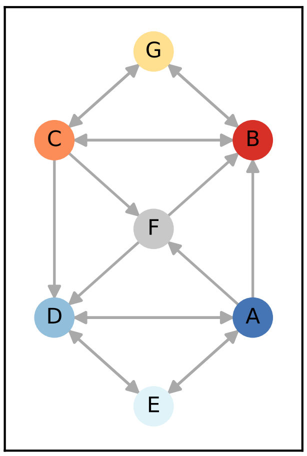
Copyright notice

Graphical Abstract

Qualitative Trend Analysis based on a Mixed-Integer Representation

Dhrubajit Chowdhury, Kris Villez



Highlights

Qualitative Trend Analysis based on a Mixed-Integer Representation

Dhrubajit Chowdhury, Kris Villez

- Qualitative trend analysis can be formulated as a mixed-integer quadratic program
- This is the first qualitative trend analysis method without heuristic rules
- The fit-complexity trade-off is decided by a novel statistical information criterion

Qualitative Trend Analysis based on a Mixed-Integer Representation

Dhrubajit Chowdhury[a], Kris Villez[a]

[a] *Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, 37830, TN, USA*

Abstract

Shape constrained spline fitting is a useful method to impose prior knowledge onto flexible semi-parametric models during parameter estimation. Most typically, the function shape is imposed through order restrictions on the regression coefficients. The intended shape is considered known or selected based on heuristic rules. In this study, we present a method to estimate the optimal set of order restrictions to segment a univariate data series into episodes with distinct shapes. This is also known as the qualitative trend analysis (QTA) problem. The obtained solution uses a trade-off between lack-of-fit and model complexity. Our practical implementation takes inspiration from the generalized order restricted information criterion (GORIC) for inequality-constrained model selection. From this, one learns *(a)* that QTA can be formulated as a mixed-integer quadratic program (MIQP) and *(b)* that the newly proposed mixed order restricted information criterion (MORIC) enables optimal segmentation. This is illustrated through didactic case studies.

Keywords: Model selection, Qualitative representation, Order restricted regression, Shape constraints, Time series segmentation, Truncated distribution

2020 MSC: 62-08

Copyright notice

Keywords: Model selection; Order restricted regression; Shape constraints; Time series segmentation; Truncated distribution

1. Introduction

Qualitative trend analysis (QTA) is the problem of segmenting noisy data series into episodes, where each episode corresponds to a unique shape describing the local trend in the data. The shapes are most often defined by the signs of the first and second derivative. Due to the complexity of this segmentation problem, a large variety of approaches have been tested, with many depending on heuristic rules to select an adequate segmentation [see e.g. 25, 46, 47]. One of these approaches is based on shape constrained spline (SCS) functions [5, 44, 49]. In this case, the optimal locations of the change-points can be identified if the sequence of shapes is known or specified. When this is not known, multiple SCS functions are fitted and the best one is selected by weighing the lack-of-fit against complexity of the segmentation in a heuristic manner [43, 50]. The primary objective for this work is to solve this model selection problem for the first time in an objective, non-heuristic, manner.

The method proposed in this work is inspired by prior work on shape constrained function model identification, which has been used for a variety number of regression problems, such as sensor calibration [18], distribution estimation [7, 27] and identification of dose-response, hazard, and growth curves [15, 9, 12, 17, 27]. These regression problems belong to the broader class of order restricted regression problems [42] and are a type of hybrid model where a flexible semi-parametric function is constrained to satisfy a known shape [23]. A spline function basis is often chosen because a finite number of constraints are necessary and sufficient to define the feasible region of the optimization problem [23, 29, 33]. We do the same in this study.

A particular challenge in QTA is that the number of episodes needed to optimally describe a data series is not known a priori. In contrast, most studies using SCS functions assume that the shape is known and simple [see e.g., 1, 22, 26, 28, 35]. To address this challenge, we adapt the generalized order restricted information criterion (GORIC) as presented in [19, 20] and design a novel model selection criterion, named the *mixed order restricted information criterion* (MORIC), to determine the optimal number of episodes in an objective fashion.

In what follows, we discuss two important methodological developments in pursuit of the objective mentioned above, which are:

1. It is shown for the first time that shape constrained spline (SCS) function fitting can be formulated as a mixed-integer quadratic program (MIQP), for which mature optimization software exists.
2. Data series segmentation based on QTA constitutes a model selection problem. This selection can be executed in an objective manner by means of the newly proposed MORIC.

The resulting method is illustrated with two case studies and compared with two pre-existing methods for QTA. This lead to important conclusions for research concerning both shape constrained function estimation and QTA.

2. Methods

2.1. Notation

A complete list of symbols is given in Table 1.

Table 1: List of mathematical symbols.

Symbol	**Description**	**Type**	**Dimension**
μ	Mean vector	Continuous	1
$\mathbf{\Sigma}$	Variance-covariance matrix	Continuous	1
$\mathbf{\Omega}$	Feasible set of integer variables	-	1
σ ($\hat{\sigma}$)	Measurement noise standard deviation (estimate)	Continuous	1
B	Number of spline coefficients	Integer	1
$\boldsymbol{b}$	Vector of spline coefficients	Continuous	B
b_k	kth spline coefficient	Continuous	1
C ($\tilde{C}$, C_r)	Number of active constraints (for data series, for rth Monte Carlo sample)	Integer	1
$\bar{C}$	Average number of active constraints	Continuous	1
$\boldsymbol{c}$	Vector of change point indicators	Boolean	$K-1$

Continued on next page

Table 1 – continued from previous page

Symbol	Description	Type	Dimensions
c_k	Indicator of change point (vertex change) at the kth spline knot	Boolean	1
$\bar{D}$	Expected degrees of freedom	Continuous	1
$\mathcal{E}$	Set of edges	Set	–
E	Number of edges	Integer	1
e	Edge index	Integer	1
$f(t, \boldsymbol{b})$	Spline function	Function	1
$\mathcal{G}$	Graph representation feasible qualitative sequences	Graph	–
$\boldsymbol{G}$	Transition matrix for graph $\mathcal{G}$	Boolean	$V \times V$
i	Measurement pair index	Integer	1
K	Number of spline knots	Integer	1
L	Maximum length of the QS	Integer	1
M	Arbitrary large positive constant	Continuous	1
N	Number of measurements	Integer	1
P	Spline degree	Integer	1
$\boldsymbol{q}$	Vector of spline knots	Continuous	K
q_k	kth spline knot	Continuous	1
R	Number of Monte Carlo samples	Integer	1

Continued on next page

Table 1 – continued from previous page

Symbol	Description	Type	Dimensions
S	Number of data points between consecutive knots	Integer	1
$s_k^{(1-)}$	kth indicator of non-positivity for first-order difference	Boolean	1
$s_k^{(1+)}$	kth indicator of non-negativity of first-order difference	Boolean	1
$s_k^{(2-)}$	kth indicator of non-positivity of second-order difference	Boolean	1
$s_k^{(2+)}$	kth indicator of non-negativity of second-order difference	Boolean	1
$\boldsymbol{s}_k$	kth vector of non-negativity and non-positivity indicators	Boolean	4
T ($\tilde{T}$, T_r)	Time needed to solve an MIQP (for data series, for rth Monte Carlo sample)	Continuous	1
t	Function input	Continuous	1
t_i	ith function input	Continuous	1
$\mathcal{V}$	Set of vertices	Set	–
$\mathcal{V}_{start}$	Set of feasible source vertices	Set	–
$\mathcal{V}_{end}$	Set of feasible sink vertices	Set	–
V	Number of states / vertices	Integer	1

Continued on next page

Table 1 – continued from previous page

Symbol	Description	Type	Dimensions
v	State / vertex	Integer	1
$\boldsymbol{x}$	Vector of states / vertex indices	Integer	K
x_k	kth state variable / vertex index	Integer	1
$\boldsymbol{y}$	Vector of function outputs	Continuous	N
$\tilde{\boldsymbol{y}}$ ($\tilde{\boldsymbol{y}}_r$)	Vector of measurements (rth Monte Carlo sample)	Continuous	N
y_i	ith function output	Continuous	1
$\tilde{y}_i$	ith measurement	Continuous	1
$\boldsymbol{Z}$	Matrix of non-negativity and non-positivity indicators	Continuous	$4 \times V$
$\boldsymbol{z}_v$	Vector of non-negativity and non-positivity indicators for vertex v	Boolean	4

2.1.1. User-provided information

The method described below will produce a sequence of contiguous episodes to describe a data series. Such a sequence of episodes, each described with *(a)* a start point, *(b)* an end point and *(c)* a shape, is named a *qualitative representation* (QR). The sequence of shapes is named the *qualitative sequence* (QS). For the proposed method to work, we assume two pieces of information are provided. The first is the univariate data series available for analysis. The second is the set of qualitative representations that are considered feasible.

Data series. The univariate data series consists of N data points ($i = 1, \ldots, N$). Each data point is a pair of values, given as $(t_i, \tilde{y}_i)$, where t_i is assumed to be measured without error. The values for t_i can be positioned arbitrarily. The second element, $\tilde{y}_i$, is a noisy measurement of the unknown variable y_i. The N-dimensional vectors $\boldsymbol{t}$, $\tilde{\boldsymbol{y}}$, and $\boldsymbol{y}$ contain t_i , $\tilde{y}_i$, and y_i in the ith position.

Feasible set of qualitative representations (QRs). The permissible sequences of shapes to describe the data series is defined with a directed graph $\mathcal{G}$, as in [47], which is interpreted as a possibilistic Markov chain [40]. Figure 1 shows an example of such a graph. The graph contains V vertices and E edges. The presence (absence) of an edge in the graph from vertex v_s to v_t signifies that the shape associated with vertex v_s can(not) be followed by the shape associated with vertex v_t. All vertices have a self-loop edge in all graphs used for this work. A unique shape, also known as a primitive, is associated with each vertex v ($v \in \mathcal{V} := \{1, \ldots, V\}$). As is typical for QTA applications, the primitives are defined as unique combinations for the signs of the first and second derivatives. Table 2 lists the 13 primitives considered in this work. For visualization purposes, each primitive is described with a unique letter. In general, multiple vertices can take on the same primitive, but not vice versa. Note that the graph in Figure 1 only permits shape sequences that imply continuity of the first and second derivative.

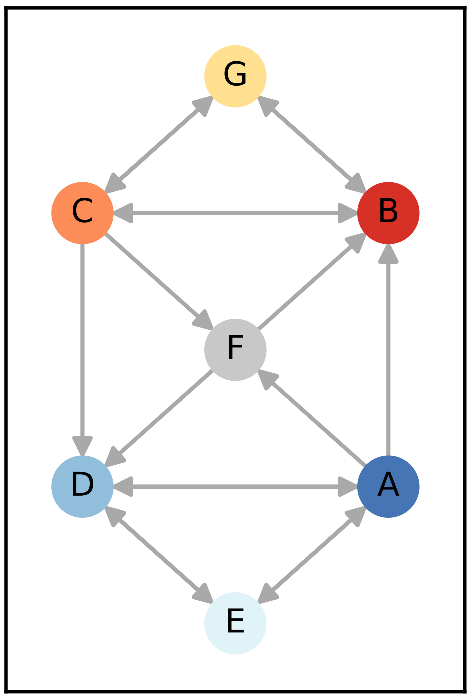


Figure 1: Exemplary graph representing feasible qualitative representations (QRs). All vertices have self-loops (not shown). This particular graph only allows vertex changes permitting smooth changes of the first and second derivative.

To enable a precise treatment of the QTA problem, the primitive for each vertex is defined by a 4-dimensional vector, $\boldsymbol{z}_v$. This vector is composed of four boolean constants, namely z_v^{1+}, z_v^{1-}, z_v^{2+}, and z_v^{2-} (in this order). Each boolean value defines non-negativity or non-positivity of one of the considered derivatives. For example, $z_v^{1+} = 1$ specifies that a non-negativity constraint is applied to the first derivative. If $z_v^{1+} = 0$, this non-negativity constraint is not imposed. Analogously, $z_v^{1-} = 1$ ($z_v^{1-} = 0$) specifies presence (absence) of a non-positivity constraint for the first derivative. Together, $z_v^{1+} = 1$ and $z_v^{1-} = 0$ describe an increasing trend. Similarly, $z_v^{1+} = 0$ and $z_v^{1-} = 1$ describe a decreasing trend; $z_v^{1+} = 1$ and $z_v^{1-} = 1$ describe a constant trend; and $z_v^{1+} = 0$ and $z_v^{1-} = 0$ describe absence of any constraint on the first derivative. The values of z_v^{2+} and z_v^{2+} describe the sign of the second derivative in an analogous manner. The boolean constants for each primitive are listed in Table 2. The set of V vectors $\boldsymbol{z}_v$ are combined as column vectors in a $4 \times V$-dimensional matrix $\boldsymbol{Z}$ such that $\boldsymbol{Z}_{\cdot,v} = \boldsymbol{z}_v$,

Table 2: List of primitives and associated signs of the first and second derivative and boolean indicator values. A question mark (?) indicates that the sign of the derivative in unspecified for the corresponding primitive.

