Computers and Chemical Engineering 179 (2023) 108402

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Online data condensation for digitalised biopharmaceutical processes

Nishanthi Gangadharan [a], Ayca Cankorur-Cetinkaya [b], Matthew Cheeks [b], Alexander F Routh [a,c], Duygu Dikicioglu [a,d,*]

[a] Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge CB3 0AS, UK
[b] Cell Culture and Fermentation Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK
[c] Institute of Energy and Environmental Flows, University of Cambridge, Cambridge CB3 0EZ, UK
[d] Department of Biochemical Engineering, University College London, London WC1E 6BT, UK

ARTICLE INFO

Keywords:
Bioprocess control
Online monitoring
Data condensation
Interruption in bioprocess monitoring
L_p^{TS} norm

ABSTRACT

Efficient control of a bioprocess relies on the ability to systematically capture and represent the process dynamics of critical process parameters. Multivariate monitoring techniques in biopharmaceuticals has resulted in the generation of large amounts of data comprising real-time measurements of critical quality and performance attributes. If exploited efficiently, these can provide an opportunity for developing better control action. For this, it is important to have a comprehensive view of the critical process parameter landscape, which can only be achieved by integrating both online and offline data into a single data matrix that can then be subjected to standard data analysis protocols. However, owing to the difference in the number of readings available for variables recorded online and offline, there is a need for new methods to achieve condensation capability. This paper introduces a novel methodology for condensing online data into an offline data matrix, which performed better when compared to traditionally employed averaging and helped increase the number of variables available for representing the design space of the process. The method was also used to understand how error propagates through online data, so as to identify an interval of tolerance in online monitoring of bioprocesses.

1. Introduction

Online monitoring is considered a fundamental requirement for obtaining real-time information from bioprocesses that can provide direct insights into process states and offer early problem detection. Data from these real-time monitoring platforms have been useful in the implementation of setpoint control for parameters such as temperature and pH that need to be tightly regulated within a specific range for quality assurance (Reyes et al., 2022). In addition, this data has the potential to bring in a wealth of information into retrospective analysis by increasing the feature space, thereby aiding in the generation of model-based control systems. Historic bioprocess data originate from multiple sources – online and offline, due to which there is a disparity in the number of readings available per day for different parameters. For instance, a parameter such as the glutamine concentration may be measured and recorded offline at regular sampling intervals resulting in a single reading per day, whereas a parameter for which continuous online readings are available, such as pH, may have hundreds of entries recorded on a daily basis.

There are efficient methods available for data compression, not to be confused with condensation, which involve modifying the data to reduce its size by re-encoding information using fewer bits than the original representation, and pattern recognition from online data using signal processing techniques such as wavelet transforms, discrete Fourier transformation, piecewise linear approximation, adaptive piecewise constant approximation (Bakshi and Stephanopoulos, 1994a, 1994b; Charaniya et al., 2008), which treat the readings as signals. These approaches fail to offer data condensation capabilities. Treating online data as signals can lead to the comparison of identical time points that may not represent similar process states, as different runs may not be temporally aligned (Charaniya et al., 2008; Huang et al., 2002; Kamimura et al., 2000). Time series from different runs may not be temporally aligned due to the occurrence of lag phase and/or variations in growth rate. A single value that captures the behaviour of online data from a time range, corresponding to the offline reading, which can then be integrated into a data matrix to be used alongside offline data is necessary for retrospective analysis, and this is the aim of data condensation.

* Corresponding author.
E-mail address: d.dikicioglu@ucl.ac.uk (D. Dikicioglu).

https://doi.org/10.1016/j.compchemeng.2023.108402
Received 12 June 2023; Received in revised form 31 August 2023; Accepted 31 August 2023
Available online 3 September 2023

One of the primary considerations while performing online data condensation, with the intention of using it for bioprocess control, is the ability to retain the identity of each individual parameter before and after condensation. This is essential to ensure that proposed control actions can later be traced back to unique parameters, which can then be achieved using conventional controllers installed in a bioreactor setting. Therefore techniques that convert parameters into latent forms, such as multi-way PLS (Nomikos and MacGregor, 1995), which might be able to offer some condensation capability, are not considered appropriate for this function. Neither does such an application exist in literature.

In literature, the only instances of condensation of data for online parameter values from bioprocesses was achieved using a moving window average method for a particular interval (Charaniya et al., 2010; Le et al., 2012). However, averaging is known to have several limitations: A central value fails to give any idea about the formation of the series. Two or more series may have the same central value but may differ widely in composition, structure and constitution. Averaging fails to represent bimodal, U-shaped or J-shaped distributions accurately. Also, averaging is highly sensitive to extreme values, and works well only when all values are equally important. When the variance is high, an average is a poor representative of the population.

To make the right inferences from data, it is important to capture the behaviour of a series. A moving average is known to smooth out variations, and fluctuations from normal behaviour can go undetected when using averaging for data condensation. From a data analysis perspective such fluctuations are of interest because they represent areas of unexpected activity that influence the behaviour and outcome of a bioprocess. Methods such as an exponential moving average, which is a weighted average that gives more importance to recent values, are not appropriate as it might downplay the influence of fluctuations occurring early on in the culture. This emphasizes the need for an alternate methodology for data condensation of time series datasets, such as those arising from online monitoring platforms in bioprocesses.

An L_p^{TS} norm was proposed as a simple distance measure that is a generalisation of the L_p norm (Minkowski norm, a metric in a normed vector space that is a generalisation of both Euclidean and Manhattan distance) for time-dependant vectors. It takes the temporal structure of sequence into account for application in the context of self-organising maps (Lee and Verleysen, 2005). However, the initial phase of this method, which calculates the area underneath the time-series function, has untapped potential that can be exploited for data condensation. The properties of this method in the context of condensation have not yet been rigorously tested against averaging. In this paper we evaluate the performance of the proposed method over existing techniques to identify the advantages of using it for the purpose of data condensation.

Although biopharmaceutical processes are tightly monitored throughout their course, errors arising from different sources such as hardware, software or manual are bound to happen and can have a detrimental impact on the product quantity and quality (Pekarsky et al., 2019). To evaluate the impact of such errors they need to be captured and represented appropriately in the data matrix, for retrospective analysis. As online data has a continuous record of system states, any such errors are efficiently recorded. However, interruption in online monitoring is not uncommon and can be caused by instances of sensor breakdown, power failure, software issues etc. An overlap of such an interruption in online monitoring with technical failure or process faults can often lead to process termination or batch loss, due to insufficient knowledge of the impact of such an event on product quality or even as a precautionary measure to avoid risk (Pekarsky et al., 2019). Therefore, it becomes important to computationally evaluate the impact of different types of error present in different percentages in online data, and how it propagates through the data matrix during condensation. Through such an analysis it would be possible to identify an interval of tolerance during which quality can still be assured in cases of an unexpected interruption in online measurements. In this paper, properties of L_p^{TS} norm in the context of data condensation and how it compares with averaging are evaluated. We also evaluate the impact of error on online data to identify a window of tolerance for errors that occur during manufacturing.

