Transforming data into information

M.B. Beck and Z. Lin
Warnell School of Forest Resources, University of Georgia, Athens, Georgia 30602-2152, USA
(E-mail: mbbeck@arches.uga.edu)

Abstract In spite of a long history of automated instruments being deployed in the water industry, only recently has the difficulty of extracting timely insights from high-grade, high-volume data sets become an important problem. Put simply, it is now relatively easy to be “data-rich”, much less easy to become “information-rich”. Whether the availability of so many data arises from “technological push” or the “demand pull” of practical problem solving is not the subject of discussion. The paper focuses instead on two issues: first, an outline of a methodological framework, based largely on the algorithms of (on-line) recursive estimation and involving a sequence of transformations to which the data can be subjected; and second, presentation and discussion of the results of applying these transformations in a case study of a biological system of wastewater treatment. The principal conclusion is that the difficulty of transforming data into information may lie not so much in coping with the high sampling intensity enabled by automated monitoring networks, but in coming to terms with the complexity of the higher-order, multi-variable character of the data sets, i.e., in interpreting the interactions among many contemporaneously measured quantities.

Keywords Biological wastewater treatment; filtering-smoothing algorithms; models; nitrification; recursive estimation; signal extraction

Introduction
To imagine one might have too many data – to be data-rich, yet information-poor – would have been a startling thought just a few years ago. Whatever the reasons for this change in our fortune, in both problem-solving opportunities and the different challenges following from these new opportunities, our concern in this paper is that of making the most sense of the great volumes of data to which we can now have access. We acknowledge the substantial interest in further developing the technologies for generating ever more intense and varied streams of water quality data; and we will presume that creating systems and procedures of analysis for acquiring insights into the practical implications of all these data, in real time, in particular, will be especially challenging. However, our focus herein is on addressing just two issues: an outline of one candidate methodological framework for such procedures of analysis, essentially founded upon the principles of recursive estimation (Young, 1984, 1998; see also Beck, 1987); and an assessment of its success for a single case study dealing primarily with aspects of biological nitrification in an activated sludge process of wastewater treatment.

Methodological framework
Suppose the behavior of the system being monitored can be defined according to the following (lumped-parameter) representation of the state variable dynamics,

\[ \frac{d\mathbf{x}(t)}{dt} = f(\mathbf{x}, \mathbf{u}, \mathbf{a}; t) + \mathbf{z}(t) \]  

(1a)

with observed outputs being defined as follows,

\[ y(t) = h[\mathbf{x}, \mathbf{a}; t] + \eta(t) \]  

(1b)
in which: \( f \) and \( h \) are vectors of nonlinear functions; \( u, x, \) and \( y \) are the input, state, and output vectors, respectively; \( \alpha \) is a vector of model parameters; \( \xi \) and \( \eta \) represent collectively all manner of unknown (unobserved) disturbances of the system’s behavior, errors of observation attaching to \([u,y]\), and conceptual (or structural) errors in the formulation of the model; and \( t \) is continuous time. Should it be necessary, spatial variability of the system’s state can be assumed to be accounted for by, for example, the use of several state variables of the same attribute of interest at several defined locations.

The data, in general, comprise the observed time series of discretely sampled measurements of all the inputs and outputs \([u,y]\) at each sampling instant in time \( t_k \) over some interval \( t_0 \) to \( t_N \), i.e., \( k = 0, 1, 2, ..., N \). They are the collection of numbers in the original time series, including the spurious and corrupting consequences arising from \([\xi,\eta]\), and those numbers missing (that would have been acquired but for some kind of failure) for some of the quantities for some of the points in time, \( t_k \). We gather these data, broadly speaking, for two rather different purposes: improving understanding; and making decisions leading to actions. The distinction drawn herein between “data” and “information” will be pragmatic.\(^1\) It is as follows: information comprises an insight or set of insights, drawn from a transformation of the data, and used to serve either of the foregoing purposes; further, this insight is distinctively different from any gained simply by inspection of the (untransformed) data; moreover, it should ideally represent a compression of knowledge and experience, from the very high dimension, order, and volume of \([u,y]\) to something much more succinct (typically a model).