Character	**Derivative signs**		**Boolean specification ($\boldsymbol{z}$)**			
	$\frac{df}{dt}$	$\frac{d^2f}{dt^2}$	$z^{(1+)}$	$z^{(1-)}$	$z^{(2+)}$	$z^{(2-)}$
A	$-$	$+$	0	1	1	0
B	$+$	$+$	1	0	1	0
C	$+$	$-$	1	0	0	1
D	$-$	$-$	0	1	0	1
E	$-$	0	0	1	1	1
F	0	0	1	1	1	1
G	$+$	0	1	0	1	1
U	$+$	?	1	0	0	0
L	$-$	?	0	1	0	0
N	?	$-$	0	0	0	1
O	?	0	0	0	1	1
P	?	$+$	0	0	1	0
Q	?	?	0	0	0	0

The directed graph $\mathcal{G}$ is equivalently described with a $V \times V$-dimensional matrix $\boldsymbol{G}$. This matrix contains boolean values to represent the edges in the graph. Columns correspond to the start vertex of the edge while rows correspond to the end vertex. Thus, $T_{w,v} = 1$ ($T_{w,v} = 0$) indicates that a change from vertex v to w is feasible (infeasible). For example, the graph $\mathcal{G}$ in Figure 1 has this transition matrix:

$$\boldsymbol{G} = \begin{array}{c} \begin{array}{ccccccc} \mathsf{A} & \mathsf{B} & \mathsf{C} & \mathsf{D} & \mathsf{E} & \mathsf{F} & \mathsf{G} \end{array} \\ \left[\begin{array}{ccccccc} 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{array}\right] \end{array} \begin{array}{c} \\ \mathsf{A} \\ \mathsf{B} \\ \mathsf{C} \\ \mathsf{D} \\ \mathsf{E} \\ \mathsf{F} \\ \mathsf{G} \end{array} . \tag{1}$$

Finally, the end user can also specify that the first primitive in a QR must match the shape of a subset of the vertices in the graph. To this end, $\mathcal{V}_{start}$ is the set of vertices describing the feasible shapes for the first episode in the QR. Similarly, $\mathcal{V}_{end}$ is the set of vertices describing the feasible shapes for the last episode in the QR. From this point onward, $\boldsymbol{G}$, $\mathcal{V}_{start}$, $\mathcal{V}_{end}$, and $\boldsymbol{Z}$ are user-specified constants.

2.2. Method A: Shape constrained spline (SCS) model fitting and selection

2.2.1. Main ideas

The proposed method is based on two fundamental elements:

- The quality of a QR to describe a data series can be quantified by fitting an SCS function to the data, such that the fitted function satisfies the QR. The QR itself can be optimized by choosing the path in the graph $\mathcal{G}$ that leads to the best fit. This path search, it turns out, can be solved as a mixed-integer quadratic program (MIQP).
- The maximal length of the path determines the complexity of the QR. With increasing length, a better fit will be obtained. Optimal complexity of the QR is determined by means of an information criterion, similar to the Akaike information criterion (AIC). This information criterion weighs the improvement in fit against the increased degrees of freedom as the complexity of the QR increases. The degrees of freedom is defined here as the number of spline coefficients minus the expected number of active constraints among the constraints involving the spline coefficients.

2.2.2. Qualitative trend analysis as a mixed integer quadratic program (MIQP)

Spline function model. We assume that the measurement errors, $\tilde{e}_i$, are sampled independently from the same normal distribution, i.e. the following measurement model is assumed:

$$\tilde{y}_i = y_i + \tilde{e}_i \tag{2}$$

$$\tilde{e}_i \sim \mathcal{N}(0, \sigma) \ . \tag{3}$$

where σ is the (unknown) measurement error standard deviation.

The data series is described with a (univariate) spline function, denoted as $f(t, \boldsymbol{b})$ where t is the function input and $\boldsymbol{b}$ is the B-dimensional vector of spline coefficients. For theory and detailed instructions on the use of spline functions, we refer to [4, 38, 39, 37]. Importantly for this work is that a spline function can be constrained to satisfy a particular shape by adding a finite number of equality and inequality constraints to the parameter estimation problem [29, 32, 33].

Spline models exhibit a number of hyper-parameters, namely the number of knots, K, the knot locations, q_k ($k = 1, \ldots, K$), and the degree of the spline function, P. The number of coefficients equals $B = K + P - 1$ for all univariate spline functions. The B-dimensional vector of coefficients is given as $\boldsymbol{b}$. In this work, the degree of the spline is $P = 1$, making $f(t, \boldsymbol{b})$ piece-wise linear in t. In all cases studied in this work, knots are placed at values contained in $\boldsymbol{t}$. More specifically, these knots are placed as follows:

$$q_k = \begin{cases} t_1 & \text{if: } k = 1 \\ t_N & \text{if: } k = K \\ t_{k+1} & \text{if: } 1 < k < K \wedge S = 1 \\ t_{(k-1)\,S+1} & \text{if: } 1 < k < K \wedge S > 1 \end{cases}, \qquad (k = 1, \ldots K) \tag{4}$$

where S specifies the smallest number of data points between consecutive knots and is always larger than or equal to 1. It follows that the total number of knots is equal to $K = N - 2$ if $S = 1$ ($q_k = t_1, t_3, \ldots, t_{N-2}, t_N$) and $K = \left\lfloor \frac{N-1}{S} \right\rfloor + 1$ otherwise ($q_k = t_1, t_{S+1}, \ldots, , t_{\left(\left\lfloor \frac{N-1}{S} \right\rfloor - 1\right) S+1}, t_N$). Most importantly, this design ensures that the total number of continuous parameters ($B = K$ coefficients and 1 standard deviation σ) is always smaller than the number of measurements ($B + 1 \leq N$), thus leading to structural identifiability of these parameters (for definitions and discussion [see e.g. 2, 34]. By default, we set $S = 1$, thus producing $K = N - 2$ knots.

Shape constrained spline (SCS) function fitting. Having selected a linear spline function, one can optimize the spline coefficients $\boldsymbol{b}$ to obtain a least-squares optimal fit while making sure the obtained spline function satisfies the imposed shape constraints. In appendix A, we show that this optimization problem takes the following MIQP form:

$$\hat{\boldsymbol{b}}, \hat{\boldsymbol{x}} := \arg\min_{\boldsymbol{b},\boldsymbol{x}} \sum_{i}^{N} \left(\tilde{y}_i - y_i\right)^2 \tag{5}$$

$$y_i = b_k + \frac{t_i - q_k}{q_{k+1} - q_k}\,(b_{k+1} - b_k) \quad (q_k \le t_i < q_{k+1}) \tag{6}$$

$$M \ge M\, s_k^{(1+)} - \frac{b_{k+1} - b_k}{q_{k+1} - q_k} \quad (k = 1, \ldots, K-1) \tag{7}$$

$$M \ge M\, s_k^{(1-)} + \frac{b_{k+1} - b_k}{q_{k+1} - q_k} \quad (k = 1, \ldots, K-1) \tag{8}$$

$$M \ge M\, s_k^{(2+)} - \left(\frac{b_{i+1} - b_k}{q_{k+1} - q_k} - \frac{b_k - b_{k-1}}{q_k - q_{k-1}}\right) \quad (k = 2, \ldots, K-1) \tag{9}$$

$$M \ge M\, s_k^{(2-)} + \left(\frac{b_{k+1} - b_k}{q_{k+1} - q_k} - \frac{b_k - b_{k-1}}{q_k - q_{k-1}}\right) \quad (k = 2, \ldots, K-1) \tag{10}$$

$$\boldsymbol{s}_k = \boldsymbol{Z}_{\cdot,x_k} := \begin{bmatrix} s_k^{(1+)} & s_k^{(1-)} & s_k^{(2+)} & s_k^{(2-)} \end{bmatrix}^T \quad (k = 1, \ldots, K) \tag{11}$$

$$T_{x_{k+1},x_k} = 1 \quad (k = 1, \ldots, K-1) \tag{12}$$

$$x_1 \in \mathcal{V}_{start} \tag{13}$$

$$x_K \in \mathcal{V}_{end} \tag{14}$$

$$c_k = \begin{cases} 0 & \text{if: } x_k = x_{k+1} \\ 1 & \text{otherwise} \quad (k = 1, \ldots, K-1) \end{cases} \tag{15}$$

$$L \ge \sum_{k=1}^{K-1} c_k + 1\,. \tag{16}$$

In the above problem, the set of variables are:

- $\boldsymbol{b}$: The B-dimensional vector of spline coefficients
- $\boldsymbol{c}$: A $K-1$-dimensional vector of boolean variables indicating the pres-

ence of a change point. c_k is 1 when there is a vertex change from knot k to knot $k+1$ and 0 otherwise.

- $\boldsymbol{s}_k$: A four-dimensional vector of booleans associated with knot k specifying the inequality constraints for the point-wise derivatives computed at this knot. The four elements are $s_k^{(1+)}$, $s_k^{(1-)}$, $s_k^{(2+)}$, $s_k^{(2-)}$. When $s_k^{(1+)} = 1$, this means the first derivative is constrained to be positive at knot k. Together, $s_k^{(1+)}$, $s_k^{(1-)}$, $s_k^{(2+)}$, and $s_k^{(2-)}$ specify positivity and negativity constraints for the first and second derivatives. The boolean vector $\boldsymbol{s}_k$ specifying the shape at knot k is obtained by selecting the x_kth column from the user-specified matrix $\boldsymbol{Z}$.

- $\boldsymbol{x}$: A K-dimensional vector of integers defining the path traveled through $\mathcal{G}$ by specifying a vertex for each knot k.

- $\boldsymbol{y}$: The N-dimensional vector of spline function evaluations at the measurement points ($y_i = f(t_i, \boldsymbol{b})$).

All remaining elements are constant while solving and consist of:

- $(t_i, \tilde{y}_i)$: The N-dimensional set of measurement pairs

- $\boldsymbol{q}$: The K-dimensional vector of spline knots

- L: The maximal number of episodes; equal to the maximum number of vertex changes plus one.

- M: An arbitrarily large positive constant used to transform bilinear constraints into an equivalent set of linear inequality constraints (see appendix A for details)
- $\boldsymbol{G}$: The transition matrix for $\mathcal{G}$
- $\boldsymbol{Z}$: The matrix defining the primitive associated with each vertex in $\mathcal{G}$
- $\mathcal{V}_{start}$: The feasible start vertices
- $\mathcal{V}_{end}$: The feasible end vertices

The capacity L is a hyper-parameter defining the maximum number of episodes. In this work, we tune it with an AIC-like criterion, as will be explained below.

Once the problem is solved, the vertex values $\hat{\boldsymbol{x}}$ represent the obtained QR and the corresponding sequence of primitives defines the QS. The number of primitives equals $\sum_{k=1}^{K-1} c_k + 1$. The maximum likelihood estimate of σ, $\hat{\sigma}$, is obtained by computing the root mean squared residual (RMSR):

$$\hat{\sigma} = \sqrt{\frac{1}{N} \sum_{i}^{N} (y_i - \hat{y}_i)^2} \tag{17}$$

where $\hat{y}_i$ are the left-hand sides of (6) evaluated at the optimum solution.

2.2.3. Tuning the shape constrained spline (SCS) model by optimal selection of the number of episodes (L)

In practical circumstances, it may not be known a priori how many episodes (L) are adequate for segmentation. To optimally select the number of primitive changes one must trade off the fit of the model, represented by the optimal value of the objective function used above, against its ability to describe an arbitrary data series. The Akaike information criterion (AIC) is one such tool which trades off model fit (likelihood) against complexity (number of parameters). Unfortunately, the AIC cannot be applied for selection between inequality-constrained models. Instead, the generalized order restricted information criterion (GORIC) allows selection among inequality-constrained models [19, 20]. The most important element of the GORIC is that it contains a penalty term which is equal to the number of parameters $(K + 1)$ minus the expected number of active constraints, $\bar{C}$. An analytical expression for $\bar{C}$ is available in limited cases [48]. In the general case, $\bar{C}$ can only estimated through Monte Carlo integration. To do so, one samples a large number (R) of random data series ($\tilde{\boldsymbol{y}}_r$, $r = 1, \ldots, R$) from a known distribution specifying a null hypothesis [19, 20]. For each of the sampled vectors, one obtains the parameter estimates $\boldsymbol{b}_r$, and determines the number of active constraints involving elements of $\boldsymbol{b}_r$, called C_r. The expected number of active constraints ($\bar{C}$) is then estimated as $\bar{C} = \sum_r^R C_r/R$. Once a value for $\bar{C}$ is available, the information criterion (IC) can be computed as:

$$IC = N\ \left(1 + \log(2\pi) + 2\ \log(\hat{\sigma})\right) + 2\,\bar{D} \tag{18}$$

$$= N\ \left(1 + \log(2\pi) + 2\ \log(\hat{\sigma})\right) + 2\ \left(N - 1 - \bar{C}\right)\ . \tag{19}$$

where the first term is the log-likelihood and the second term is a penalty term that is linear in the estimated degrees of freedom, $\bar{D} = N - 1 - \bar{C}$. As discussed in [19, 20], the degrees of freedom is an estimate for the effective dimension of the feasible set of K-dimensional vectors of spline coefficients ($\boldsymbol{b}$) and the (unconstrained) noise standard deviation (σ). Choosing an adequate model is then reduced to selecting L with the smallest value for this model selection criterion.

Unfortunately, application of the GORIC implies that the parameter estimation problem takes the form of a convex QP. To obtain an equivalent criterion for our case (MIQP), we simply replace the QP (as studied in [20]) with the MIQP described above. To this end, $\bar{C}$ is also computed through Monte Carlo integration via the following steps:

1. Sample R random instances of the data series ($\tilde{\boldsymbol{y}}_r$, $r = 1, \ldots, R$) from the standard multivariate normal distribution:

$$\tilde{\boldsymbol{y}}_r \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\ \ (r = 1, \ldots, R); \tag{20}$$

with $\boldsymbol{\mu} = \mathbf{0}_{N\times N}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_{N\times N}$.

2. With each random instance ($r = 1, \ldots, R$), *(a)* set $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{y}}_r$, *(b)* solve the above MIQP problem (5-16) and *(c)* report C_r, the number of active constraints among the constraints involving elements of $\boldsymbol{b}$, i.e. among equations (7-10);
3. Estimate the expected number of active constraints as $\bar{C} = \sum_r^R C_r/R$.