2. Theory/calculation

2.1. Comparison of averaging vs L_p^{TS} norm using synthetic data

When variance is high, an average is a poor representative of the population. This is a likely scenario in any long-term continuous monitoring setting, such as those observed in bioprocess operations. Condensing the values using averaging, limits the ability to capture variability, resulting in detrimental fluctuations going undetected. Capturing and incorporating such fluctuations into the data matrix is an important step to ensure sensible models that are true to the process. The L_p^{TS} norm methodology takes into account the temporal structure of the sequence by involving the previous and next value of x_i in the i^{th} term of the sum, instead of x_i alone. Assuming a constant sample period τ, the proposed norm is given by (Lee and Verleysen, 2005),

$$L_p^{TS} = \left(\sum_{i=1}^{D} (A_{i-1} + A_{i+1})^p \right)^{\frac{1}{p}} \tag{2.1}$$

where,

$$A_{i-1} = \begin{cases} \tau/2 \ |x_i|, \ \ if \ x_i x_{i-1} \geq 0 \\ \tau/2 \dfrac{x_i^2}{|x_i| + |x_{i-1}|}, \ \ if \ x_i x_{i-1} < 0 \end{cases} \tag{2.2}$$

and,

$$A_{i+1} = \begin{cases} \tau/2 \ |x_i|, \ \ if \ x_i x_{i+1} \geq 0 \\ \tau/2 \dfrac{x_i^2}{|x_i| + |x_{i+1}|}, \ \ if \ x_i x_{i+1} < 0 \end{cases} \tag{2.3}$$

The variable p is assumed to be a positive integer. This can provide us with a single condensed value for the series, which can then be incorporated into the data matrix to use alongside offline variables for retrospective analysis (Fig. 1).

To demonstrate how variability in the dataset can go undetected while using averaging, a test was designed using synthetic datasets. These were generated to mimic tightly controlled parameters such as pH and parameters that were not so tightly controlled such as dissolved oxygen. Their averages and L_p^{TS} norm values were compared along with their temporal profiles, to demonstrate how averaging fails to capture fluctuations.

2.2. Considerations for data condensation

Chinese Hamster Ovary (CHO) cell cultures in fed batches were monitored for a period of 14–16 days and therefore a large amount of online data was accumulated for each day of the culture. At a rate of one reading per six minutes, each parameter had 240 values per day. However, considering the slow rate of growth of mammalian cells especially compared to other cell types, such as bacteria, it is quite likely to see extended periods of static parameter readings, unless instigated by an unexpected fluctuation. Parameters such as pH were tightly regulated to be within a specific range, meaning that the online data typically had prolonged periods of steady activity. As the aim of data condensation is to capture regions of distinct fluctuations in parameters that can be unique to each culture, it becomes important to identify an interval that is appropriate for condensation. In this exercise, online data was divided into different intervals before condensation to determine whether an interval that was generalisable for CHO cell culture online parameter readings existed.

Prior to employing condensed data for retrospective analysis, it is

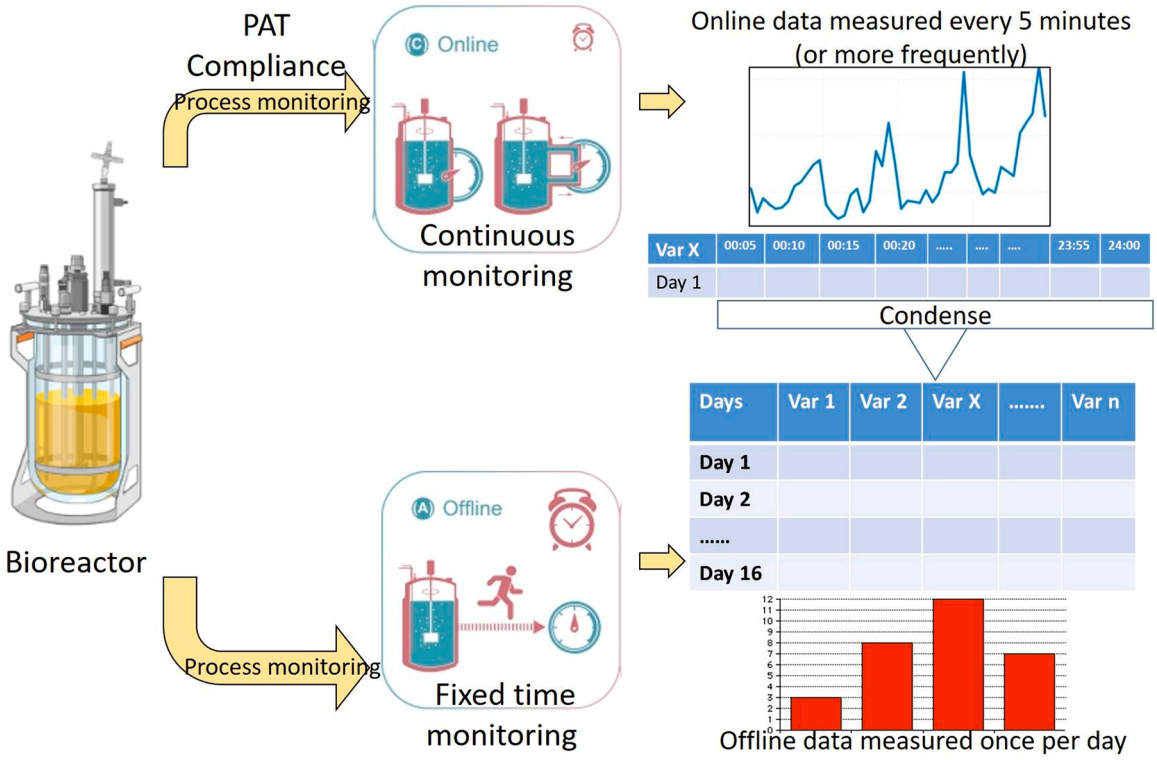


Fig. 1. Online data condensation in biopharmaceutical manufacturing. Parameter 'Var X' is monitored online and recorded every 5 min for the entire duration of the culture, which is then condensed to one value per day to fit into the data matrix alongside other parameters monitored offline. The monitoring intervals described above are just an example, not a rule. Diagram of bioreactor was created using BioRender, and the figures for continuous and fixed time monitoring were adapted from (Schlembach et al., 2021).

imperative to ensure that the methodology used for condensation did not alter parameter relationships. Correlations between parameters are known to change on different days of the culture (Gangadharan et al., 2021). A comparison of correlations on each day of the culture while employing averaging and L_p^{TS} norm was used to reveal any potential bias introduced by the data condensation methodology. Ideally, on a given culture day the parameter relationships should remain unchanged by the data condensation method.

2.3. Impact-of-error analysis

Unexpected fluctuations can have large impact on the outcome of a culture. However, evaluating this using data is quite difficult, as information from failed cultures are often discarded. Therefore, one often relies on simulated errors to study their effect. If the condensed value reflects the error, it indicates the impact the change has on the overall behaviour of a parameter. This provides us with an opportunity to identify an interval of tolerance for errors in online data. It also doubles up as a test to identify which condensation method is best at capturing fluctuations in online data. To demonstrate this, errors of increasing complexity, from logarithmic to exponential were introduced into the online data at different percentages starting from 1.25% up to 10%. The error,ε_i, was introduced as a function of the observed value, y_i, at time point *i*.

If,

$$y_i = f\left(x_i, z_i m_i, \ldots q_i\right) \tag{2.4}$$

where error function,

$$\varepsilon_i = g(y_i) \tag{2.5}$$

then new function becomes,

$$\hat{y}_i = y_i + \varepsilon_i \tag{2.6}$$

The observed value y_i was influenced by multiple parameters. The new function, $\hat{y}_i$ was then condensed using averaging and the L_p^{TS} norm and compared with condensed values of y_i using respective methods to evaluate the impact of the error on online data.

