Preprints of the
9th International Symposium on Advanced Control of Chemical Processes
The International Federation of Automatic Control
June 7-10, 2015, Whistler, British Columbia, Canada

Analysis of problems induced by imprecise dating of measurements in oil and gas production

Nicolas Petit *

* *(e-mail: nicolas.petit@ mines-paristech.fr).*

Abstract: In this paper we discuss the negative impact on monitoring algorithms of working with imprecisely dated data. Two examples from the world of the oil & gas industry are presented and serve to illustrate that this problem can be of practical importance. First analytical results show that when signals with significant time variations are monitored, the impact of dating of measurements can be as troublesome (or even worse) than measurement noises.

Keywords: process control, monitoring algorithms, synchronisation of data, information technology, distributed database

1. INTRODUCTION

Industrial information technology (IT) have steadily grown and become ubiquitous and powerful over the last decades. Due to this impressive push, the question of data-based monitoring of processes could be expected to have become easier. Indeed, IT has enabled the availability of massive streams of data serving in sophisticated data analysis, which has resulted in the development of advanced algorithms. However, the situation is not that straightforward. In this article, we wish to point out a serious and intrinsic limitation of IT that has been, so far, underestimated: erroneous dating of data. The focus of this article is put on the oil & gas industry, but similar conclusions could be drawn in numerous other fields.

Concerns about dating of measurements is not a new. There is little doubt in the mind of production and process engineers that mis-synchronisation of measurements has some impact on all the algorithms they use to exploit data. In fact, because of this belief, numerous solutions are usually implemented to mitigate this problem. The most widely used solution is synchronisation of clocks across the IT network. Synchronisation procedures are employed to have a single time-reference shared by the various subsystems. Unfortunately, these procedures are built on assumptions that are impossible to guarantee, strictly speaking. For this reason, synchronisation can not be achieved with an arbitrary accuracy (Noble (2012)). For example, in the state-of-the-art NTP synchronization algorithm, synchronization is only correct when both the incoming and outgoing routes between the client (the computer to be synchronized) and the server (considered as reference) have symmetrical nominal delay. Otherwise, the synchronization has a systematic bias of $\Delta/2$ where Δ is the difference between the forward and backward travel times. In practical cases, this synchronicity is extremely rare, not to say impossible: for latency in both directions of a no trivial path to be equal and stable, it is required that the bit rate of all links in the path are identical, the traffic flow in both directions, the IP routing in both directions is identical, the routing policies at every device in the path in both directions is identical, and the host systems at both ends are behaving identically. This situation is unrealistic in any real-world industrial application.

In the general context of industrial process control (Luyben et al. (1998)), the devices needing synchronisation are in a very large number: e.g. for a refinery, at least hundreds of computers and tens of thousands of sensors are under consideration. The nature and the quality of the network employed has also a great importance. For example, in the quickly evolving “digital oilfield” applications, networks are composed of various types of connections (Ethernet, Wireless, Fiber, VSTA) with great variability in their bandwidth and latency. These facts make synchronisation of such networks a serious problem, not to say an almost impossible task, given the performance of currently employed technologies.

For these reasons, one observes relatively large mis-synchronisation of data in plant-wide applications. A natural question is to determine whether mis-synchronisation is large enough to cause any problem. More precisely and quantitatively, one can formulate the following: *What is the cost of working with imprecisely dated measurements?*

To try to answer this question, we consider two simple applications, putting into play basic monitoring applications. As will be shown, the mis-synchronisation is the root cause of mis-interpretation in the monitoring tasks.

The paper is organized as follows. In Section 2, a general description of current IT systems and recent “Big Data” technologies is proposed. Two examples are treated in Section 3 and Section 4, respectively. The first problem under consideration, which receives analytical investigations, is a material flow monitoring application, the second one is the calculation of allocation factor in a oil-field. Section 5 contains conclusions.

2. GENERAL FACTS FOR INDUSTRIAL DATA MANAGEMENT SYSTEMS

2.1 Domains and current architectures as sources of mis-synchronisation

Before developing the examples under consideration, it is necessary to give a brief overview of current technologies in industrial data management systems. In most current implementations, the typical constitutive elements of an IT system are historian databases, production databases, and application servers. At a company level, the IT system is usually decomposed into two areas: the *Process domain* and the *Office domain.*

In the process domain, one finds the sensors, SCADAs (supervisory control and data acquisition), PLCs (Programmable logic controller), RTUs (Remote terminal units). Data is collected and stored in Historian databases which are part of the process-domain. These systems provide convenient support for the tasks of real-time on-site operations surveillance.