In a geometric sense, the main difference is that the GORIC requires the feasible region for $\boldsymbol{b}$ to be a convex polyhedron, whereas the proposed approach abandons this requirement. Consequently, $\bar{C}$ estimates the expected number of constraints on $\boldsymbol{b}$ that are active for a randomly generated instance of the MIQP. The proposed model selection criterion for spline function fitting with a mixed order restriction is named the *mixed order restricted information criterion* (MORIC), given the use of mixed-integer programming.

2.2.4. Implementation of the shape constrained spline (SCS) model

All results were obtained with Python (v3.8.5) [36], including the *numpy* (v1.19.0), *networkx* (v2.4) and *matplotlib* (v3.2.2) packages [10, 11, 14]. All instances of the MIQP described above were solved with Gurobi [8], which is called from Python with the *gurobipy* (v5.0.2) package.

2.3. Method B and C: Qualitative path estimation (QPE) and piece-wise polynomial fitting (PWP)

We compare the proposed method with two existing methods for QTA. These are the *qualitative path estimation* (QPE) method proposed in [47] and the interval-halving method for piece-wise polynomial (PWP) fitting proposed in [3]. Details on their implementation, application, and tuning are provided in the *Supplementary Materials*.

2.4. Case studies

Two univariate data series are studied in detail to evaluate and compare the methods for qualitative trend analysis. The first data series constitutes an illustrative examples while the second data series quantifies the improvement of the newly proposed MORIC-based method relative to previously established methods.

2.4.1. Case I. Titanium data series

The Titanium data set was reported originally in [4] and has been used to demonstrate SCS function fitting methods in [45] and [50]. In this study, it is selected to illustrate the MORIC-based model selection method (Section 2.2.3). Due to its small size, this particular data set allowed us to set L at every value between 1 and K while allowing sufficient time to compute the corresponding degrees of freedom ($\bar{D}$). The number of Monte Carlo samples is $R = 180$ in this case.

2.4.2. Case II. Nitrite control data series

The nitrite control data set was published originally in [43] and was used there to illustrate a trend-based control mechanism. To this end, the online control system repeatedly evaluates the lack-of-fit of SCS functions, each with a different capacity (L) but with the same user-specified graph $\mathcal{G}$. The model selection criterion was based on a heuristic however. In this study, we use these data to compare the MORIC-based model selection method (Section 2.2.3) against the original heuristic method. In addition, this data set helps to evaluate the potential of the MORIC for online applications in process monitoring, diagnostics, and control. The number of Monte Carlo samples is $R = 1200$ in this case.

3. Results

3.1. Case I. Titanium data series

Provided information. QTA is used to analyze the data series ($\boldsymbol{t}$, $\boldsymbol{y}$) with length $N = 49$ as shown in Figure 2(a) into a series of episodes, each associated with a unique shape, defined as primitive in the above. Next to the data, the newly proposed method (method A, SCS) requires a user-defined graph $\mathcal{G}$. This graph is shown in Figure 1. The start and end vertices are unconstrained ($\mathcal{V}_{start} = \mathcal{V}_{end} = \mathcal{V}$). The QPE method makes use of the same graph while the PWP method does not require this kind of user-specified information.

Results with method A (SCS). Problem (5-16) is first solved $K = 47$ times by setting L equal to $1, 2, \ldots, K$. The resulting RMSR ($\hat{\sigma}$) is shown as a function of L in Figure 3(a). Note that increasing L means that RMSR decreases and thus forms a set of Pareto-optimal solutions trading model fit (RMSR) against model complexity (L). Figure 3(b) shows the number of active constraints ($\tilde{C}$). Note that there is no strictly monotonic relationship in this case. This shows that simply counting the number of active constraints of the competing models is not sufficient to gauge model complexity. Figure 3(d) shows the amount of time needed to solve each of the candidate SCS models ($\tilde{T}$). This requires less than a minute for each.

Problem (5-16) is then solved $K = 47$ times by setting L equal to $1, 2, \ldots, K$ for each of $R = 180$ randomly generated vectors $\tilde{\boldsymbol{y}}_r$. Figure 3(b) shows the number of active constraints for each random vector C_r as well as the average $\bar{C}$, as a function of L. The average number of active constraints decreases with increasing L. This is expected given that increasing the number of episodes (L) increases the feasible region for the regression coefficients. The computational requirements to solve the random instances of (5-16) are quite large, taking an hour on average for intermediate values of L ($10 \leq L \leq 16$). The total required time to compute and select an SCS model amounts to 135 days in this case.

In the proposed method, the trade-off between model fit (RMSR, Figure 3(a)) and model complexity ($\bar{C}$, Figure 3(b)) is determined by selecting the value for L producing a minimum value for the MORIC (Eq. 2.2.3). Figure 3(c) shows the MORIC as function of L. The minimum value of the MORIC is found at $\hat{L} = 4$ and represents the optimal length of the segmentation. The corresponding function, the first-order differences, and the resulting segmentation are shown in Figure 2(a-c). While there is no ground truth result to compare to in this case, the resulting segmentation is sensible to the human eye.

Results with method B (QPE) and C (PWP). The results obtained with the pre-existing methods are discussed in detail in the *Supplementary Materials.* We provide a summary here. Both methods produce a single QR only so there is no need to select one. Both methods are computationally efficient, requiring less than a second to compute. While fast, both methods produce a more complex segmentation. The QPE method (B) generates a QR of

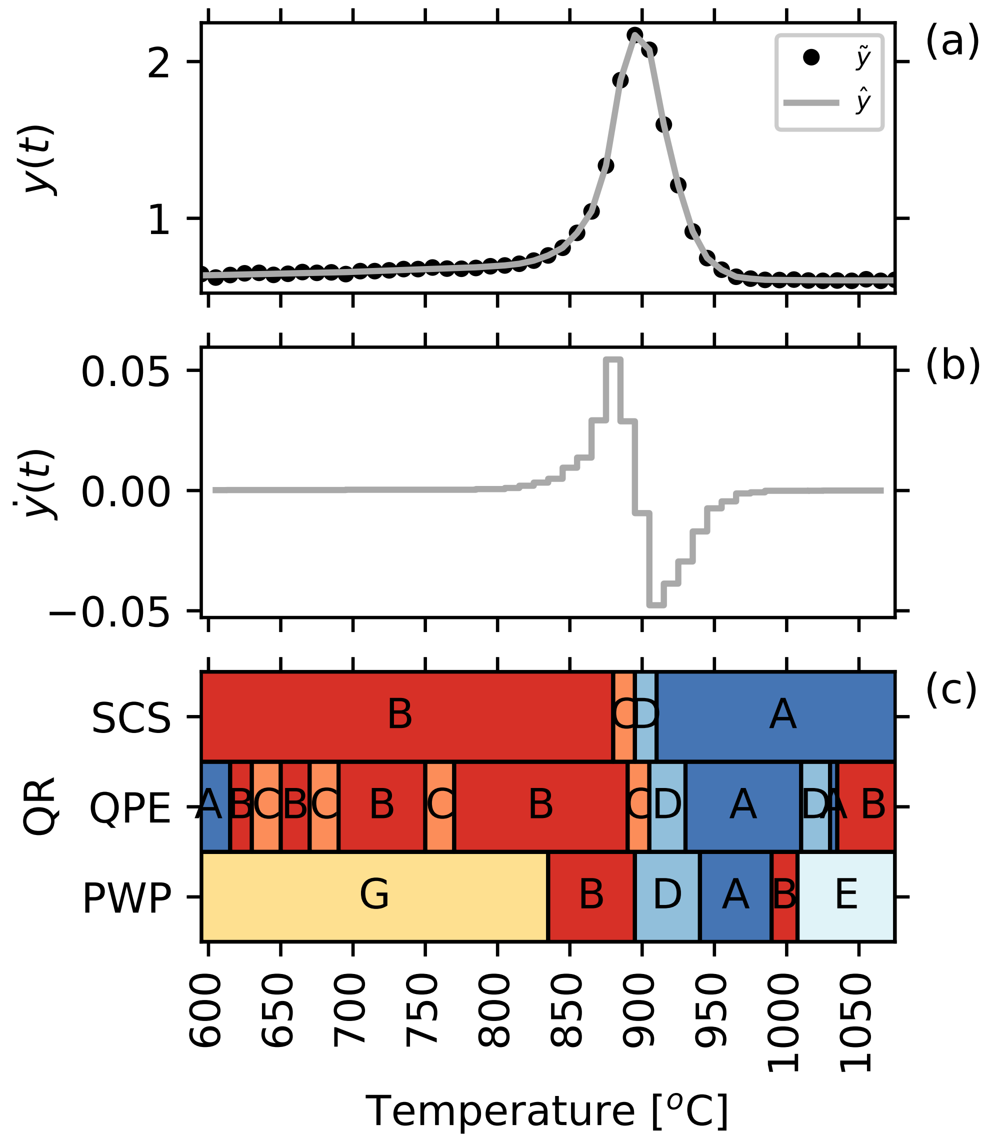


Figure 2: Results obtained with the Titanium data series: *(a)* data series ($\tilde{y}$) and model obtained with method A (SCS, $\hat{y}$); *(b)* estimated first-order difference obtained with method A (SCS); *(c)* qualitative representation (QR) with method A (SCS), B (QPE), and C (PWP).

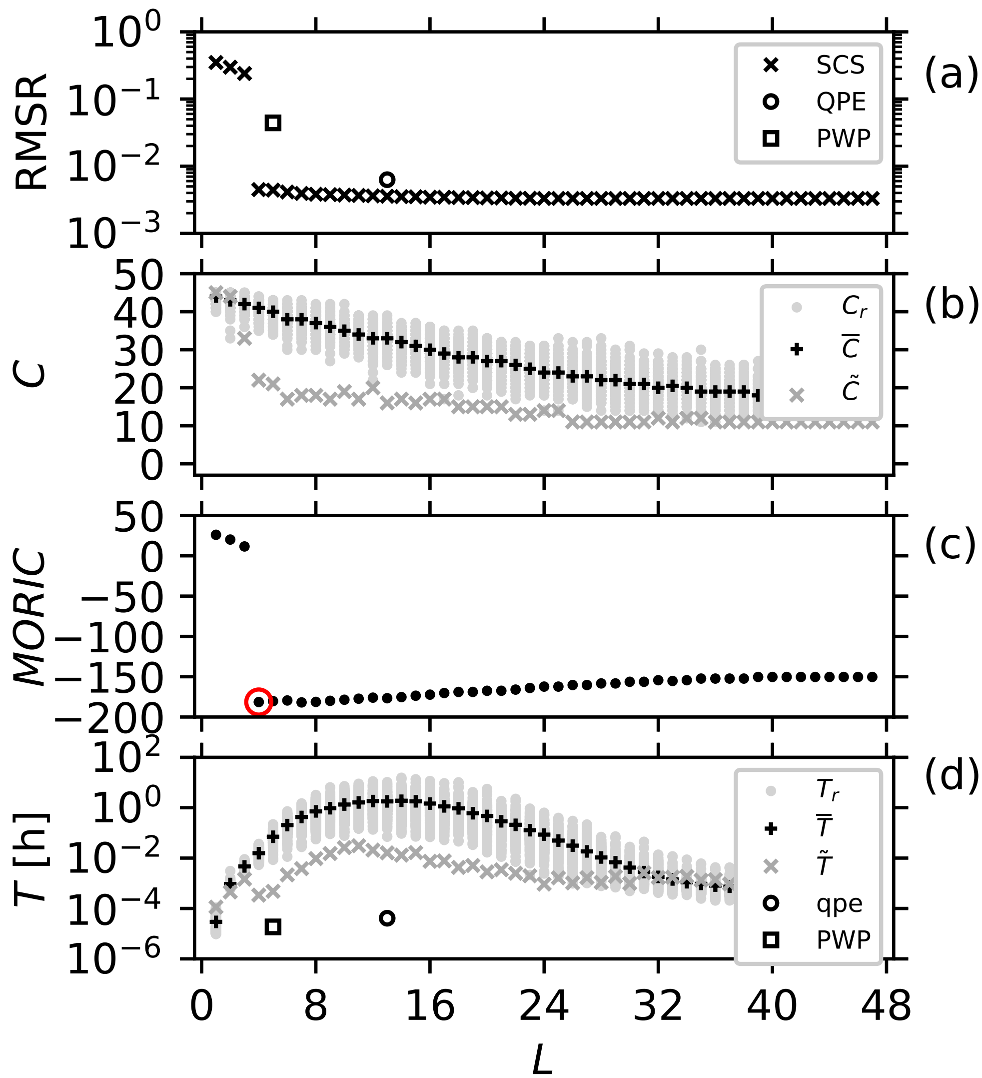


Figure 3: Results obtained with the Titanium data series: *(a)* Lack-of-fit ($RMSR$) as a function of L obtained with the method A (SCS), B (QPE), and C (PWP); *(b)* Number of active constraints as a function of L for the data series ($\tilde{C}$), random instances (C_r), and average across random instances ($\bar{C}$); *(c)* Mixed-order restricted information criterion ($MORIC$) as a function of the maximum number if segments (L); and *(d)* Time needed to compute the lack-of-fit ($RMSR$) and number of active constraints as a function of L with method A (SCS) for the data series ($\tilde{T}$), random instances (T_r), and average across random instances ($\bar{T}$), with method B (QPE), and with method C (PWP).

length $L = 13$ and the PWP method (C) produces $L = 6$ episodes. Figure 3(a) shows the RMSR values as a function of corresponding length of the segmentation. One can see that these values are larger than the values obtained with the SCS method for the same number of episodes. Moreover, the data points are above the Pareto curve obtained with the SCS method. This means that the QRs produced by the QPE and PWP methods deliver a poorer combination of fit and complexity, regardless of the model selection criterion applied to select the SCS model. The QRs obtained with each method are visualized in Figure 2(c).