3. Materials and methods

3.1. Nature of the data

Data for this analysis was procured from an online data management system at AstraZeneca and comprised of continuous readings of eight parameters from 22 cultures of antibody producing CHO cells. Online parameter readings of different cultures were recorded at different frequencies – either every minute or every six minutes. In order to avoid a frequency mismatch, parameter readings collected at 6-minute intervals were employed in this study. Eight parameters: DO.Out, DO.PV, F.PV, FO_2.PV, pH.Out, T.Out, XCO_2.PV, XO_2.PV, where, DO.Out and DO.PV are the measurements of Dissolved Oxygen controller response and process value respectively, at a specific time point; F.PV is the Total Gas Flow; FO_2.PV is the Flow Rate of Oxygen; pH.Out is the measurement of pH; T.Out is the measure of Temperature; XCO_2.PV is the Percentage of CO_2 in total gas flow; XO_2.PV is the percentage of oxygen in the total gas flow (where, X_i.Out is the response of the controller to the change in a particular parameter X_i and X_i.PV is the actual value of a process variable X_i), which had continuous readings for all 14 days of the culture, were used. Gaps were filled by using the Last Observation Carried Forward (LOCF) method (Heyting et al., 1992), as observations suggested values were recorded only if there was a change.

3.2. Comparison of averaging vs L_p^{TS} norm using synthetic data

Data was generated using a random number generator *rnorm()* function in the *stats* package of R (R Core Team, 2020). To mimic a tightly controlled variable such as pH, three different series of twenty values each with a mean of 7.5 and different standard deviations (0.5, 2 and 5) were generated. Their means and L_p^{TS} norm values were calculated. L_p^{TS} was implemented as shown in Eq. (2.1), (2.2) and (2.3) in MATLAB R2019b, using $p = 2$, (Euclidean norm, since it corresponds to the intuitive notion of length and distance in a three-dimensional world). As the L_p^{TS} norm calculates the area underneath the parameter vs time curve, the output is usually a number much higher than observed for any given parameter, when the average of the values that constitute the function were taken. In order to maintain comparable magnitudes between the numbers obtained from averaging and L_p^{TS} norm, the condensed value from L_p^{TS} norm was divided by the number of intervals. The resulting structure of the series and their L_p^{TS} norm values were evaluated. A second test case was designed with three series of 100 values each, using the same random number generator, with a mean value of 150 and standard deviations of either 5, 25 or 60. The resulting structure of the series and L_p^{TS} norm values were evaluated, with $p = 2$.

3.3. Online data condensation

Parameter behaviour was visualised using the *mvtsplot* package in R (Peng, 2008) for each day of the culture, in order to identify any occurrence of inflection points, so that condensation can be limited to a specific interval of interest. In instances where a 24-hour interval did not display inflections, four equal 6-hour intervals (i.e., quarters) denoted by Q1, Q2, Q3, and Q4, in addition to the 24-hour interval, were employed for condensation. An internal normalisation option was selected with all other settings kept at their default values for visualisation. Values from each interval were condensed using averaging and the L_p^{TS} norm methods (A simple MATLAB code is provided in the supplementary). The significant differences between the two approaches in condensing the data were evaluated using the Student's *t*-test. This was bootstrapped ($n = 2000$) using the *boot.t.test* function in the *MKinfer* package in R (Kohl, 2022), to ensure that potential non-normality of the data does not impair the test validity because the Student's *t*-test assumes that a variable in question is normally distributed. If the sample size is moderately large, a two-sample *t*-test is robust to non-normality due to the central limit theorem (Kim and Park, 2019; Kwak and Kim, 2017; Lumley et al., 2002). This means that if the sample size is not too small and the distribution is not extremely skewed, a *t*-test can yield satisfactory results. A correlation matrix was generated using the *findCorrelation* function in the *caret* package in R (Max Kuhn Contributions from Jed Wing et al., 2019), for each day of the culture as described in (Gangadharan et al., 2021) using condensed values generated by either method. In this manner parameter dependencies were evaluated.

3.4. Impact-of-error analysis

Five different types of error (logarithmic, linear, quadratic, cubic, and exponential), were introduced at different percentages (1.25%, 2.5%, 5%, 10%) into online readings. The error functions were given by:

Logarithmic

$$\varepsilon_i = \log y_i \tag{2.7}$$

Linear

$$\varepsilon_i = 0.2(y_i) + 3 \tag{2.8}$$

Quadratic

$$\varepsilon_i = 0.5\left(y_i^2\right) + 3(y_i) - 6 \tag{2.9}$$

Cubic

$$\varepsilon_i = 0.06\left(y_i^3\right) - 2 \tag{2.10}$$

Exponential

$$\varepsilon_i = 0.005(e^{y_i}) \tag{2.11}$$

where, y_i is the observed value of a parameter at time point *i*. The error function ε_i was then added to the observed values to introduce 1.25%, 2.5%, 5% or 10% error. The values were then condensed using averaging and L_p^{TS} norm methods. The condensed values obtained from original and error-induced datasets were compared with each other. A Bootstrapped Welch two sample *t*-test for samples of unequal variance were performed to identify if there is a significant difference in condensed value when an error is introduced (Fig. 2). All the 'NaN' and 'Inf' values, representing measurement or sensor errors during experimentation, which were introduced into condensed values due to error addition, were replaced with zeros before the *t*-test. Bootstrapping ($n = 2000$) was achieved using the *boot.t.test* function in the *MKinfer* package in R (Kohl, 2022). The rest of the analysis was implemented using MATLAB R2021b.

4. Results and discussion

4.1. Comparison of averaging vs L_p^{TS} norm using synthetic data

Before employing an L_p^{TS} norm for data condensation, it was necessary to compare its properties with regularly employed averaging. To replicate a scenario where variance is high, two different test cases were designed, one for parameters that are tightly controlled and one for those that are not. Series A1, A2 and A3 were designed to replicate the data profiles of a tightly controlled parameter such as pH, with a mean value of 7.5, but varying standard deviations (SD = 0.5, 2 and 5) (Fig. 3a). Although the average for these three series were very similar, due to the differences in SD the distribution of values were quite different (Supplementary Figures S1a-c). While averaging failed to capture such subtle differences in variance, the L_p^{TS} norm successfully captured differences between the three (Fig. 3b). If the average of the series were to be used for condensation, parameter behaviour of all three cultures would be perceived to be the same, whereas they were not. The L_p^{TS} norm makes capturing such variances possible, thereby bringing more accurate information into the data matrix for retrospective analysis. Series B1, B2 and B3 were designed to mimic a less tightly controlled parameter (e.g. Na^+) with a mean of 150, but different standard deviations (5, 25 and 60) (Fig. 3c). The series have different composition although their mean values were similar as shown in Supplementary Figures S1d-f. The L_p^{TS} norm was once again successful in capturing such differences (Fig. 3d). This analysis clearly demonstrated that the L_p^{TS} norm has the ability to capture subtle differences in parameter behaviour, as compared to averaging, which is important for representing relationships accurately in the data matrix used for retrospective analysis.

