, other
Company	The company, which has provided the industrial data. If no information is put forward the default value is anonymous.
Normalisation	Yes/No. The data may be normalized to unit variance and standard deviation to protect the anonymity of the data.
Contributor	Name, surname and affiliation if available.
Year of origin	2017
Appears in publication	Authors, year, title, journal, volume, pages

Tab. 2. Columns of .csv file for each SISO PID control.

Time	Time	YYYY-MM-DD hh:mm:ss
Setpoint	SP	In same scale as PV
Process variable	PV	In same scale as SP
Controller Output	OP	Between 0 and 100

4 FURTHER ADDITIONS

When discussing the data repository within the research community many ideas for other uses of the data repository came to mind. For now, the data is strictly limited to industrial single-input single-output PID controller data. This is done in order to keep the data manageable, clear and establish a reputation for consistency and usability. The ideas for extensions and additions are captured in this section for future development and reference.

4.1 Plant-wide disturbances

Disturbances often travel through the connected process equipment and affect a number of process variables. The process variables then all fluctuate with the same pattern of the disturbance and deviate from their setpoints and it becomes difficult to identify the root cause. Finding and diagnosing the root cause of so called plant-wide disturbance is a specialized but important area of CPM. The datasets contain a number of measurements in an isolated section of a plant. These measurements do not necessarily have to be control loops but can be pure sensor data. In fact, for plant-wide disturbance analysis, only the process variables are usually investigated.

4.2 Industrial model predictive control data

Model predictive control (MPC) is the control layer on top of the PID loops and concerned with a optimizing the setpoints for a number of variables in a process. MPC data contains process model information – derived from first principles or step testing – as well as trajectories and is rich in information. The analysis of historical MPC data for performance evaluation is more complex than for PID and arguably less advanced as a result.

4.3 Simulated plant data

Similar to the Tennessee Eastman Process made available by Downs and Vogel, there are a number of simulated processes that are frequently used for fault detection. This starts with a simple continuous stirred tank reactor (CSTR) process and sometimes involves complex processes. Making the simulated data and the underlying models available to the research community provides different insights.

4.4 Monitoring methods

Besides sharing data, methods developed by academics can be shared on this platform. Many methods are complex and can be implemented differently. For example, any method that builds on examining the probability density function (PDF) from historical data requires a method to estimated the PDF. Estimation methods are simple histograms but ideally use Kernel functions. The choice of the Kernel function will impact – even if slightly – on the estimated PDF and therefore on the detection method. As a result, two implementations of the very same method will give different

results. Using one optimized implementation that is available on a platform allows the comparison of different methods on different data sets.

4.5 Evaluation criteria

Once the methods have been added another consideration is to cite a specific dataset and to include a brief summary of the performance of the proposed method. Defining performance metrics for control monitoring methods (a task not yet formally addressed in literature) will further promote relevant research. Some aspects that such performance metrics should cover:

- Level of user-input required (qualitative):
  - Whether the method requires user-defined parameters, how many parameters are required, and how robust performance is for incorrect parameter selection;
  - Computational performance (qualitative or quantitative):
- Processing requirements for training and application:
  - Memory requirements for training and application
  - Accuracy and precision (qualitative and quantitative);
- Ability of method to detect and identify control faults correctly (robustness).

5 CONCLUSIONS

This paper describes a voluntary initiative by academic researchers to further the use of data-driven methods for control loop performance monitoring. The host institution - South African Council for Automation and Control (SACAC) is a non-profit organisation with the interest of furthering the use of new technologies in automation and bridging the gap between industry and academia.

The authors firmly believe that the field of data-driven methods for CPM can be greatly enhanced by having comparable and standardised data sets that can be used for testing. As a result, the authors hope that production companies will be increasingly and consistently use CPM tools to tackle control loop performance problems.

As a first step, only SISO PID data of the most common fault categories are available on the platform. All colleagues are invited to submit their data sets and contribute to the discussion. The database is still in the process of being set up so there is room to improve, correct and optimize the data content and access.

REFERENCES

Bauer, M., Horch, A., Xie, L., Jelali, M., & Thornhill, N. (2016). The current state of control loop performance monitoring–A survey of application in industry. *Journal of Process Control*, 38, 1-10.

Choudhury, A. A. S., Shah, S. L., & Thornhill, N. F. (2008). *Diagnosis of process nonlinearities and valve stiction: data driven approaches.* Springer Science & Business Media.

Downs, J. J., & Vogel, E. F. (1993). A plant-wide industrial process control problem. *Computers & Chemical Engineering*, 17(3), 245-255.

Huang, B., & Shah, S. L. (1999). *Control loop performance assessment: Theory and applications*. Springer Science & Business Media.

Jelali, M., & Huang, B. (Eds.). (2009). *Detection and diagnosis of stiction in control loops: state of the art and advanced methods*. Springer Science & Business Media.

Jelali, M. (2012). *Control performance management in industrial automation: assessment, diagnosis and improvement of control loop performance*. Springer Science & Business Media.

Lee, J., Kao, H. A., & Yang, S. (2014). Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia Cirp*, 16, 3-8.

Ordys, A., Uduehi, D., & Johnson, M. A. (Eds.). (2007). *Process control performance assessment: from theory to implementation*. Springer Science & Business Media.

Vilanova, R., & Visioli, A. (2012). *PID control in the third millennium*. London: Springer.

Åström, K. J., & Hägglund, T. (1995). *PID controllers: theory, design, and tuning (Vol. 2)*. Research Triangle Park, NC: Instrument society of America.