3.2. Case II. Nitrite concentration data series

Provided information. QTA is used to analyze the data series ($\boldsymbol{t}$, $\boldsymbol{y}$) with length $N = 1025$ as shown in Figure 4(a) into a series of episodes, each associated with a unique shape, defined as primitive in the above. Next to the data, the newly proposed method (method A, SCS) requires the user to define a graph. This graph is shown in Figure 4(b). The first vertex is constrained to be the vertex with primitive U ($\mathcal{V}_{start} = \{1\}$) while any vertex can be the last vertex ($\mathcal{V}_{end} = \{1, 2, 3\}$). The transition matrix for this graph is:

$$\boldsymbol{G} = \begin{array}{cc} \begin{array}{ccc} \mathsf{U} & \mathsf{D} & \mathsf{A} \end{array} & \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} & \begin{array}{c} \mathsf{U} \\ \mathsf{D} \\ \mathsf{A} \end{array} \end{array} \tag{21}$$

where the primitives (U, D, and A) indicate the corresponding vertices.

The QPE method makes use of the same graph and constraints while the PWP method does not involve such a graph or constraints. To emulate the trend-based control application of [43], the QTA problem is solved multiple times. It is executed first with the first 40 points in the data series and repeated by adding 8 more data points in the data series until the end of the data series is met. This leads to 123 instances of the QTA problem.

Results with method A (SCS). In [43], destabilizing nitrite accumulation events in a nitrification reactor are avoided by tracking the trend in a nitrite concentration signal. Following a pH setpoint drop, the trend is expected to satisfy the QS *UDA* before the pH setpoint is increased again. Initially, the data series is expected to be described well as a simple *U* shape. Somewhat later, the *UD* shape is expected to become more adequate. Finally, when the *UDA* shape is the most sensible description, it is assumed that the nitrite concentration has dropped below a safety-critical level and that the pH setpoint can be increased without jeopardizing process stability. For this reason, [43] formulate the decision problem as a selection problem between three SCS functions, with shapes *U*, *UD*, and *UDA*.

The problem (5-16) is solved for every QS ($L = 1, \ldots, 3$). Figure 4(c) shows the evolution of the RMSR ($\hat{\sigma}$) for each QS as a function of the accumulated number of data points (N). One can see that the RMSR for the *U* and *UD* models increases as more data points become available while the RMSR of the *UDA* model hovers around the same value throughout. Importantly, the feasible set of parameters $\boldsymbol{b}$ in the *U* model is a subset of the the feasible set of parameters $\boldsymbol{b}$ in the *UD* model. The same holds for the feasible set for the *UD* and *UDA* models. Thus, the models represent a series of nested models. Consequentially, for a given number of collected data records (N), the RMSR value for a simpler model (e.g. *U*) is necessarily higher than or equal to the RMSR value for a more complex model (e.g. *UD*). In turn, this means the minimal RMSR value is not an adequate guide

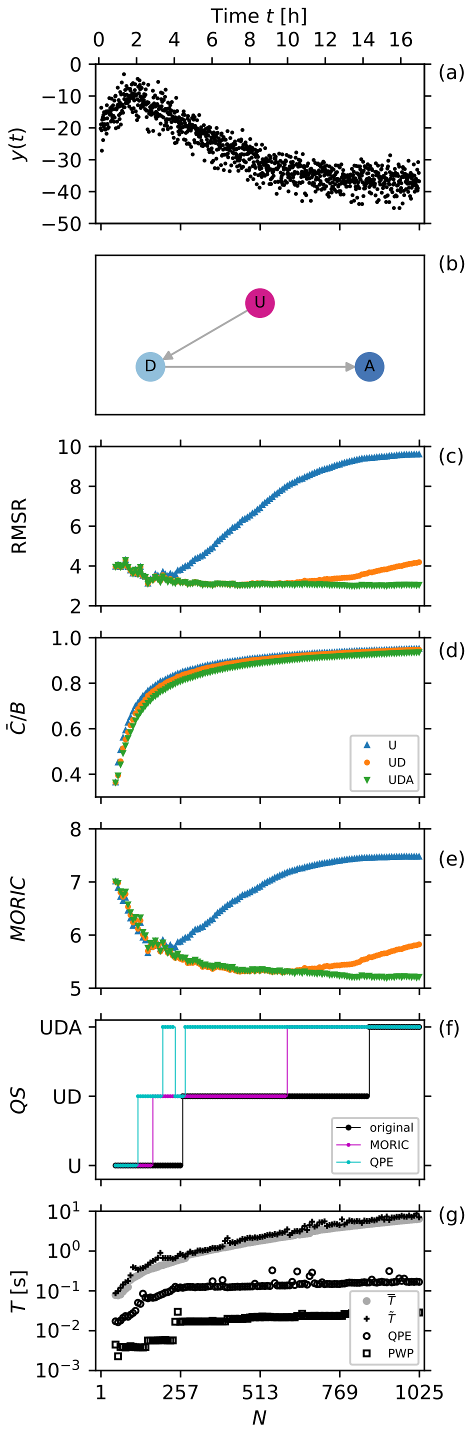


Figure 4: Nitrite concentration data series analyzed with method A (SCS) and B (QPE): *(a)* Data series ($\tilde{y}$); *(b)* Graph $\mathcal{G}$, shown here without self-loops (SCS and QPE); *(c)* Root mean squared residual ($RMSR(N)$) with SCS as a function of the number of accumulated data points (N); *(d)* Average number of active constraints ($\bar{C}$) with SCS divided by the number of spline coefficients (B) as a function of the number of accumulated data points (N); *(e)* Mixed-order restricted information criterion ($MORIC(N)$) with SCS as a function of the number of accumulated data records (N); *(f)* Selected qualitative sequence (QS) with SCS and QPE as a function of the number of accumulated data points (N) as in the original study; *(g)* Computational requirements for analysis of the data series with SCS ($\tilde{T}$), average for a single random instance with SCS ($\bar{T}$), and total time for QPE and PWP.

for model selection. The improvement of fit must be traded off against the model complexity, i.e. the size of the feasible parameter set.

To decide which model is most adequate at any given time, [43] use a set of two heuristic rules: *(i)* QS *UD* is preferred over QS *U* when $RMSR_{UD} < 1.2\,RMSR_{U}$ and *(ii)* QS *UDA* is preferred over QS *UD* when $RMSR_{UDA} < 1.2\,RMSR_{UD}$. The selected QS is shown as a function of the collected data records (N) in Figure 4(f). In this case, the *UDA*-based model is accepted for the first time after $N = 817$ data records have been collected. Applying a wait time of 2h, as in [43], means that the next control action follows after $N = 937$ data records are collected, i.e. at 15 hours and 37 minutes into the control cycle. The total time needed to solve the 3 problems (5-16) ($L = 1, \ldots, 3$), $\tilde{T}$, is shown as a function of the number of data points (N) in Figure 4(g). One can see that $\tilde{T}$ increases with N but remains below 10s. This well below the 1 minute interval between data points and explains how [43] could evaluate this method experimentally.

The heuristic rules can be replaced by statistical decision rules based on the MORIC. To this end, the average number of constraints $\bar{C}$ is computed for every model and every number of data points. Figure 4(d) shows this number of constraints divided by the number of spline function coefficients. One can see that this fraction is high (above 0.9) when the complete data series is available. This shows that the effective degrees of freedom for the models fit to the complete data series is far smaller ($B - \bar{C} \leq 9$) than the number of spline coefficients ($K = 126$). The MORIC, computed from $RMSR$ and $\bar{C}$, is shown in Figure 4(e). This enables execution of the following rules: *(i)* QS *UD* is preferred over QS *U* when $MORIC_{\mathsf{UD}} < MORIC_{\mathsf{U}}$ and *(ii)* QS *UDA* is preferred over QS *UD* when $MORIC_{\mathsf{UDA}} < MORIC_{\mathsf{U}}$. The resulting decision is shown in Figure 4(f). The MORIC-based acceptance of the *UDA*-based model occurs for the first time after $N = 553$ data points are collected, requiring 264 fewer data points than the decision based on heuristic rules. Applying the same wait time of 2h means that the setpoint increase would follow after $N = 673$ data points are collected, or equivalently, at 11 hours and 13 minutes into the control cycle. This corresponds to a reduction of the cycle length of more than 25%. In turn, this means that the efficiency, i.e. the volume of wastewater treated per time unit, could increase by 20% or more when the trend-based controller is equipped with the MORIC instead of the heuristic rules. The computational cost is substantially higher now due to the need to solve random instances of problem (5-16). The average computational cost for a random instance ($\bar{T}$) is similar to the equivalent problems using the original data series ($\tilde{T}$). However, due to using $R = 1200$ random instances to compute $\bar{C}$, the total computational cost is around 2h in the worst case. Thus, the MORIC-based controller requires that all random problem instances are solved a priori and only the problem instances involving the measured data are solved in real-time.

Results with method B (QPE). The QPE method was designed to be a fast method for QTA and this is obvious in Figure 4(g) as its computational cost is less one second for all instances. This makes online deployment easy to achieve. Similar to the SCS method, QPE constrains the resulting QS to take one of three forms: U, UD, and UDA. This means the control logic is very similar. As in [47], the QS with maximum posterior likelihood is selected and, once selected, the control logic is the same as above. Figure Figure 4(f) shows the selected QS as a function of time. The QS UDA is considered best for the first time after 3h and 20 minutes but the simpler QS UD is selected after 4h. The QS UDA is selected again at 4.5h without return to a simpler QS after. Using the same wait time as in [43] means that the setpoint increase would follow after $N = 345$ data points are collected, or equivalently, at 5 hours and 45 minutes into the control cycle. This would correspond to a reduction of the cycle length of more than 50%. However, this decision would likely be premature and therefore risk-prone.

Results with method C (PWP). Similar to the QPE method, the PWP method is designed to deliver fast segmentation results. Figure 4(g) shows that this method requires less than 0.2 seconds in all instances. However, one cannot limit the resulting segmentation to a set of predefined QS (e.g., U, UD, and UDA). This is illustrated by Figure 5, which shows a histogram of the 12 distinct QS produced by solving the 123 instances of the problem. Because one cannot predict which form the resulting segmentation will take, one cannot design a simple feedback control logic based on this QTA method.

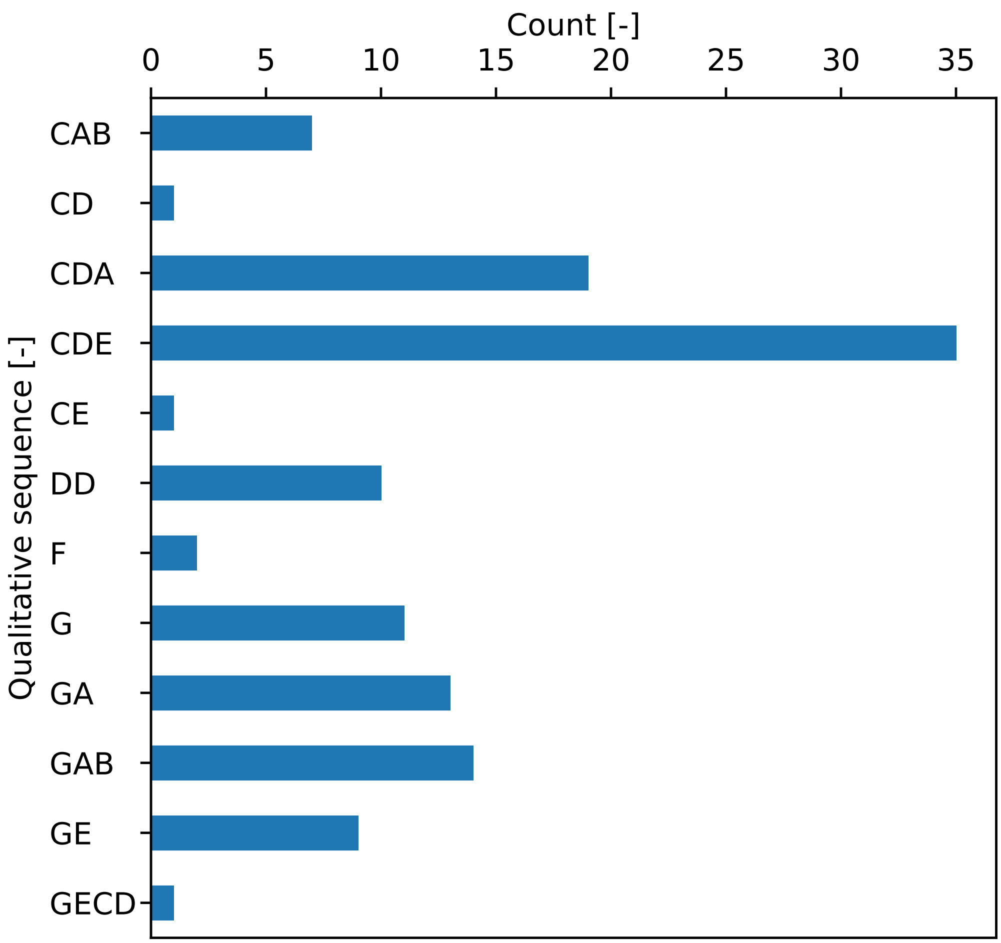


Figure 5: Nitrite concentration data series analyzed with method C (PWP): Histogram showing the QS and their frequency. A total of 12 different QS are obtained over the course of the time series, thus making an effective control strategy infeasible.

4. Discussion

4.1. Strengths and weaknesses of the SCS method

4.1.1. Strengths

Among the strengths of the SCS-based method for QTA, one can count:

1. *Mathematically precise definition of QTA.* The MIQP formulation of SCS function fitting, as expressed by (5-16), offers a compact, precise, and transparent interpretation of the QTA problem. Moreover, the MIQP formulation allows to clearly distinguish the problem definition (equations) from the applied solution strategy (algorithm for optimization). This is unlike many existing approaches to QTA based on heuristics, where the distinction between problem definition and solution strategy is not always obvious.
2. *Optimal function fitting and segmentation.* A provable and globally optimal solution to SCS function estimation is available for data series of moderate length, thanks to the existence of state-of-the-art optimization software and efficient computing hardware.
3. *Optimal selection of competing models.* The MORIC, like the GORIC, is conceptually equivalent to the AIC. This means that an intuitive and sound method for selection among competing SCS models is available.