4.2. Identifying the best interval for condensation and online data condensation

The interval to which the condensation is confined needs to be determined as a prerequisite to the process, in order to use it as representative of the parameter behaviour at any given period. Matching the condensation interval to the frequency of offline reading, which usually happens once every 24 h, is one option. The question remains whether an online parameter exhibited significant variation throughout the entire interval. If the parameter profile exhibited no significant variation during a major fraction of the interval, a rational approach would be to focus on a time frame which represented an observable variation. For this purpose, inflection points in the data were identified. Visualisation of the data did not suggest the presence of any distinctive regions

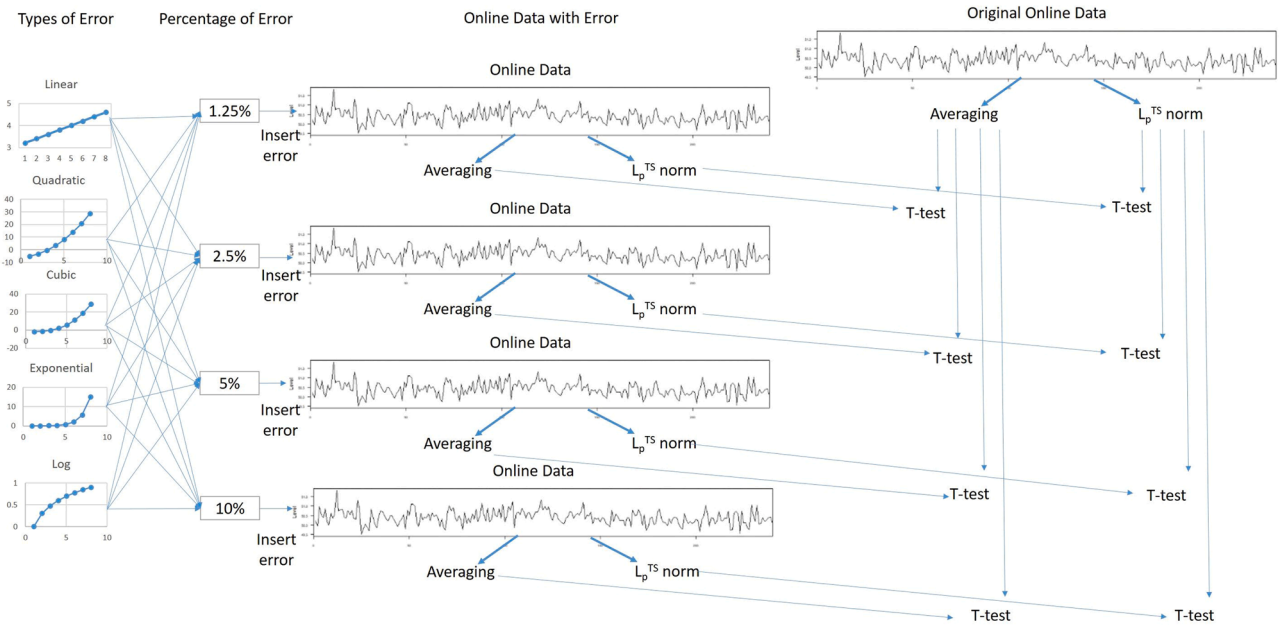


Fig. 2. Diagrammatic representation of Impact-of-error analysis. Different types of errors: log, linear, quadratic, cubic, and exponential, were introduced at 1.25%, 2.5%, 5% and 10%, and the difference between original and error induced values while condensing using averaging and L_p^{TS} norm were compared using a two-sample *t*-test, bootstrapped at $n = 2000$.

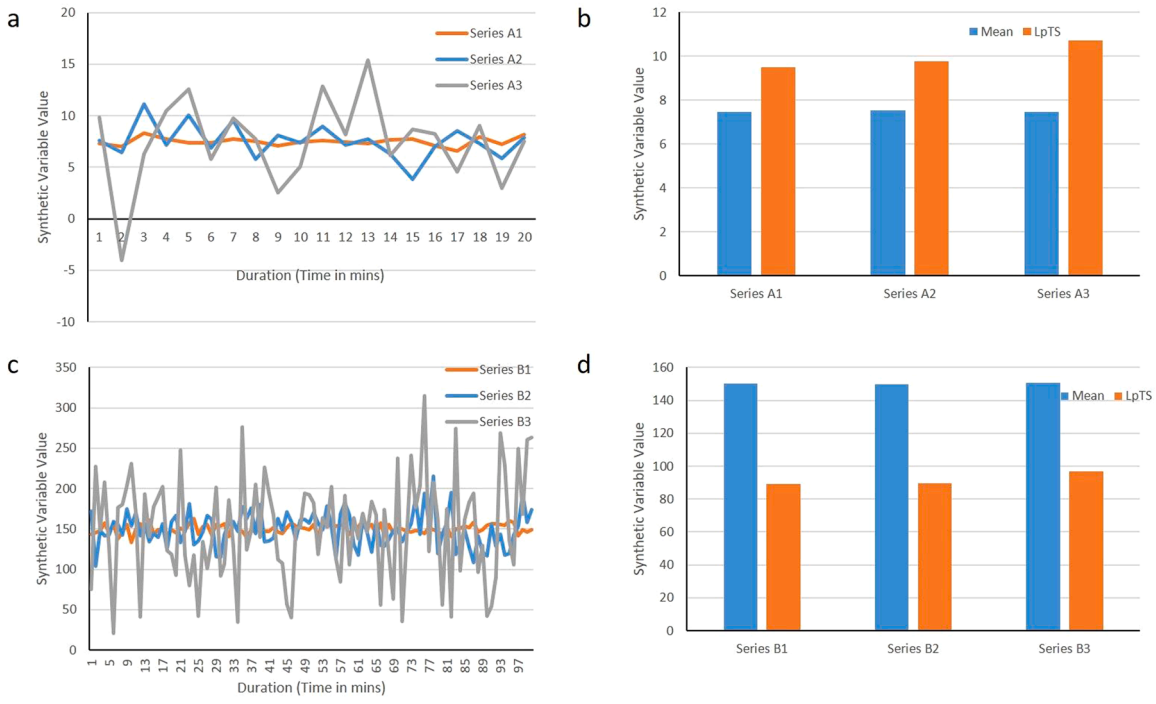
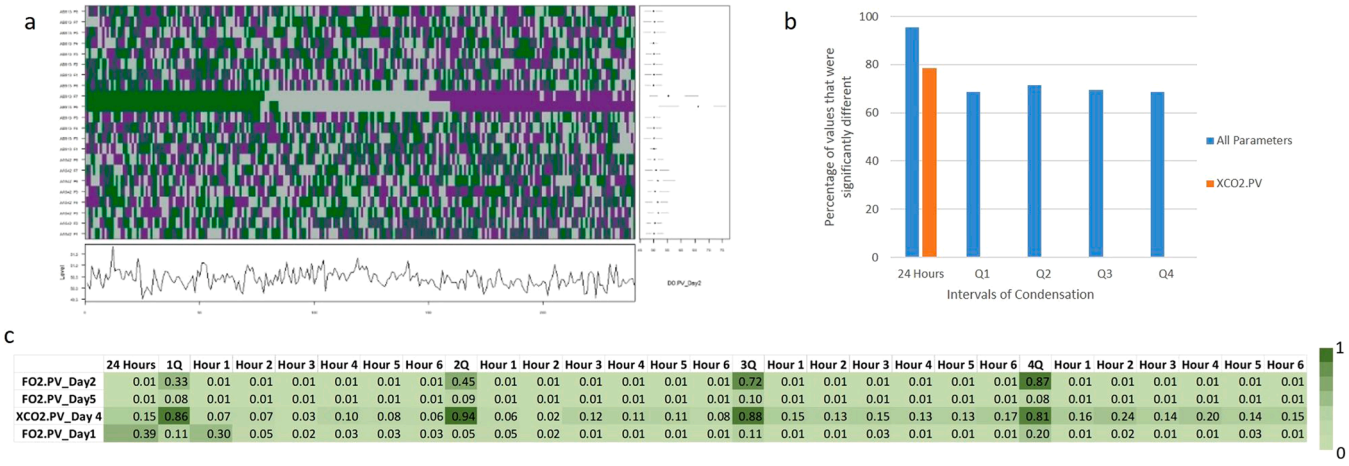