A model of some form inevitably underpins the origin of the insight, although by no means must it be expressed in the formalities of Eq. (1). Unless the process of generating insight and taking decisive action are automated, as in (real-time) control, the model will be a mental model embedded in the mind of the observer/decision-maker. It takes the form of a bundle of reasoning, wherein causes \( (u) \) can be related to effects \( (y) \) – understanding in itself – from which comes the insight of how to manipulate some of the former to alter the values of some of the latter (the decision/action). For example, in order to distill the immense complexity of the formation of a bulking sludge into its essence, some highly insightful work was invested in forming a mental model of the competitive growths of filamentous and floc-forming bacteria: as that of the growth of two populations of dogs of two quite different physiologies and feeding habits (Jones and Franklin, 1985). From this high-level conceptual description a set of rules could be derived, by which to navigate the associated biological treatment process away from the potential failure of a bulking sludge.

The faster the occurrence of the disturbances impinging on the system \( (u) \), however, the more it will be necessary for the scheme of generating insights and making decisions to be automated, tending in the limit to what is familiar to us as (real-time) control. It would otherwise be impossible – through our own mental reasoning – to deduce the consequences of the disturbances before it became too late to enact any compensating controls. It is readily apparent that automated water quality monitoring underpins real-time control.

To summarize, against the reference frame of some sort of more or less informal mental model, or high-level conceptual description, we have the following progressive chain of transformations, from the raw, unexpurgated data, into the ever more refined terms of the inner workings of a formal mathematical model:

1. **raw data, \([u,y]\);**

---

\(^1\) Perhaps to the disappointment of those searching for a satisfying, philosophical or technical definition, as to the distinctive differences among the various elements in the chain of apprehending the nature of the world about us, that is, the chain defined by \{data-information-knowledge-wisdom\}.
smoothed and interpolated data, \([\hat{u}, \hat{y}]\), i.e., the products of preliminary processing designed to discriminate against the consequences of \([\xi, \eta]\), albeit weakly;

(iii) extracted (trend, periodic, seasonal) components of the data;

(iv) input–output, transfer-function models, or low-order models (LOMs), bearing parameters \((\gamma)\);

(v) data-based mechanistic models (Young, 1998), bearing parameters \((\gamma)\) capable of interpretation within the terms of Eq. (1);

(vi) the higher-order models (HOMs) of Eq. (1); and

(vii) the very high-order models (VHOMs) with states and parameters \([x, \alpha]\) numbering in the hundreds and thousands.

Here, the smoothed and interpolated data, \([\hat{u}, \hat{y}]\), the extracted signal components, and the parameters \((\gamma)\) associated with the data-based models are defined in scalar terms as (Young, 1998):

\[
\hat{u}(t_k) = u^T(t_k) + u^S(t_k) + u^P(t_k) \quad (2a)
\]

\[
\hat{y}(t_k) = y^T(t_k) + y^S(t_k) + y^P(t_k) + g\{u, \gamma; t_k\} \quad (2b)
\]

where superscripts \(T, S, \text{ and } P\) denote trend, seasonal and periodic components, respectively, and the transfer-function model relating input to output, denoted by the expression \(g\{u, \gamma; t_k\}\), is embraced as a part of Eq. (2b); by virtue of the boldface \(u\) the possibility of a multiple input/single output form of the model is acknowledged. More familiarly, one would encounter this model as Eq. (2b) without the other components of the output signal (other schemes of signal extraction are illustrated in the work of Quilty and Whitfield, 2001).

The nub of our problem is then: given very high-volume sets of data and a prior mental model of the observed system’s behavior, what insights – for the purposes of understanding or decision-making – are generated as we pass through the sequence of data transformations so defined above. In particular, the scope of the data base is distinctively defined by access to an automated network for monitoring water quality employing a high sampling frequency. It follows further therefrom that we shall be predisposed towards the use of methods of data processing and transformation designed to function efficiently on-line, in real-time, such as recursive estimation algorithms, for example (Young, 1984; see also Dochain and Vanrolleghem, 2002).

Case study

We report on just one case study, implemented with the Environmental Process Control Laboratory (EPCL) of the University of Georgia. In the context of the present discussion it is important to note that the EPCL was designed expressly for the purpose of providing near-continuous, fully automated monitoring of the multi-variable dynamics of environmental systems. Figure 1 shows time series for observations of the ammonium-N concentration at three locations – the crude sewage, first pocket of the biological reactor, and the third pocket thereof – across an activated sludge unit at the Athens Number 2 Wastewater Treatment Facility. Inspection of this figure ought to evoke the response: what does all this mean for understanding the behavior of the associated system? It is clear that each trace contains patterns of variation that are not the result of pure chance, i.e., there are patterns capable of some kind of deterministic interpretation. It is also readily apparent that the constituent time series are correlated in ways that ought to sustain a secure interpretation through a coherent and consistent model (mental or otherwise). Yet intense scrutiny of any one trace would be quite insufficient, because the other dozen or so traces (not shown)
are strongly suggestive of the consequences of the dense network of significant interactions typical of multi-variable, biochemical systems. In other words, the mesh of formal expressions for the tens of inputs, states, and outputs covered by \( [f: h: x: q] \) in Eq. (1) is by no means sparse or “diagonal” in its entries. Treating all the data as a whole is likely to be important and inevitable. Yet it is highly improbable that employing any (prior) multi-variable model of the form of Eq. (1) to do so will be successful, without first reducing this whole into a collection of more tractable parts, all the inadequacies of this notwithstanding.