4.1.2. Weaknesses

The most important weakness of the proposed method is that solving MIQPs is NP hard. This has two consequences:

1. *Exponential scaling of computational cost.* In the worst case, the required computational resources will scale exponentially with the length of the studied data series (N) and the maximal length of the QS (L). In turn, this limits the range of the proposed methods to cases where computational resources are available and cheap.
2. *Numerical approximation.* The MORIC, like the GORIC, requires an expensive, numerical approximation to compute its penalty term. While this can be computed off-line, this limits the broad applicability of the MORIC in current form due to the exponential scaling discussed above.

4.2. Future work

1. *Uncertainty analysis.* The MIQP formulation is interpreted as a maximum likelihood solution. In turn, this corresponds to finding the mode of a mixture of truncated multivariate normal densities. Thus, SCS function fitting has close ties to the sampling from mixtures of truncated densities [30, 31]. In turn, this means uncertainty analysis of SCS functions is within reach.
2. *Methods delivering accuracy at low computational cost.* The recognition that QTA can be formulated as an MIQP highlights why QTA problems are resistant to a solution that is simultaneously efficient and optimal, this despite the apparent simplicity of qualitative segmentation by the human eye. This explains the variety of historical methods to solve the QTA problem. Today, the practitioner is has the choice between a

provably optimal method at exceptionally high computational cost. of which the SCS method with MORIC selection is an example, or one of the historical methods whose computational cost is minimal yet may not deliver acceptable results while requiring expert-based tuning. New methods in between these two extremes would be very valuable.

3. *Benchmarking* The proposed MIQP method can be understood as a provably optimal yet computationally demanding method for QTA. In our opinion, this means the proposed method likely resides at the extreme right of a Pareto front describing all available QTA methods by plotting accuracy against computational efficiency. This remains to be evaluated empirically however, possibly by means of benchmark data sets.

4. *Analytical expression for the degrees of freedom.* In Section 3.1, the average active number of constraints ($\bar{C}$) was shown to exhibit a relatively smooth profile as a function of the model capacity (L). This could mean that alternative estimates of $\bar{C}$ (or D) exist. In our opinion, an exact yet tractable solution is unlikely to exist. However, approximate analytical expression may be available. If true, the computational efforts could be reduced drastically.

5. *Closed-loop validation.* Section 3.2 indicated that the MORIC may lead to significant efficiency improvements over the existing control scheme based on qualitative trend analysis tested in [43]. Naturally, these projected improvements should be evaluated experimentally.

6. *Data-driven identification of qualitative process models.* The graph $\mathcal{G}$ introduced above is very similar to the state graphs obtained through simulation of qualitative differential equations (QDE) [21], also known as qualitative simulation (QSIM). So far, QDEs have only been identified by abstraction from ordinary differential equations or by explicit encoding of prior knowledge. Combining *(a)* QSIM to produce graphs taking the form of $\mathcal{G}$ and *(b)* the MORIC-based mechanism to select an optimal graph $\mathcal{G}$ might enable the identification of a QDE model based on empirical observations. Similar concepts may be conceived for data-based identification of other qualitative dynamic models [6, 16, 24, 41]
.

5. Conclusion

In this study, a new method for qualitative trend analysis has been proposed and evaluated. As a result, the following novel conclusions can be made:

- Qualitative trend analysis can be formulated as a spline function fitting problem involving continuous and integer parameters. It is recognized for the first time that this type of time series segmentation task can be solved as a mixed integer quadratic program (MIQP).
- The newly proposed mixed order restricted information criterion (MORIC) enables the selection of an adequate model among a set of shape con-

strained spline models.

- The MORIC is the first objective model selection criterion for qualitative trend analysis and is therefore a valuable alternative to historical approaches based on heuristic criteria.

A. Appendix: Development of the shape constrained spline (SCS) function model

We assume the definitions and notations provided in the main text, Section 2.1 in particular. In what follows, we discuss the step-wise construction of the SCS function estimation problem when a spline function of degree 1 is assumed, as in the main text.

A.1. Construction of the model in bilinear form

Objective function and spline basis. The best-fitting spline function is obtained by minimizing the least-squares fit to the data series as follows:

$$\hat{\boldsymbol{b}} := \arg\min_{\boldsymbol{b}} \sum_{i}^{N} \left(f(t_i, \boldsymbol{b}) - \tilde{y}_i\right)^2 \tag{22}$$

$$f(t_i, \boldsymbol{b}) = b_k + \frac{t_i - q_k}{q_{k+1} - q_k}\,(b_{k+1} - b_k) \quad (q_k \le t_i < q_{k+1}) \quad (i = 1, \ldots, N) \tag{23}$$

where $\hat{\boldsymbol{b}}$ is the B-dimensional vector of spline coefficients.

Shape constraints. To control the shape of the linear spline function, the trend (increasing, decreasing) and curvature (convex, concave) in particular, it is sufficient to constrain the sign of point-wise differences. This is achieved by adding the following constraints to the above problem:

$$0 \leq +\frac{b_k - b_{k-1}}{q_k - q_{k-1}}\, s_k^{(1+)} \quad (k = 2, \ldots, K) \tag{24}$$

$$0 \leq -\frac{b_k - b_{k-1}}{q_k - q_{k-1}}\, s_k^{(1-)} \quad (k = 2, \ldots, K) \tag{25}$$

$$0 \leq +\left(\frac{b_{k+1} - b_k}{q_{k+1} - q_k} - \frac{b_k - b_{k-1}}{q_k - q_{k-1}}\right) s_k^{(2+)} \quad (k = 2, \ldots, K-1) \tag{26}$$

$$0 \leq -\left(\frac{b_{k+1} - b_k}{q_{k+1} - q_k} - \frac{b_k - b_{k-1}}{q_k - q_{k-1}}\right) s_k^{(2-)} \quad (k = 2, \ldots, K-1) \tag{27}$$

where $s_k^{(1+)}$, $s_k^{(1-)}$, $s_k^{(2+)}$, and $s_k^{(2-)}$ are boolean indicator variables specifying the sign of the first- and second-order differences at the knot q_k. For example, $s_k^{(1+)} = 1$ and $s_k^{(1-)} = 0$ specifies that the first derivative is non-negative while $s_k^{(1+)} = 0$ and $s_k^{(1-)} = 1$ specifies that the first derivative is non-positive. When both $s_k^{(1+)} = 1$ and $s_k^{(1-)} = 1$, the first derivative is constrained to be zero. The values for $s_k^{(2+)}$ and $s_k^{(2-)}$ specify the sign of the second derivative in the same fashion. For future reference, we combine the four boolean indicators into a single vector:

$$\boldsymbol{s}_k := \begin{bmatrix} s_k^{(1+)} & s_k^{(1-)} & s_k^{(2+)} & s_k^{(2-)} \end{bmatrix}^T \tag{28}$$

Feasible set of shapes. We aim to limit the complexity of the shape of the fitted spline function by constraining the feasible sets of signs for the first and second derivatives. To obtain a precise formulation, the unique shapes available for the problem at hand are defined first. These shapes are named primitives and defined by four-dimensional vectors $\boldsymbol{z}$. See Table 2 for details. The vectors are combined as columns in a $4 \times V$ matrix $\boldsymbol{Z}$. The selection of a primitive and assignment of the derivative signs is then written as

$$\boldsymbol{s}_k = \boldsymbol{Z}_{\cdot,x_k} \;\; (k = 1, \ldots, K)\,. \tag{29}$$

where x_k is an integer indicating the primitive selected at q_k and $\boldsymbol{Z}_{\cdot,x_k}$ selects the column vector with index x_k.

The complexity of the shape can be limited in three ways. First, one can limit which primitives are allowed to follow each other. In mathematical terms this becomes:

$$\boldsymbol{G}_{x_{k+1},x_k} = 1 \;\; (k = 1, \ldots, K - 1) \tag{30}$$

where is a $V \times V$-dimensional boolean matrix defined by the user. If $\boldsymbol{G}_{i,j} = 1$ then the ith primitive can be followed by the jth primitive. If $T_{i,j} = 0$, then such a sequence is not permitted. The matrix $\boldsymbol{G}$ can be visualized as a graph as discussed in the main text.

Second, the first and last primitive are constrained to belong to a subset of the available primitives. One writes:

$$x_1 \in \mathcal{V}_{start} \tag{31}$$

$$x_K \in \mathcal{V}_{end} \tag{32}$$

where $\mathcal{V}_{start}$ and $\mathcal{V}_{end}$ determine the primitives considered feasible at the start and end of the data series.

Third and last, one can limit the number of times the primitives change. This is written as:

$$c_k = \begin{cases} 0 & \text{if: } x_k = x_{k+1} \\ 1 & \text{otherwise} \quad (k = 1, \ldots, K-1) \end{cases}$$
$$L \geq \sum_{k=1}^{K-1} c_k + 1 \tag{33}$$

where c_k is a boolean variable indicating the presence of a primitive change and L is a user-set parameter limiting the number of times the selected primitive changes from one knot to the next.

Complete problem. The complete problem can now be written as follows:

$$\hat{\boldsymbol{b}}, \hat{\boldsymbol{x}} := \arg\min_{\boldsymbol{b},\boldsymbol{x}} \sum_{i}^{N} \left(f(t_i, \boldsymbol{b}) - \tilde{y}_i\right)^2 \tag{34}$$

$$f(t_i, \boldsymbol{b}) = b_k + \frac{t_i - q_k}{q_{k+1} - q_k}\,(b_{k+1} - b_k) \quad (q_k \leq t_i < q_{k+1}) \;\; (i = 1, \ldots, N) \tag{35}$$

$$0 \leq +\frac{b_k - b_{k-1}}{q_k - q_{k-1}}\, s_k^{(1+)} \quad (k = 2, \ldots, K) \tag{36}$$

$$0 \leq -\frac{b_k - b_{k-1}}{q_k - q_{k-1}}\, s_k^{(1-)} \quad (k = 2, \ldots, K) \tag{37}$$

$$0 \leq +\left(\frac{b_{k+1} - b_k}{q_{k+1} - q_k} - \frac{b_k - b_{k-1}}{q_k - q_{k-1}}\right) s_k^{(2+)} \quad (k = 2, \ldots, K-1) \tag{38}$$

$$0 \leq -\left(\frac{b_{k+1} - b_k}{q_{k+1} - q_k} - \frac{b_k - b_{k-1}}{q_k - q_{k-1}}\right) s_k^{(2-)} \quad (k = 2, \ldots, K-1) \tag{39}$$

$$\boldsymbol{s}_k = \boldsymbol{Z}_{\cdot,x_k} \quad (k = 1, \ldots, K) \tag{40}$$

$$\boldsymbol{G}_{x_{k+1},x_k} = 1 \quad (k = 1, \ldots, K-1) \tag{41}$$

$$x_1 \in \mathcal{V}_{start} \tag{42}$$

$$x_K \in \mathcal{V}_{end} \tag{43}$$

$$c_k = \begin{cases} 0 & \text{if: } x_k = x_{k+1} \\ 1 & \text{otherwise} \quad (k = 1, \ldots, K-1) \end{cases} \tag{44}$$

$$L \geq \sum_{k=1}^{K-1} c_k + 1 \tag{45}$$

where t_i and y_i represent the data series, $\boldsymbol{G}$, $\mathcal{V}_{start}$, $\mathcal{V}_{end}$ and , $\boldsymbol{Z}$ are user-defined inputs, and $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{x}}$ are the optimization variables.

A.2. Reformulation of bilinear constraints

The above problem contains $4K - 6$ bilinear constraints (36-39) and thus takes the form of a mixed-integer bilinear program. Fortunately, these constraints can be made linear by applying the Big-M approach [see e.g. 13]. To this end, the magnitude of the first- and second-order differences are bounded to be smaller than an arbitrary large absolute constant, named M. Then, the above problem can be reformulated without bilinear constraints as:

$$\hat{\boldsymbol{b}}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{x}} := \arg\min_{\boldsymbol{b},\boldsymbol{x}} \sum_{i}^{N} \left(f(t_i, \boldsymbol{b}) - \tilde{y}_i\right)^2 \tag{46}$$

$$f(t_i, \boldsymbol{b}) = b_k + \frac{t_i - q_k}{q_{k+1} - q_k}\ (b_{k+1} - b_k) \quad (q_k \leq t_i < q_{k+1}) \ \ (i = 1, \ldots, N) \tag{47}$$

$$-M \leq +\frac{b_k - b_{k-1}}{q_k - q_{k-1}} - M\, s_k^{(1+)} \quad (k = 2, \ldots, K) \tag{48}$$

$$-M \leq -\frac{b_k - b_{k-1}}{q_k - q_{k-1}} - M\, s_k^{(1-)} \quad (k = 2, \ldots, K) \tag{49}$$

$$-M \leq +\left(\frac{b_{k+1} - b_k}{q_{k+1} - q_k} - \frac{b_k - b_{k-1}}{q_k - q_{k-1}}\right) - M\, s_k^{(2+)} \quad (k = 2, \ldots, K-1) \tag{50}$$

$$-M \leq -\left(\frac{b_{k+1} - b_k}{q_{k+1} - q_k} - \frac{b_k - b_{k-1}}{q_k - q_{k-1}}\right) - M\, s_k^{(2-)} \quad (k = 2, \ldots, K-1) \tag{51}$$

$$\boldsymbol{s}_k = \boldsymbol{Z}_{\cdot, x_k} \quad (k = 1, \ldots, K) \tag{52}$$

$$\boldsymbol{G}_{x_{k+1}, x_k} = 1 \quad (k = 1, \ldots, K-1) \tag{53}$$

$$x_1 \in \mathcal{V}_{start} \tag{54}$$

$$x_K \in \mathcal{V}_{end} \tag{55}$$

$$c_k = \begin{cases} 0 & \text{if: } x_k = x_{k+1} \\ 1 & \text{otherwise} \quad (k = 1, \ldots, K-1) \end{cases} \tag{56}$$

$$L \geq \sum_{k=1}^{K-1} c_k + 1 \tag{57}$$

This makes all constraints in the optimization problem linear. As the objective function is quadratic and the problem involves both continuous and integer variables, the above optimization problem is now a mixed-integer quadratic program (MIQP).