Fig. 3. Synthetic data generated to mimic parameters that are tightly controlled (a) Series A1, A2 and A3 with same mean, but different variances. x axis shows duration and y axis shows the value of the synthetic variable. Orange represents series A1, blue represents series A2 and grey represents series A3. (b) Comparison of average vs L_p^{TS} norm. x axis shows the different series A1, A2 and A3 and y axis shows the condensed synthetic variable value. Blue and orange bars represent condensed values obtained while using mean and L_p^{TS} norm respectively. (c) Synthetic data generated to mimic parameters that are not so tightly controlled. Series B1, B2 and B3 with same mean, but different variances. x axis shows the duration and y axis shows the value of the synthetic variable. Colours orange, blue and grey represents series B1, B2 and B3 respectively. (d) Comparison of average vs L_p^{TS} norm. x axis shows series B1, B2 and B3 and y axis shows the condensed synthetic variable values. Colours blue and orange represents condensed values obtained while employing mean and L_p^{TS} norm respectively.

represented by inflection points for the investigated parameters on any day of the culture. Instead, parameter fluctuations were distributed across the duration of the interval (Fig. 4a, Supplementary Figures S2-S15). Based on this observation, condensation was performed on data collected for each parameter at 24-hour intervals. Additionally, the daily collected data were further condensed into four equal 6-hour intervals; Q1, Q2, Q3, and Q4 in order to investigate the possibility of using smaller intervals to assist the interpretation of how the condensation



c

	24 Hours	1Q	Hour 1	Hour 2	Hour 3	Hour 4	Hour 5	Hour 6	2Q	Hour 1	Hour 2	Hour 3	Hour 4	Hour 5	Hour 6	3Q	Hour 1	Hour 2	Hour 3	Hour 4	Hour 5	Hour 6	4Q	Hour 1	Hour 2	Hour 3	Hour 4	Hour 5	Hour 6
FO2.PV_Day2	0.01	0.33	0.01	0.01	0.01	0.01	0.01	0.01	0.45	0.01	0.01	0.01	0.01	0.01	0.01	0.72	0.01	0.01	0.01	0.01	0.01	0.01	0.87	0.01	0.01	0.01	0.01	0.01	0.01
FO2.PV_Day5	0.01	0.08	0.01	0.01	0.01	0.01	0.01	0.01	0.09	0.01	0.01	0.01	0.01	0.01	0.01	0.10	0.01	0.01	0.01	0.01	0.01	0.01	0.08	0.01	0.01	0.01	0.01	0.01	0.01
XCO2.PV_Day 4	0.15	0.86	0.07	0.07	0.03	0.10	0.08	0.06	0.94	0.06	0.02	0.12	0.11	0.11	0.08	0.88	0.15	0.13	0.15	0.13	0.13	0.17	0.81	0.16	0.24	0.14	0.20	0.14	0.15
FO2.PV_Day1	0.39	0.11	0.30	0.05	0.02	0.03	0.03	0.03	0.05	0.05	0.02	0.01	0.01	0.01	0.01	0.11	0.01	0.01	0.03	0.01	0.01	0.01	0.20	0.01	0.02	0.01	0.01	0.03	0.01

Fig. 4. Evaluation of the performance of L_p^{TS} norm method for online data condensation. (a) Heat map generated using mvtsplot showing parameter fluctuations distributed across an example 24-hour interval for DO.PV. Here Day 2 data is presented. Heat maps for all the other days and online parameters are provided in the Supplementary figures S2-S15. The x-axis represents time, and the y-axis shows the different cultures. Purple, grey and green denote low, medium and high values within a culture, respectively. Colours are not comparable across cultures. The box plot on the right-hand side panel presents the data in each time-series and the one on the bottom panel displays the median values across all time-series for each time point. (b) Performance of condensation on intervals of different sizes for all variables (blue) and parameter XCO_2.PV (orange). x axis shows the percentage of values that were significantly different (p-value<0.05) while comparing the condensed values from averaging method and L_p^{TS} norm method using *t*-test, and y axis shows the different intervals of condensation. The distribution of p-values from each test case is available in supplementary figure S16. (c) Heat map showing the p-values obtained from *t*-test on hourly intervals of arbitrarily selected set of parameters for which 24-hour condensation resulted in significant and non-significant differences while the constituent quarters did not. Panel on the right-hand side shows the gradient of the heat map.

methodology handles the data.

The data condensed by averaging and using the L_p^{TS} norm were compared with similar results obtained from bootstrapped and non-bootstrapped t-tests. The condensed values calculated by the two methods were significantly different in more than 95% of the cases for a 24-hour interval (Fig. 4b, Supplementary Table S1 and Figure S16a). Consequently, the effect of condensation interval size on the algorithm's ability to capture data variations was evaluated for all eight parameters on all 14 days of the 22 cultures, focusing on shorter intervals during the day. Significant differences were observed in fewer cases when 6-hour intervals were employed (68.75%, 71.43%, 69.64%, 68.75% of the cases for Q1, Q2, Q3, and Q4, respectively) (Fig. 4b, Supplementary Table S2 and Figures S16b-e). This suggests that at longer time intervals relevant in an industrial setting, such as 24-hours, an averaging method, which smooths inherent variation risks losing information embedded within the data. This is problematic for subsequent data pre and post-processing steps leading to the identification of possible control actions that could be suggested.

The above observation raises questions as to why this variation in behaviour existed between condensed values computed from different intervals. In order to avoid ambiguity in selecting the condensation interval, an interesting observation within the sample set was investigated. One such instance was that of the parameter 'percentage of CO_2 in total gas flow' (XCO_2.PV), where the 24-hour condensation values calculated by averaging and L_p^{TS} norm yielded significant differences for most days of the cultivation, whereas the differences were not significant for each of Q1, Q2, Q3, and Q4 on any of the 14 days of cultivation (Fig. 4b, Supplementary Table S3 and Figures S16f-j). Understanding this is key to explaining how the selection of a condensation interval affects the ability of the methodology to effectively capture trends. This observation was explored by investigating one-hour intervals at random for those instances where the 24-hour condensation yielded significant (and non-significant) differences between the L_p^{TS} norm and averaging values, while the comparisons made for the constituent quarters did not. Interestingly, condensation at hourly intervals resulted in a greater number of significantly different comparisons than it did for the analysis of 6-hour intervals (Fig. 4c).