Let us begin, therefore, with some simpler, elementary transformations, along the lines set out above. Figure 2(a) shows an 18-day segment of the data (labeled \( y_2 \)) already given in Figure 1(c) (for the period 12–30 March, 1998), together with its smoothed estimate \( (\hat{y}_2) \) and (lower-frequency) trend \( (y^T_2) \). Extracted (higher-frequency) seasonal components, \( u^S \), \( y^S_1 \) and \( y^S_2 \), for the three signals (in Figure 1) of ammonium-N in the crude sewage, first pocket, and third pocket, respectively, are given in Figure 2(b) for the period of 18–24 March. For orientation here, the symbol \( u \) has been used to denote an input variable, while \( y_2 \) refers to an output variable; for \( y_1 \) the notation is ambivalent, since this quantity can be regarded as an output from the perspective of its relationship with \( u \) or as an input when causally related to \( y_2 \). The extracted components, i.e., the trend and seasonal components, have been generated using the microCAPTAIN software, wherein all the parameters \( (\gamma) \) employed to characterize these components (in Eq. (2)) are considered to be time-varying random walk processes, collectively represented within a Generalized Random Walk (GRW) model, itself embedded within a combined filtering–smoothing algorithm (Young, 1998). In short, the parameters are assumed to vary with time and are estimated recursively, i.e., in a manner designed to work naturally in real time; strictly speaking, these smoothed recursive estimates are expressed, therefore, as \( \hat{y}(t_k | t_N) \). The smoothed estimates of the input and output variables would bear the same argument and, for the results of Figure 2(a), \( \hat{y}_2 \) is simply the sum of its two components, \( y^T_2 \) and \( y^S_2 \).

Under this first transformation of the raw data \([u, y] \), what insights might we draw from Figure 2? Looking at Figure 2(a), the smoothed estimates of the signal \( (\hat{y}_2) \) are clearly

**Figure 1** Biological wastewater treatment case study (part). Observed time series for ammonium-N concentration during 1 February–4 April, 1998 (sampling interval 15 minutes): (a) in the crude sewage; (b) in the first pocket (outer ring) of the bioreactor; (c) in the third pocket (inner ring).
capable of reconstructing the values of ammonium-N in the effluent from the biological reactor (its third pocket) free of outliers and over periods of missing observations. This technical success is far from insignificant, especially when applied to the inputs ($u$), since any computational realization of the model of Eq. (1) (as in Liu and Beck (2000) or Liu (2000)) will require complete enumeration of these quantities over all the sampling intervals, including coverage of the inevitable gaps. The trend ($y_{T}$) shows roughly stable performance, at a baseline of 7–8 g m$^{-3}$, with a transient upward excursion over $t_{75}$ to $t_{125}$ suggestive of a temporary loss of nitrification in the bioreactor. Figure 2(b), we might speculate, illuminates something of the effective mixing volumes for each of the segments of the bioreactor and something else about the effective degradation rate of the ammonium-N brought into the system with the crude sewage. To see this, consider the following. Growth of a nitrifying population is a relative slow process, possibly revealed through the evidence of the (lower-frequency) trend components of the signals (not shown). Actual removal (and degradation) of the ammonium-N is a function of contact between the biomass and this substrate, which arguably is dominated by transport phenomena predominant over the shorter term and revealed through fluctuations in the (higher-frequency) seasonal, i.e., diurnal, components of the signals (Figure 2(b)). The amplitude ratios and relative phase lags of the components, which notably vary with time, might be obtained from juxtaposing these extracted components – as opposed to the raw data – as input–output transfer functions, hence providing results in a “transformed space” and thus provoking insights not immediately obvious from $[u, y]$.