A.3. Additional remarks.

The definition of L, $\boldsymbol{G}$, $\mathcal{V}_{start}$, and $\mathcal{V}_{end}$ may render the feasible space for $\hat{\boldsymbol{x}}$ empty in some cases, in turn leading to an infeasible optimization problem. For the problem to be feasible, at least one sequence of primitives (x_k) must satisfy (52-57). We name this the reachability requirement and assume it is always met through the user-provided inputs.

If the reachability requirement is met, then the above MIQP is always feasible (without proof). As a result, it holds that $0 \leq C_r \leq B$ always. In turn, this means that the estimated degrees of freedom, $\bar{D} = N - 1 - \bar{C}$, is bound to lie between 1 and $B + 1 = K + 1 = N - 1$.

Acknowledgments

The authors thank the developers of Python and the associated *matplotlib*, *networkx*, and *numpy* packages. The authors thank Per Rutquist (Tomlab Software AB) for suggesting the Big-M approach and thank Ingrid Busch, Nick Diangelis, and Bernd Gärtner for helpful comments. The authors thanks Shubham Kashyapi, Lokesh Rajulapati, Raghunathan Rengaswamy for detailed information on the interval-halving method for piece-wise polynomial representation.

References

[1] Balabdaoui, F., Rufibach, K., and Santambrogio, F. (2010). Least-squares estimation of two-ordered monotone regression curves. *Journal of Non-*

parametric Statistics, 22(8):1019–1037.

[2] Bonvin, D., Georgakis, C., Pantelides, C. C., Barolo, M., Grover, M. A., Rodrigues, D., Schneider, R., and Dochain, D. (2016). Linking models and experiments. *Industrial & Engineering Chemistry Research*, 55(25):6891–6903.

[3] Dash, S., Maurya, M. R., Venkatasubramanian, V., and Rengaswamy, R. (2004). A novel interval-halving framework for automated identification of process trends. *AIChE Journal*, 50(1):149–162.

[4] de Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag, New York, NY, USA.

[5] Derlon, N., Thürlimann, C. M., Dürrenmatt, D. J., and Villez, K. (2017). Batch settling curve registration via image data modeling. *Water Research*, 114:327–337.

[6] Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence*, 24(1-3):85–168.

[7] Gaylord, C. K. and Ramirez, D. E. (1991). Monotone regression splines for smoothed bootstrapping. *Computational Statistics Quarterly*, 6(2):85–97.

[8] Gurobi Optimization, LLC (2021). Gurobi optimizer reference manual.

[9] Habtzghi, D. and Datta, S. (2012). One sample goodness of fit tests in presence of shape restrictions on the hazard rate function. *Sankhya B*, 74:171–194.

[10] Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J., editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

[11] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R'ıo, J. F., Wiebe, M., Peterson, P., G'erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

[12] Hazelton, M. L. and Turlach, B. A. (2011). Semiparametric regression with shape constrained penalized splines. *Computational Statistics & Data Analysis*, 55(10):2871–2879.

[13] Hooker, J. (2000). *Logic-based methods for optimization: combining optimization and constraint satisfaction*. John Wiley & Sons, New York, NY, USA.

[14] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9:90–95.

[15] Jankowski, H. K. and Wellner, J. A. (2009). Computation of nonparametric convex hazard estimators via profile methods. *Journal of Nonparametric Statistics*, 21(4):505–518.

[16] Kansou, K. and Bredeweg, B. (2014). Hypothesis assessment with qualitative reasoning: Modelling the Fontestorbes fountain. *Ecological Informatics*, 19:71–89.

[17] Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, 46:1071–1085.

[18] Knafl, G., Sacks, J., Spiegelman, C., and Ylvisaker, D. (1984). Nonparametric calibration. *Technometrics*, 26(3):233–241.

[19] Kuiper, R. M., Hoijtink, H., and Silvapulle, M. J. (2011). An Akaike-type information criterion for model selection under inequality constraints. *Biometrika*, 98(2):495–501.

[20] Kuiper, R. M., Hoijtink, H., and Silvapulle, M. J. (2012). Generalization of the order-restricted information criterion for multivariate normal linear models. *Journal of Statistical Planning and Inference*, 142(8):2454–2463.

[21] Kuipers, B. (1986). Qualitative simulation. *Artificial Intelligence*, 29(3):289–338.

[22] Liao, X. and Meyer, M. C. (2017). Change-point estimation using shape-restricted regression splines. *Journal of Statistical Planning and Inference*, 188:8–21.

[23] Mašić, A., Srinivasan, S., Billeter, J., Bonvin, D., and Villez, K. (2017). Shape constrained splines as transparent black-box models for bioprocess modeling. *Computers & Chemical Engineering*, 99:96–105.

[24] Maurya, M. R., Rengaswamy, R., and Venkatasubramanian, V. (2003). A systematic framework for the development and analysis of signed digraphs for chemical processes - 1. Algorithms and analysis. *Industrial & Engineering Chemistry Research*, 42(20):4789–4810.

[25] Maurya, M. R., Rengaswamy, R., and Venkatasubramanian, V. (2007). Fault diagnosis using dynamic trend analysis: A review and recent developments. *Engineering Applications of Artificial Intelligence*, 20(2):133–146.

[26] Meyer, M. C. (2012). Nonparametric estimation of a smooth density with shape restrictions. *Statistica Sinica*, 22(2):681–701.

[27] Meyer, M. C. and Habtzghi, D. (2011). Nonparametric estimation of density and hazard rate functions with shape restrictions. *Journal of Nonparametric Statistics*, 23(2):455–470.

[28] Meyer, M. C. and Woodroofe, M. (2004). Consistent maximum likelihood estimation of a unimodal density using shape restrictions. *Canadian Journal of Statistics*, 32(1):85–100.

[29] Nesterov, Y. (2000). *High Performance Optimization, Applied Optimization*, volume 33, chapter Squared functional systems and optimization problems, pages 405–440. Kluwer Academic Publishers.

[30] Pakman, A. and Paninski, L. (2013). Auxiliary-variable exact Hamiltonian Monte Carlo samplers for binary distributions. In *Advances in Neural Information Processing Systems (NIPS2013)*, pages 2490–2498.

[31] Pakman, A. and Paninski, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542.

[32] Papp, D. (2011). *Optimization models for shape-constrained function estimation problems involving nonnegative polynomials and their restrictions*. PhD thesis, Rutgers University, New Brunswick, NJ, USA.

[33] Papp, D. and Alizadeh, F. (2014). Shape-constrained estimation using nonnegative splines. *Journal of Computational and Graphical Statistics*, 23(1):211–231.

[34] Petersen, B., Gernaey, K., Devisscher, M., Dochain, D., and Vanrolleghem, P. A. (2003). A simplified method to assess structurally identifiable parameters in Monod-based activated sludge models. *Water Research*, 37(12):2893–2904.

[35] Pflug, G. C. and Wets, R. J. B. (2013). Shape-restricted nonparametric

regression with overall noisy measurements. *Journal of Nonparametric Statistics*, 25(2):323–338.

[36] Python Software Foundation (2021). Python language reference, version 3.8.5.

[37] Ramsay, J. and Hooker, G. (2017). *Dynamic Data Analysis: Modeling Data with Differential Equations*. Springer, New York, NY, USA.

[38] Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer, New York, NY, USA.

[39] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York, NY, USA.

[40] Sabbadin, R. (2001). Possibilistic Markov decision processes. *Engineering Applications of Artificial Intelligence*, 14(3):287–300.

[41] Shaich, D., Becker, R., and King, R. (2001). Qualitative modelling for automatic identification of mathematic models of chemical reaction systems. *Control Engineering Practice*, 9(12):1373–1381.

[42] Silvapulle, M. J. and Sen, P. K. (2005). *Constrained statistical inference: Order, inequality, and shape constraints*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

[43] Thürlimann, C. M., Udert, K. M., Morgenroth, E., and Villez, K. (2019). Handling sensor drift for stabilizing nitrite control during nitrification of

high strength wastewater by means of qualitative trend analysis. *Water Research*, 165:114958.

[44] Thürlimann, C. M. and Villez, K. (2017). Input estimation as a qualitative trend analysis problem. *Computers & Chemical Engineering*, 107:333–342.

[45] Turlach, B. A. (2005). Shape constrained smoothing using smoothing splines. *Computational Statistics*, 20:81–103.

[46] Venkatasubramanian, V., Rengaswamy, R., and Kavuri, S. N. (2003). A review of process fault detection and diagnosis – Part II: Qualitative models and search strategies. *Computers & Chemical Engineering*, 27(3):313–326.

[47] Villez, K. (2015). Qualitative path estimation: A fast and reliable algorithm for qualitative trend analysis. *AIChE Journal*, 61(5):1535–1546.

[48] Villez, K. (2020). Computation and illustration of the generalized order restricted criterion (GORIC) for shape constrained spline functions (ORNL/TM-2020/1541). Technical report, Oak Ridge National Laboratory.

[49] Villez, K. and Habermacher, J. (2016). Shape anomaly detection for process monitoring of a sequencing batch reactor. *Computers & Chemical Engineering*, 91:365–379.

[50] Villez, K., Venkatasubramanian, V., and Rengaswamy, R. (2013). Generalized shape constrained spline fitting for qualitative analysis of trends. *Computers & Chemical Engineering*, 58:116–134.

S. Qualitative Trend Analysis based on a Mixed-Integer Representation – Supplementary Materials

S.1. Python package

The Python package associated with this study (v1.0 and higher) is named pyQTA and includes all code and data used to produce the results in this work, including parameter estimation for shape constrained spline function fitting and model selection by means of the MORIC. The most recent version of the package is publicly available at `https://gitlab.com/krisvillez/sklearn-qta` [Note to editor: public release pending peer review result].

S.2. Additional results with the SCS model

The next paragraphs discuss some additional results obtained with shape constrained spline models not presented in the main text.

S.2.1. Nitrogen concentration data set

Figure S.1 shows the selected QR for the nitrite concentration data series as a function of the number of data points analyzed.

S.2.2. Refinery data set

This section describes results obtained with the shape-constrained spline method for the refinery data described in [39]. This data set has been used for illustrative purposes in many studies: [39] use it to illustrate how discontinuous behavior can be described by a spline function; [37] use it to explain dynamic model identification methods; and [49] use it to illustrate a shape constrained spline function fitting in the presence of discontinuities. In this work, the data series is used to illustrate the methods for shape constrained spline function fitting, i.e. the material in Section 2.2.2.

Data series. The data series consists of $N = 193$ tray level measurements in an oil refinery distillation column. The measurements (y_i) are recorded equidistantly in time. The data $(\boldsymbol{t}, \boldsymbol{y})$ are shown in Figure S.2(a).

User-specified inputs. The assumed QS to describe these data is given as FBGC, meaning that the graph $\mathcal{G}$, after removal of self-loops, is the directed

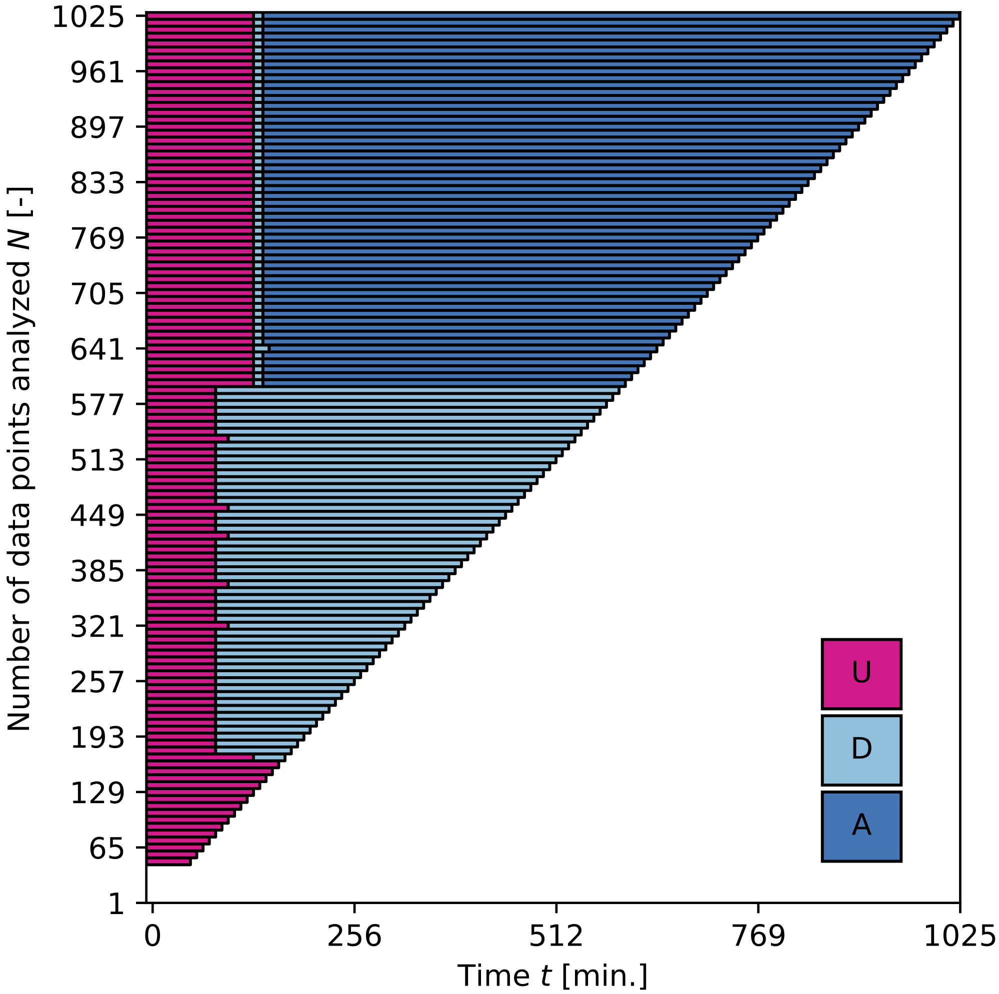


Figure S.1: Nitrite concentration data series – Results with the SCS method: Evolution of the selected qualitative representation (QR) as a function of the number of data points available for analysis.

acyclic graph shown in Figure S.2(d). The corresponding transition matrix $\boldsymbol{G}$ is:

$$\boldsymbol{G} = \begin{array}{c} \begin{array}{cccc} \mathsf{F} & \mathsf{B} & \mathsf{G} & \mathsf{C} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{array} \begin{array}{c} \\ \mathsf{F} \\ \mathsf{B} \\ \mathsf{G} \\ \mathsf{C} \end{array} \tag{58}$$

where the primitives (F, B, G, and C) indicate the corresponding vertices.