This can be explained using the mean value theorem for definite integrals. This theorem states that, if $y = f(x)$ is a continuous function on the closed interval $[a, b]$, there is at least one point c in the interval $[a, b]$ where $f(x)$ attains its average value $\bar{f}$:

given by,

$$f(c) = \bar{f} = \frac{1}{(b-a)} \int_a^b f(x).dx \tag{2.12}$$

Geometrically, this means that there is a rectangle whose area exactly represents the area of the region under the curve $y = f(x)$. The value of $f(c)$ represents the height of the rectangle and the difference $(b - a)$ represents the width. If these values were employed to calculate the area of the function, the value per interval would be $f(c)$ which is close to the value calculated from the L_p^{TS} norm (i.e., Area of the function / Number of intervals). This means that if the average of a set of numbers were equal to the average of the function describing that set of numbers, then there would be no significant difference between the condensation values calculated by the L_p^{TS} norm and by averaging. In order for the values obtained from the average of a set of values and the average of its function to be similar, a function should exist such that the set of values have very little or no outliers. In this analysis, the intervals for which no significant differences were observed between the average value and the L_p^{TS} norm value, were those where the set of values could be perfectly described by a function. Ordinarily this would not be the case as it would be difficult to fit a parameter's behaviour into a perfectly describable function when many parameters are interacting non-linearly, such as in a bioprocess. However, the intervals that exhibit such a perfect fit are likely to be those where parameter values display linear behaviour, because of less influence from other parameters. This could depend on parameter interactions on each day and other factors that are likely not known. Therefore, it is necessary to adopt a method that is superior in capturing the function behaviour, such as L_p^{TS} norm as opposed to averaging.

This is not merely the identification of an interval of condensation, where a function-based computation would yield a different value than that of averaging, but one that represents the true function i.e., data behaviour. In this example, it was observed that the behaviour was captured better during condensation in hourly than 6-hourly intervals.

However, there are other factors contributing to the interval selection strategy for data condensation. In the event that the condensed data were to be merged with existing offline data recordings for further data processing, the condensation would be carried out in intervals to match the offline data record timings. In this dataset, 24-hour intervals, which matched with the offline sampling frequency, also captured behaviour sufficiently well and did not require any compromises. If encountered with instances where the 24-hour interval fails to capture the behaviour of a series well, condensation could be restricted to smaller intervals of observable activity.

4.3. Implication of data condensation on bioprocess dataset

The implication of adopting different data condensation strategies on upstream bioprocess knowledge was investigated. Using a suitable condensation strategy allowed differentiation between different cultures. For instance, the temporal profiles of one example parameter (DO.Out on Day 1) for two different cultures (Fig. 5a) yielded the same average value (0.41), whereas the L_p^{TS} norm could distinguish the differences in the two separate cultures based on the condensation values of 0.42 and 0.48. This has important implications in retrospective data analysis. Utilising averaging would enforce similarity between two cultures despite their differences, as shown by their profiles. Furthermore, using a suitable condensation strategy eliminated the risk of two different parameters, which had different behaviour, being assigned the same condensed value. For example, a comparison of two different parameters DO.Out on Day 1 and pH.Out on Day 12 from two different cultures yielded the same average value (0.72) despite their temporal profiles suggesting otherwise, whereas the L_p^{TS} norm could distinguish between the two (0.62 and 0.30) (Fig. 5b). This also has important implications in data analysis as the similarity of temporal behaviour of parameters across cultures are frequently used to determine how different cultures behave. In short, the L_p^{TS} norm as a methodology, respects the temporal data profiles while condensing values and provides a more reliable single value as representation of its profile, which is essential for extracting accurate information out of data mining.

Since the L_p^{TS} norm is more likely to introduce numerical variability, it is important to ensure that the methodology does not have any impact on the way in which analysis would be conducted leading to misinterpretation of the results. One of the most important aspects of data mining is identification of the relationships between parameters, and correlation analysis provides a useful measure of this (Charaniya et al., 2008; Gangadharan et al., 2021). In order to explore this notion, the differences in correlations between parameters, whose values were condensed using averaging or the L_p^{TS} norm were compared. Results suggested that the differences introduced to the data, and subsequently into the correlation analysis were not substantial when the L_p^{TS} norm was employed for data condensation, negating the possibility of any undesirable impact on parameter behaviour (Fig. 6a-b, Supplementary Figures S17-S23). Despite the similarities between the correlation values obtained when either the L_p^{TS} norm or averaging was employed for data condensation, the ability of the L_p^{TS} norm to describe the inherent behaviour of the data becomes important once these values are embedded into the master data matrix.

4.4. Impact-of-error analysis

To evaluate the impact of error on online data, errors of different types were introduced at different percentages into the original dataset for all the available parameters. Depending on the nature of the parameter, the response to errors at different percentages were found to vary. Low magnitude errors such as those arising from logarithmic and linear distributions, did not introduce a significant difference in the majority of cases, when present in smaller percentages in parameters such as F.PV, Temp.Out, XCO_2.PV and XO_2. PV (Fig. 7c, f, g, h) irrespective of the method used for condensation. However, as the error started to increase towards 10%, F.PV and Temp.Out started to exhibit significant differences from their original condensed value. For the same type of error, parameters DO.Out and FO_2.PV showed significant differences approximately 20% of the time, whereas, pH.Out and DO.PV showed significant differences 50–80% of the time (Fig. 7a, b, d, e). In a majority of cases, the L_p^{TS} norm was able to capture significant differences more often than averaging (Fig. 7a-h, Supplementary Tables S4 – S11). For larger errors such as those arising from a quadratic, cubic and exponential distribution, significant differences were observed from the original value, irrespective of the percentage at which the error was present, except in the case of pH.Out, which exhibited no significant difference for errors arising from cubic and exponential distributions, even at higher percentages.

This seemingly perplexing response of pH is due to a combination of reasons. The parameter pH.Out in the dataset was a collective of both CO_2 and base addition. In order to avoid negative values, the dataset was split into two, one for CO_2 and the other for base addition. However, impact-of-error analysis was performed only on the CO_2 addition data (as base was only added sparsely). This data had zero values in places where base addition had occurred. This specific nature of the dataset, along with the nature of the cubic and exponential error functions might have resulted in the generation of errors smaller than that considered significant.

From this exercise, it is evident that even a small error can have a significant impact on a parameter behaviour depending on the period of exposure and nature of the parameter itself. For the dataset used in this study, a 1.25% error corresponds to 18 min of culture duration. Even such a short duration of exposure seems to reflect in the condensed value

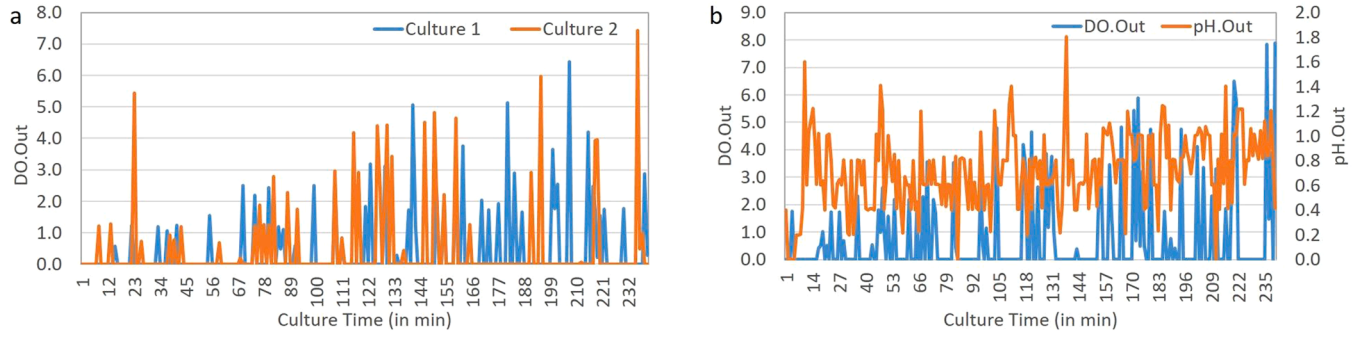


Fig. 5. Comparison between standard averaging and L_p^{TS} norm for data condensation on the outcome and the interpretation of results. (a) An example set of temporal DO.Out profiles for two different cultures from the same day, yielding the same value upon condensation using averaging but different values using L_p^{TS} norm. The x-axis shows culture time in minutes and the y-axis shows DO.Out in percentage. Blue and Orange colours depict data from two different cultures. (b) An example pair of temporal profiles of parameters (DO.Out in blue and pH.Out in orange) showing dissimilarity for two different cultures, yielding the same value upon condensation using averaging but different values using L_p^{TS} norm. The x-axis shows culture time in minutes and the y-axis shows DO.Out in percentage and pH.Out.