Figure 3 presents the results of just such a transfer-function model, based on Eq. (2b), in which the identification of $g(u, y; t)$ is now brought into play, in a rather more sophisticated transformation. Conceptually, what is seen here is this. The best form of transfer function relating the variations in ammonium-N concentration passing from the first ($y_{1}$) to the third ($y_{2}$) pocket of the bioreactor turns out to be equivalent to an active-mixing volume model (Young and Lees, 1993), i.e., a combination of a continuously stirred tank reactor (CSTR) and plug flow reactor (PFR) element. Figure 3 shows the time-varying estimates of the
time-constant ($\tau^\hat{}$) and steady-state gain ($\kappa^\hat{}$) of the CSTR element; in addition, Figure 3(a) provides the accompanying one-standard deviation error bounds attaching to $\tau^\hat{}$, while Figure 3(b) includes the trend component extracted from the observed signal for the mixed liquor suspended solids (MLSS) concentration in the bioreactor over this period. This reconstruction of the “inner” properties of $[\tau^\hat{}, \kappa^\hat{}]$ is now more than one stage of transformation removed from the “outer” properties of $[u, y]$. Given a knowledge of the observed variations in the flow of the mixed liquors passing through the bioreactor (not shown) it is theoretically possible to pass from $[\tau^\hat{}, \kappa^\hat{}]$ into the quasi-conceptual (mechanistic) space of $[\hat{v}, \hat{r}]$, where $\hat{v}$ represents an estimate of the (time-varying) effective mixing volume of a portion of the bioreactor and $\hat{r}$ the corresponding time-varying estimate of the approximate first-order degradation-rate constant for the removal of ammonium-N. Since we are presuming this “removal” to be a function of microbial activity, as opposed to pure chemistry, $[\hat{v}, \hat{r}]$ are not quite the same as our innermost space of parameters, represented as $\alpha$ in Eq. (1). Apparently, from Figure 3(a), the dynamics of the system are accelerating over this 18-day period, while the overall removal of ammonium (in Figure 3(b)) exhibits a weekly oscillation – as does the extracted trend component of the MLSS concentration, inversely and albeit somewhat out of phase. This latter insight requires its own translation. The estimates of $\kappa$ in Figure 3(b) are at times greater than 1.0, suggesting there is actually more ammonium-N present in the output than the input, whereas at other times, e.g., at about $t_{125}$, they are roughly 0.8, i.e., indicative of a 20% reduction in ammonium-N concentration between the first and third pockets of the reactor. Of further note, over the period $t_{75}$ to $t_{105}$ no ammonium-N is removed from the system. Overall, one could infer (from Figure 3(b)) that removal declines (rises) when the MLSS declines (rises) in sympathy with the degree of contact between the biomass and the substrate.

These results, in particular those suggestive of some correlation (or causation) between the trend component of the MLSS concentration and the steady-state gain of the ammonium-N transfer function model, can be transformed yet again. In essence, the temporal fluctuations in a parameter ($\gamma$) estimate can be made a formal function of the variations in
one of the system’s state \( (x) \) variables, using expressions in which appear (ideally) only other parameters \( (\alpha) \) “truly invariant” with time (see, for example, Beck and Young (1976)). In the specific setting of the software used here (microCAPTAIN) implementing this further transformation is referred to as state-dependent parameter (SDP) modeling (Young, 1998). Exactly how to proceed from the results of the transformation (of an SDP model), cast in the space of the LOM \( (g\{u,y,t_k\}) \) in Eq. 2(b), to the formal state-space HOM of Eq. (1) remains unclear at present, all the prior theory on the mechanisms of the nitrifying behavior of an activated sludge biomass notwithstanding. It is the subject of ongoing research.