In addition, the first and last state are constrained to correspond to the first and last primitive in the QS:

$$\mathcal{V}_{start} = \{1\} \tag{59}$$

$$\mathcal{V}_{end} = \{4\} \tag{60}$$

In this case, L is set to an arbitrary large number exceeding the length of the QS ($L > 4$). Consequently, Eq. 16 has no impact on the location of the optimal solution.

Results. Solving problem (5-16) took less than 4 seconds and leads to the segmentation shown in Figure S.2(c). The fitted function is shown in Figure S.2(a) while the first order differences are shown in Figure S.2(b). It is verified easily that both the fitted spline function and the resulting segmentation satisfy the constraints imposed through $\mathcal{G}$, $\mathcal{V}_{start}$, and $\mathcal{V}_{end}$.

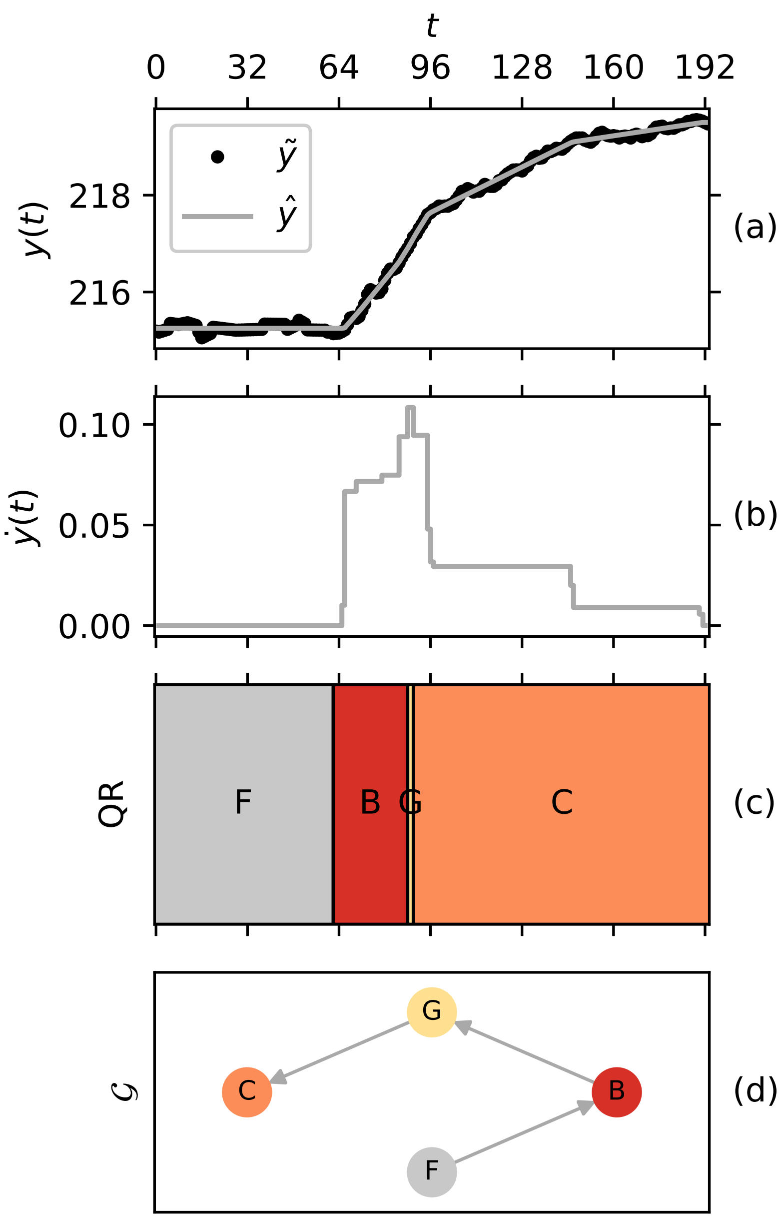


Figure S.2: Results obtained with the refinery data series: *(a)* Data series ($\tilde{\boldsymbol{y}}$) and fitted function ($\hat{\boldsymbol{y}}$); *(b)* First derivative of fitted function; *(c)* Qualitative representation (QR); *(d)* Graph $\mathcal{G}$, shown here without self-loops.

S.3. Tuning and detailed results with qualitative path estimation (QPE)

S.3.1. Tuning

Similar to the proposed SCS method, QPE requires specification of a graph $\mathcal{G}$, start vertices $\mathcal{V}_{start}$, and end vertices $\mathcal{V}_{end}$ describing the feasible QSs. The QPE method does not support the use of primitives that specify zero-valued derivatives so the primitives E, F, G are not permitted. The graphs used in this study are explained below.

In addition, the QPE method requires the specification of the noise standard deviation, σ_ϵ. In contrast to the original method, we estimate its value as the median absolute difference:

$$\begin{aligned} \hat{\sigma}_\epsilon &= \text{Median}\left(E\right) \\ E &= \left\{\left|\tilde{y}_2 - \tilde{y}_1\right|, \ldots, \left|\tilde{y}_{k+1} - \tilde{y}_k\right|, \ldots, \left|\tilde{y}_N - \tilde{y}_{N-1}\right|\right\} \end{aligned} \tag{61}$$

Finally, the QPE method allows to specify a prior likelihood for the first and final vertex. By default, we assume a uniform prior for each vertex in $\mathcal{V}_{start}$ and $\mathcal{V}_{end}$, i.e. each vertex is considered equally probable.

S.3.2. Case I: Titanium data set

Data series and user-specified inputs. The data series is the same as in the main text and shown in Figure S.3(a). The user-specified inputs consists of the graph shown in Figure S.3(e) and the feasible start and end vertices, which are all vertices in the graph. The prior likelihood for the start and end vertices are uniform.

Results. Figure S.3(b) shows the point-wise probability for each vertex/primitive computed in the first step of the QPE method. In the second step, these are analyzed by means of the Viterbi algorithm which takes the user-specified graph into account as well as the prior probabilities. The resulting Viterbi path is shown in Figure S.3(c). The corresponding QR is visualized in Figure S.3(d).

Following segmentation with the QPE method, an SCS model is fit while making sure the selected vertices ($\hat{x}$) match the Viterbi path. This means problem (5-16) is solved as convex quadratic program (QP) with fixed values for $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{s}}$. The result of this is shown in Figure S.4. The resulting function fits the data well (Figure S.4(a)) but the first derivative exhibits spurious fluctuations ((Figure S.4(b)) as can be expected from the applied QR ((Figure S.4(c)).

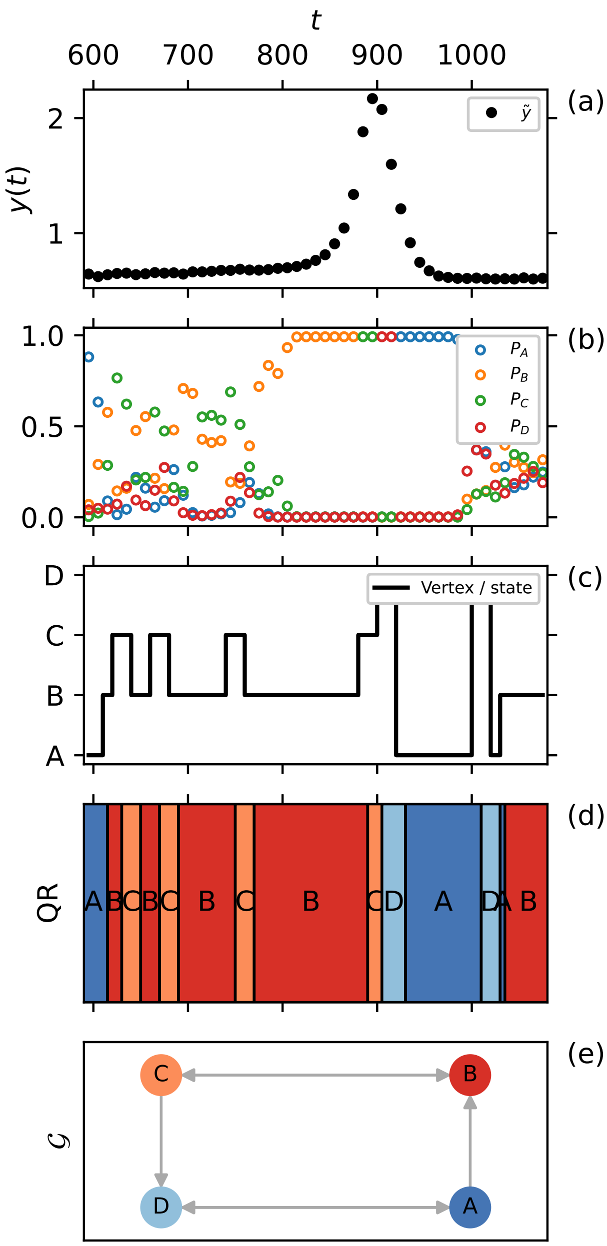


Figure S.3: Titanium data set – Results with the QPE method: *(a)* data series; *(b)* point-wise probabilities; *(c)* Viterbi path; *(d)* qualitative representation (QR); *(e)* applied graph to constrain the solution.

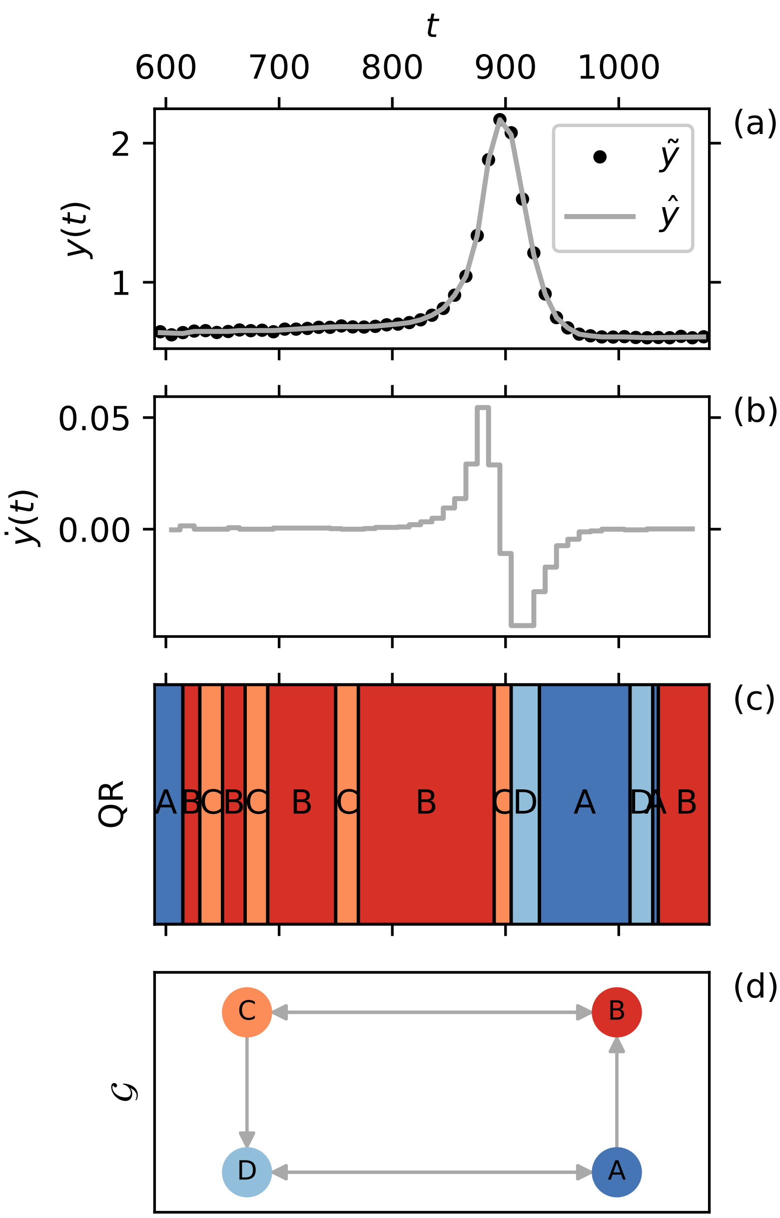


Figure S.4: Titanium data set – Results with the SCS method when using the QPE-based segmentation: *(a)* data series and fitted function; *(b)* first derivative; *(c)* qualitative representation (QR) used as a constraint on the spline function model.