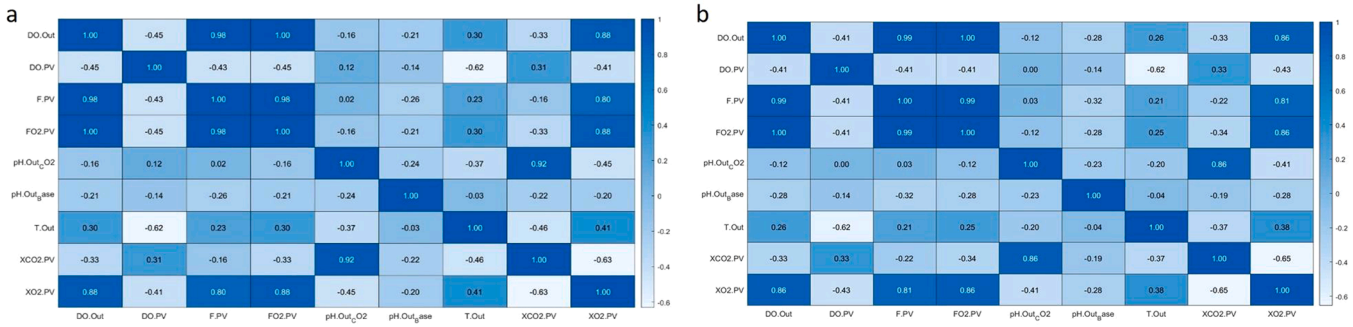


Fig. 6. Correlation matrices showing similar parameter correlations while using condensed values from averaging (a) and L_p^{TS} norm (b). Panel on the right-hand side shows the gradient of the heat map. Due to the specific nature of data generated from the process, parameter pH.Out had to be split into two controller outputs – one for CO_2 addition (pH.Out_CO_2Addition) and one for Base addition (pH.Out_BaseAddition), to avoid manifestation of negative values in the input before calculating correlations. The correlation matrices for the remaining parameters are provided in the supplementary document (Figures S17-S23).

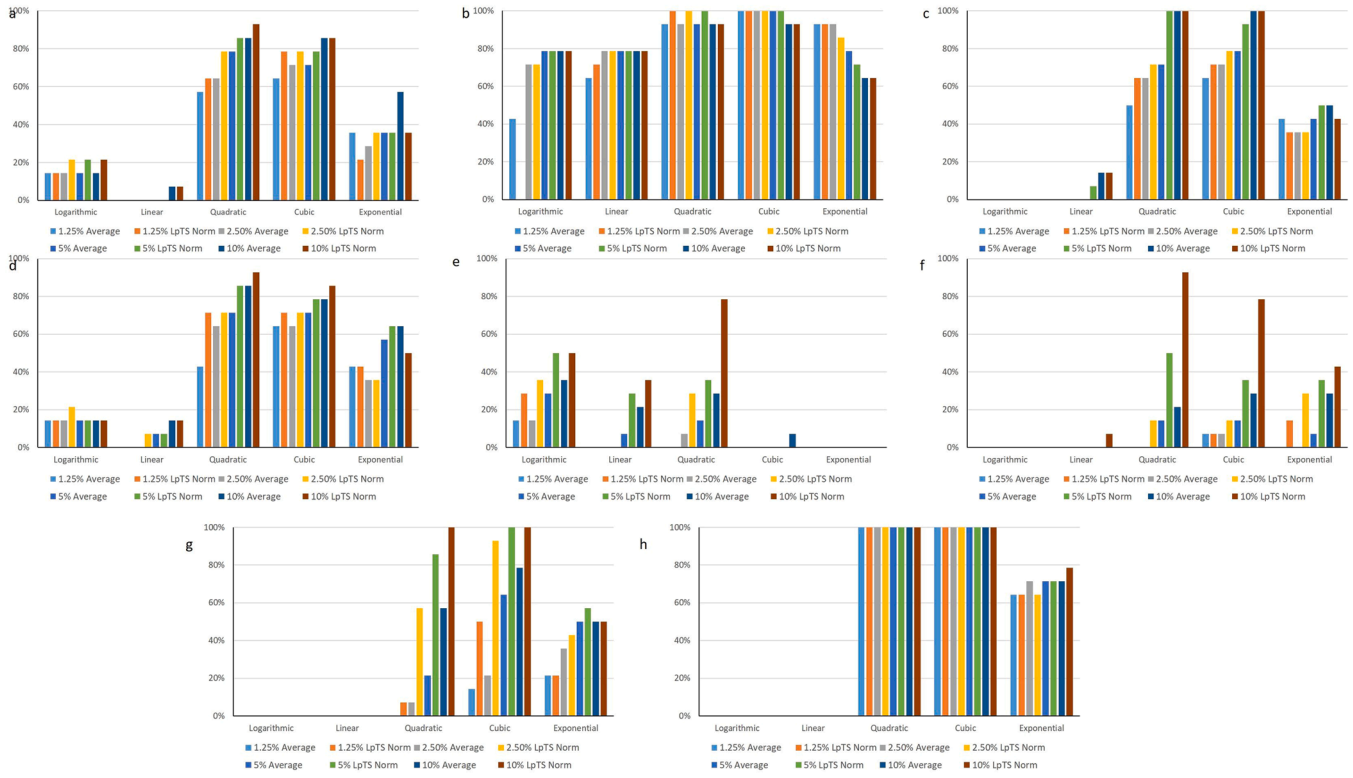


Fig. 7. Results from Impact-of-Error analysis, showing the number of values (in percentage) that were significantly different between condensed values of original dataset and the dataset with different types and percentages of error while employing averaging and L_p^{TS} norm for different parameters. (a) DO.Out (b) DO.PV (c) F.PV (d) FO_2.PV (e) pH.Out (f) Temp.Out (g) XCO_2.PV (h) XO_2.PV.

in several instances, indicating that continuous monitoring is critical for mammalian cell culture processes, to ensure no deviation from expected trajectory. The impact of the reflected change in parameter behaviour, during condensation due to exposure to an error, on the overall culture performance can only be decided during the subsequent stages of analysis. However, this study highlights how important it is to have continuous monitoring for bioprocesses to reduce the risk as even a short duration of exposure to unexpected activity can alter parameter behaviour. In addition, this study has shown that the L_p^{TS} norm can capture significant differences in parameter behaviour, caused by exposure to error very efficiently.