To take stock of our current position in the chain of transformations, let us summarize in cryptic form whence we have come (within this case study) and whither we are headed. Thus we began with the raw data \([u,y]\) (Figure 1); next came smoothing and interpolation of these data to give \([\hat{u},\hat{y}]\) (in Figure 2(a)); then extraction of the component signals \([u^T,\hat{u}^T,y^T,y^S]\) (primarily Figure 2(b)); followed by the identification of an input–output transfer function model, with generically labeled parameters \( (\gamma) \), specifically illustrated as \([t,k]\) in Figure 3; from which it would be possible to derive a set of parameters \( (\alpha) \) with a “mechanistic” interpretation, specifically \([\hat{v},\hat{r}]\); hence, from these beginnings of a state-space model, \([x,\alpha]\), we could in principle proceed to the results of Figure 4. Indeed, summed in this last set of results are the outcomes of two further steps (or transformations) in a process of matching a HOM – eventually of the order of 100 or more state variables and 100 or so parameters in \([f,h,x,\alpha]\) – to the same 18-day period as above for the entire set of observed variables (except for two sequences of on-line respirometric observations; Liu and Beck (2000), Liu (2000)). Step one of the process demonstrated the inadequacy of treating nitrification as though it were mediated by just a single biomass of nitrifiers, as the most elementary, prior, conceptual (or mechanistic), model (the observations in Figure 4 are of ammonium-N in the first and third pockets of the bioreactor). Step two follows from a further composite re-structuring of the HOM, which was entirely driven by the need to mimic the loss of nitrification – apparent in Figure 2(a) and, again, in Figure 3(b) – and arguably the result of a minor precipitation-induced surge of combined sewage over the period \( t_{75} \) to \( t_{105} \). The (posterior) high-level conceptual description of the restructuring of the model runs as follows: flow of mixed liquors through the bioreactor is vertically stratified; in fact, the liquors may not be all that well “mixed”, because the highly viscous, relatively dense, recycled biomass is introduced into the bioreactor from a bridge, whereafter it probably behaves initially as a submerged waterfall, descending to the bottom of the

![Figure 4](https://iwaponline.com/wst/article-pdf/47/2/43/423913/43.pdf)

**Figure 4** State-space modeling of ammonium-N behavior in wastewater treatment system during 12–30 March, 1998 (sampling interval 2 hours): (a) first pocket of bioreactor; (b) third pocket of bioreactor (from Liu, 2000)
tank; the biomass is treated conceptually as particulate matter in the bioreactor\(^2\); when a diluted, less dense, crude sewage enters the reactor, still less contact than normal is made in the stratified flows between the lower (dense) nitrifying biomass and its upper (less dense) substrate.

At each of the steps along the path of transforming the raw data, across Figures 1–4, insights have been obtained; Figure 4 alone encompasses substantial evolution in the structure \([f,h]\) of a formal model (Eq. (1)), from a HOM to something approximating a VHOM. These insights, however, have been ones of understanding, as opposed to insights leading directly to decisions and actions.

Conclusions
The salient and abiding experience of the past three decades has been that of having to work with mathematical analyses of sets of data on the behavior of aquatic environments that have been sampled too infrequently in time: from control of biological wastewater treatment systems, to surface water acidification, to watershed nutrient loading analysis, to lake eutrophication (Beck, 1987). This need no longer be the case. In fact, many of the transformations presented herein have been conducted on the basis of aggregating, even discarding, observations at their highest frequency of sampling. Change of this order is dramatic: from too few to too many data. It solves some of our problems, yet reveals others of an unexpected nature. Given data of the volume and quality of Figure 1, accepting a high-order model as “valid” – because the data are too sparse to deny this judgement – becomes untenable. Mismatches between the model and data are obvious and of a non-random origin, capable of rectification, but at very great effort (see, for example, Liu, 2000). For the case study of this paper, the predominant challenge has been its multi-variable character, as opposed to interpretation simply of the higher-frequency fluctuations to which access has been enabled through an automated monitoring platform (in our case, the Environmental Process Control Laboratory). To extract information in these situations cannot presently be reduced to entirely automated procedures operating at the same speed as the monitoring system (and perhaps such automated interpretation is itself a risky prospect). Furthermore, the insights acquired from real-time data do not have to be directed at immediate, real-time goals. In the case study of wastewater treatment the most significant insights liberated by the transformations of the high-frequency data relate to re-design of the plant – by installing bed-mounted mixers for regulating the settling and resuspension of biomass in the bioreactor – and the deferring of capacity expansion through better operation. In other words, high-frequency data, when properly interpreted, can guide lower-frequency decisions (we re-design and re-build infrastructure infrequently). Elsewhere, we have found the converse to be just as relevant: high-frequency control actions (varying within the day) can be successfully applied day after day in order to avoid a lower-frequency convergence of conditions likely to promote (high-frequency) process failure from a bulking sludge (Chen and Beck, 1993). In other words, a vital attribute of automated monitoring is persistence in its maintenance over the longer term, thereby illuminating significant cross-spectrum connections between what governs behavior in the short-term and what in the long-term.

Acknowledgement
M.B. Beck is currently Visiting Professor in the Department of Civil and Environmental Engineering at the Imperial College of Science, Technology and Medicine, London.

---

\(^2\) This seems unsurprising, but the prior HOM, based on the standard model of Henze \textit{et al.} (2000), treats the biomass in the reactor (as opposed to the clarifier) as a solute, not subject therefore to either settling or resuspension (Liu, 2000).
References