S.3.3. Case II: Nitrogen concentration data set

Data series and user-specified inputs. The data series is the same as in the main text and shown in Figure 4(a). The user-specified inputs consists of the graph shown in Figure 4(b) and the feasible start and end vertices are the same as applied for the SCS method, i.e. start vertices constrained to be the first vertex (U primitive) and feasible end vertices any of the three vertices (primitives U, D, and A). The prior likelihood for the start vertex is naturally uniform (only one vertex in $\mathcal{V}_{start}$). The prior likelihood for the end vertex is defined by a broken-stick prior, favoring a shorter QS over longer ones:

$$L_{prior}\left(x_K = v\right) = \sum_{i=v}^{V} \frac{1}{i}, \quad v = 1, \ldots, V \tag{62}$$

where $v = 1, \ldots, 3$ match the primitives U, D, and A.

Results. Figure S.5 shows the 123 QRs obtained with the QPE method. As soon as 330 data points are collected, the resulting QR is essentially the same.

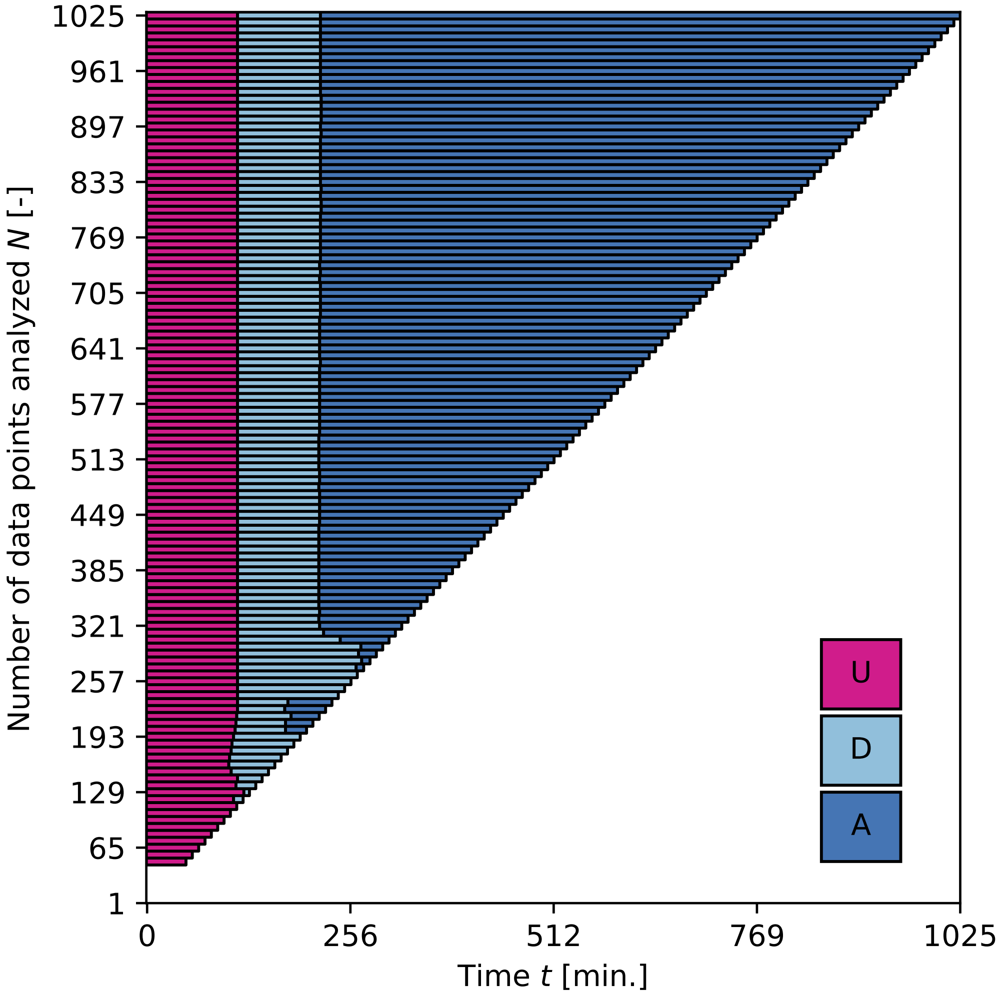


Figure S.5: Nitrite concentration data series – Results with the SCS method: Evolution of the selected qualitative representation (QR) as a function of the number of data points available for analysis.

S.4. Tuning and detailed results with piece-wise polynomial segmentation through interval-halving (PWP)

The interval-halving method of [3] is based on a fast procedure to identify a piece-wise polynomial curve which describes a data series. The polynomial pieces of the curve can take zero, one, or two as degrees (constant, linear, quadratic segments). After identifying the polynomials, a QR is obtained by inspecting the signs of the first and second derivative of the polynomial segments and identifying the primitives A, B, C, D, E, F, and G (see Table 2). Since a single quadratic polynomial may exhibit both decreasing and increasing sections, it is possible that a single polynomial leads to two episodes in the resulting QR.

S.4.1. Tuning and implementation

The PWP method requires specification of a noise standard deviation and a number of significance levels. In this study, the noise standard deviation is estimated as with the PWP method Eq. 61. The remaining parameters and their values are listed in Table S.4.1. These match the values selected in the original work [3]. Note that the original method enables detection of jumps, i.e. large step changes, as well. However, the detailed implementation cannot be inferred from [3]. For this reason, this is omitted in this work.

S.4.2. Case I: Titanium data set

Data series and user-specified inputs. The data series is the same as in the main text and shown in Figure S.6(a). The user-specified inputs consists of

Table S.1: Tuning parameters in the PWP method.

Description	**Symbol**	**Value**
Significance level for the F-test to decide whether to increase the polynomial order	α	0.05
Significance level for the t-test to decide whether a constant trend is acceptable in the first of a pair of polynomial segments	α_1	0.8
Significance level for the t-test to decide whether a constant trend is acceptable in the second of a pair of polynomial segments	α_2	0.5
Minimum number of data points in a single unimodal segment	N_{min}	8

the tuning parameters described above.

Results. Figure S.6(a-b) show the boundaries of the identified polynomials with dashed lines. There are five polynomials (unimodal segments), of which the first and last have degree 1 while the inner ones have degree 2. To obtain the QR (Figure S.6(c)), the signs of the derivatives are analyzed. A single episode is produced when the derivative signs do not change within the support window of the polynomial. This is true for all polynomial segments except for the fourth one, where the first derivative crosses zero and thus changes sign. This means that the QR is ultimately composed of six episodes. The corresponding QS equals *GBDABE*.

S.4.3. Case II: Nitrogen concentration data set

Data series and user-specified inputs. The data series is the same as in the main text and shown in Figure 4(a). The user-specified inputs consists of the tuning parameters described above.

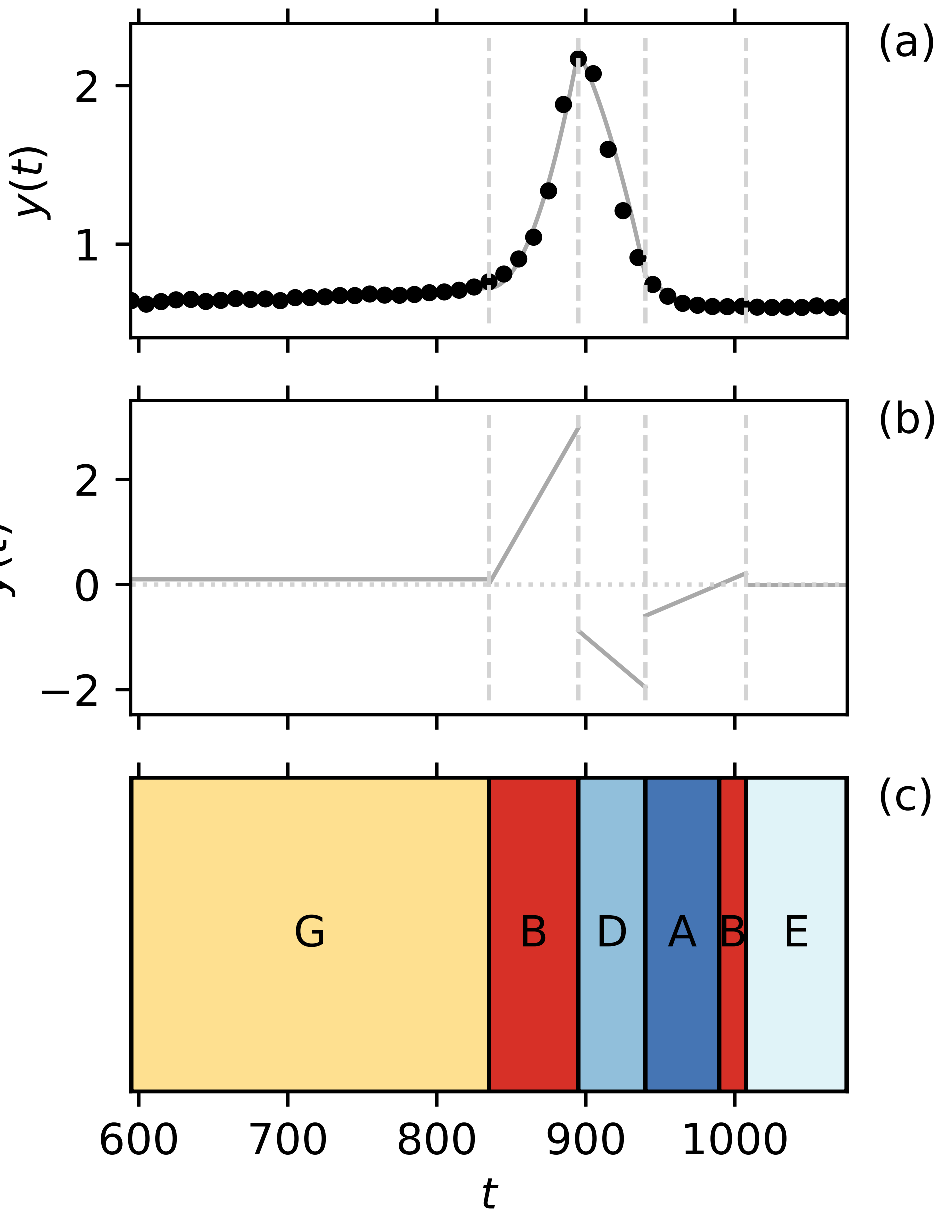


Figure S.6: Titanium data set – Results with the PWP method: *(a)* data series and fitted piwece-wise polynomial function; *(b)* first derivative; *(c)* qualitative representation (QR).

Results. Figure S.7 shows the 123 QRs obtained with the PWP method. The resulting QRs vary significantly as both the selected primitives and their location changes as more data points are collected. This is explained by the fact that this method cannot account for user-provided information on the feasible set of primitives and the order they should appear in.

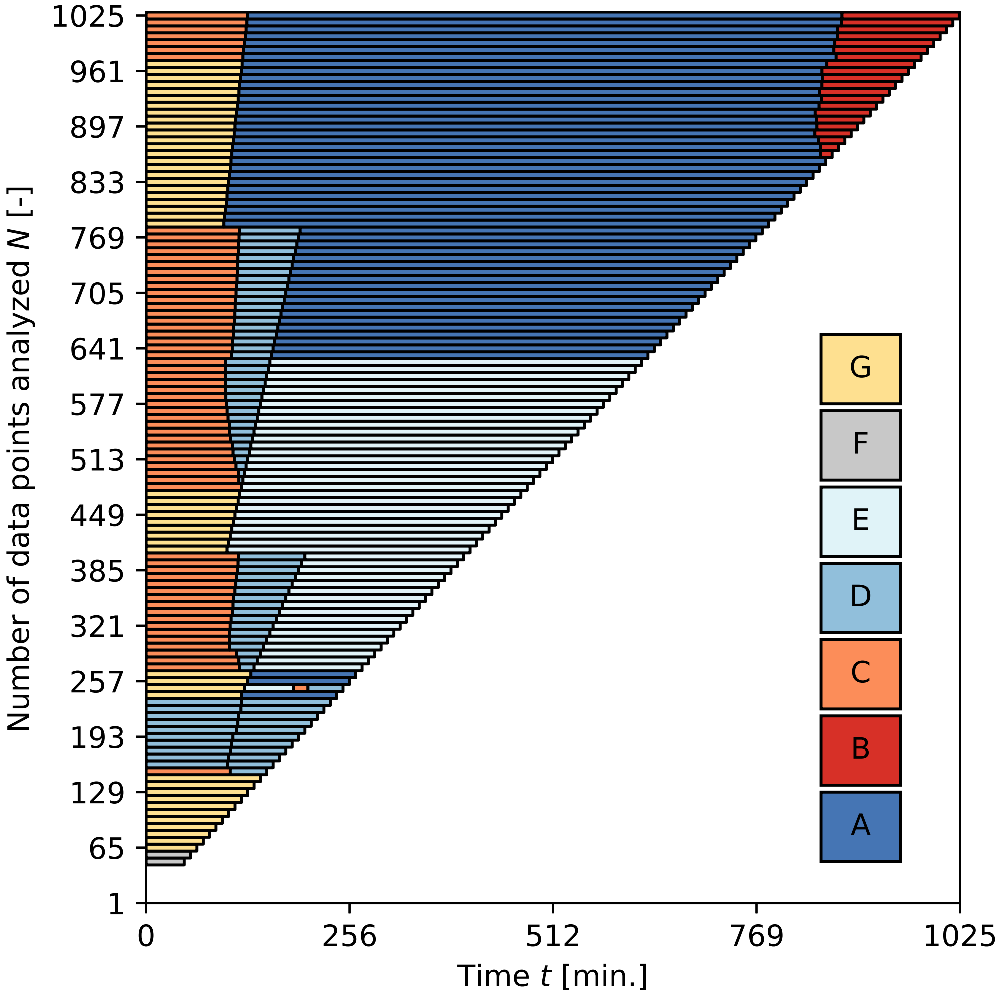


Figure S.7: Nitrite concentration data series – Results with the PWP method: Evolution of the selected qualitative representation (QR) as a function of the number of data points available for analysis.