5. Conclusion

Real-time measurements of dozens of critical process parameters are monitored continuously for PAT purposes, and they assist in taking control actions in case the monitored values deviated from their set points. However, very little has been done to harness the full potential of this data in building models for predictive control. This is mainly due to a limitation in the availability of methods that can efficiently perform data condensation, to enable its integration into a data matrix to make it comparable with offline readings. One of the important challenges in this task is to execute condensation in such a way as to capture the temporal profile of these online parameters, using a single value. This is important as they are unique to each culture and can add depth to the information that is extracted from them.

In this study, an alternative method for data condensation of online bioprocess data was explored. The L_p^{TS} norm method was compared with conventionally employed averaging for data condensation and was shown to be better at capturing parameter behaviour. Initially, a

synthetic dataset was used to demonstrate how the L_p^{TS} norm captures the essence of the series better than averaging when the variance is high. Later, the same methods were applied to online data from a mammalian cell culture process. The resulting condensed values did not alter parameter relationships when tested using correlations. An additional impact-of-error analysis shed light onto how efficiently the L_p^{TS} norm can capture differences in parameter behaviour when exposed to error. This demonstrated that the L_p^{TS} norm is a better alternative to averaging for data condensation and can aid in increasing the feature space to generate reliable predictive models to optimise process performance or implementing elaborate control strategies using model predictive control. In addition, the work highlighted the importance of continuous monitoring of variables to ensure product quality, as exposure to even a small error can have deleterious consequences. The proposed data condensation methodology has applicability in areas beyond bioprocess engineering, in scenarios where the objective is to obtain a single value that captures parameter behaviour more efficiently than averaging. Any field that uses data for retrospective analysis and has a combination of online and offline information recorded at different rates can benefit from this method.

CRediT authorship contribution statement

Nishanthi Gangadharan: Formal analysis, Methodology, Software, Validation, Writing – original draft, Visualization. **Ayca Cankorur-Cetinkaya:** Conceptualization, Data curation, Resources, Validation, Supervision. **Matthew Cheeks:** Conceptualization, Resources, Project administration. **Alexander F Routh:** Writing – review & editing, Supervision, Project administration. **Duygu Dikicioglu:** Conceptualization, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgement

The authors thank David Sewell for assisting data collation from AstraZeneca databases. This work was funded by AstraZeneca as part of the University of Cambridge – AstraZeneca Beacon Collaborative project and by the Biotechnology and Biological Sciences Research Council, Grant number: BB/L013770/1. The authors would like to acknowledge support from EPSRC CDT Bioprocess Engineering Leadership (Grant Number EP/L01520X/1). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2023.108402.

References

Bakshi, B.R., Stephanopoulos, G., 1994a. Representation of process trends-III. Multiscale extraction of trends from process data. Comput. Chem. Eng. 18, 267–302. https://doi.org/10.1016/0098-1354(94)85028-3.

Bakshi, B.R., Stephanopoulos, G., 1994b. Representation of process trends-IV. Induction of real-time patterns from operating data for diagnosis and supervisory control. Comput. Chem. Eng. 18, 303–332. https://doi.org/10.1016/0098-1354(94)85029-1.

Charaniya, S., Hu, W.S., Karypis, G., 2008. Mining bioprocess data: opportunities and challenges. Trends Biotechnol. 26, 690–699. https://doi.org/10.1016/j.tibtech.2008.09.003.

Charaniya, S., Le, H., Rangwala, H., Mills, K., Johnson, K., Karypis, G., Hu, W.-S., 2010. Mining manufacturing data for discovery of high productivity process characteristics. J. Biotechnol. 147, 186–197. https://doi.org/10.1016/j.jbiotec.2010.04.005.

Gangadharan, N., Sewell, D., Turner, R., Field, R., Cheeks, M., Oliver, S.G., Slater, N.K.H., Dikicioglu, D., 2021. Data intelligence for process performance prediction in biologics manufacturing. Comput. Chem. Eng. 146, 107226 https://doi.org/10.1016/j.compchemeng.2021.107226.

Heyting, A., Tolboom, J.T.B.M., Essers, J.G.A., 1992. Statistical handling of drop-outs in longitudinal clinical trials. Stat. Med. 11, 2043–2061. https://doi.org/10.1002/sim.4780111603.

Huang, J., Nanami, H., Kanda, A., Shimizu, H., Shioya, S., 2002. Classification of fermentation performance by multivariate analysis based on mean hypothesis testing. J. Biosci. Bioeng. 94, 251–257. https://doi.org/10.1016/s1389-1723(02)80158-x.

Kamimura, R.T., Bicciato, S., Shimizu, H., Alford, J., Stephanopoulos, G., 2000. Mining of biological data I: identifying discriminating features via mean hypothesis testing. Metab. Eng. 2, 218–227. https://doi.org/10.1006/mben.2000.0154.

Kim, T.K., Park, J.H., 2019. More about the basic assumptions of *t*-test: normality and sample size. Korean J. Anesthesiol. 72, 331–335. https://doi.org/10.4097/kja.d.18.00292.

Kohl, M., 2022. MKinfer: inferential Statistics.

Kwak, S.G., Kim, J.H., 2017. Central limit theorem: the cornerstone of modern statistics. kja 70, 144–156. https://doi.org/10.4097/kjae.2017.70.2.144.

Le, H., Kabbur, S., Pollastrini, L., Sun, Z., Mills, K., Johnson, K., Karypis, G., Hu, W.S., 2012. Multivariate analysis of cell culture bioprocess data—Lactate consumption as process indicator. J. Biotechnol. 162, 210–223. https://doi.org/10.1016/j.jbiotec.2012.08.021.

Lee, J.A., Verleysen, M., 2005. Generalization of the L p norm for time series and its application to Self-Organizing Maps. In: WSOM 2005, 5th Workshop on Self-Organizing Maps. Paris France, pp. 733–740.

Lumley, T., Diehr, P., Emerson, S., Chen, L., 2002. The importance of the normality assumption in large public health data sets. Annu. Rev. Public Health 23, 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546.

Max Kuhn Contributions from Jed Wing, A., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Core Team, the R., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T., Max Kuhn, M., 2019. Package "caret" Title Classification and Regression Training Description Misc functions for training and plotting classification and regression models.

Nomikos, P., MacGregor, J.F., 1995. Multi-way partial least squares in monitoring batch processes. Chemom. Intell. Lab. Syst. 30, 97–108. https://doi.org/10.1016/0169-7439(95)00043-7.

Pekarsky, A., Konopek, V., Spadiut, O., 2019. The impact of technical failures during cultivation of an inclusion body process. Bioprocess Biosyst. Eng. 42, 1611–1624. https://doi.org/10.1007/s00449-019-02158-x.

Peng, R.D., 2008. A method for visualizing multivariate time series data. J. Stat. Softw. 25, 1–17. https://doi.org/10.18637/jss.v025.c01.

R Core Team, 2020. R: a language and environment for statistical computing.

Reyes, S.J., Durocher, Y., Pham, P.L., Henry, O., 2022. modern sensor tools and techniques for monitoring, controlling, and improving cell culture processes. Processes 10. https://doi.org/10.3390/pr10020189.

Schlembach, I., Grünberger, A., Rosenbaum, M.A., Regestein, L., 2021. Measurement techniques to resolve and control population dynamics of mixed-culture processes. Trends Biotechnol 39, 1093–1109. https://doi.org/10.1016/j.tibtech.2021.01.006.