In the office domain, one finds Production database management systems (PDMS) with technologies based on SQL (Beaulieu (2005)). Databases contain various documents: production events, networks descriptions, daily reports. To grant data access, a replica of the process domain historian database is used.

In large-scale production systems, the network is geographically distributed from production sites to offices. This is particularly true for the oil & gas industry. Numerous parts of the networks (connection lines or subnetworks) may have very limited bandwidth, and communication outages can happen at any connection in the network.

At a company-wide scale, the amount of data necessary to perform real-time monitoring and analysis exceeds data volumes that existing architectures can deal with easily. The reasons for this are that *i)* the SCADAs are designed to provide security of operations and are not scalable, *ii)* the production network is bandwidth limited and unreliable (e.g. VSAT), *iii)* accessing a central database for daily applications yields high latency, *iv)* current PDMS can not do large-scale data replication across sites, *v)* SQL limit data availability because of locks issues in case of network partition.

Because of the technology gaps and heavy and slow interconnections between process domain and office domain IT technologies, mis-synchronisation of data occurs.

2.2 Big data technologies for geographically distributed production IT system

Currently, office domain data storage is organized in cluster operating hardware and software chosen for providing maximum robustness. However, this is not the only aspect that should be paid attention to.

Outside the industrial world (e.g. in e-business), modern IT systems are chosen for their ability to scale rapidly and grant easy and fast access to data, at a company-wide level. At this scale, data should be read and written from multiple replicated databases at once. The whole system must be robust to hardware failures. This is all the most important as the probability of individual failure is high when the number of elements increases. The IT has to deal with seek time improving more slowly than transfer rate and inhomogeneous bandwidth distribution. These are two important sources of latency if not treated appropriately. For scalability, an IT system should be chosen to guarantee that if one doubles the amount of data to be treated and doubles the size of the hardware handling this data (clusters), then accessing (and processing) the data should run just as fast (not slower), and without any increased probability of failure.

The preceding issues have received numerous answers from the world of Computer Science "Big Data" where the amount of data grows quickly. There are several properties that need to be enforced by the IT architecture to grant fault tolerance and scalability. Interestingly, these are automatically obtained if one considers specific programming languages. Among these languages are Erlang (Cesarini and Thompson (2009)), Hadoop (White (2012)). Synchronization of databases, is, at the considered company-wide scale, a problem of many-to-many talking. For this problem, standard (currently employed) software technologies are not well suited. Rather, one should consider the concept of (geographically) distributed databases, which are state-of-the-art in the Internet: Google's BigTable (Chang et al. (2008)), Amazon's Dynamo (DeCandia et al. (2007)) among others. These use redundant storage on clusters of commodity hardware and are instrumental in the scaling to petabytes-large datasets across locations. "Single point of failure" issues are avoided. Write can be done on any node thanks to smart conflict resolution. In these technology, a central concept is Message Oriented Middleware (MOM) which facilitates integration of heterogeneous and decoupled systems. It is based on asynchronous message passing to orchestrate data access and exchanges between application sharing common data sources. Message passing across applications and publish/subscribe handles transformation of data formats, and decouples systems in terms of space, time and synchronicity. It is a highly scalable many-to-many communication system from senders to receivers. In the industrial context of this article, publishers would be sensors, algorithms, applications outputs, while subscribers would be algorithms, applications.

Several studies focused on the oil & gas industry have reported that the "Big Data" technology is capable of handling data of dozens of fields, each with hundreds of wells. This permits to draw similar conclusions for large chemical plants and refineries.

However, they are not used yet in industry. For this reason, one currently has to deal with mis-synchronisation of data.

3. BULK FLOW MONITORING

To illustrate why mis-synchronisation is a true problem, let us now consider a first application, introduced in Magnis and Petit (2013). Consider a pipeline that is monitored to detect leaks (which can be also thefts (Dudek (2005); API (1995)). The situation is pictured in Figure 2. To achieve the monitoring (see Figure 3), the inlet an outlet of the transport pipe are equipped with flow meters. These devices sample the flow variable and communicate the

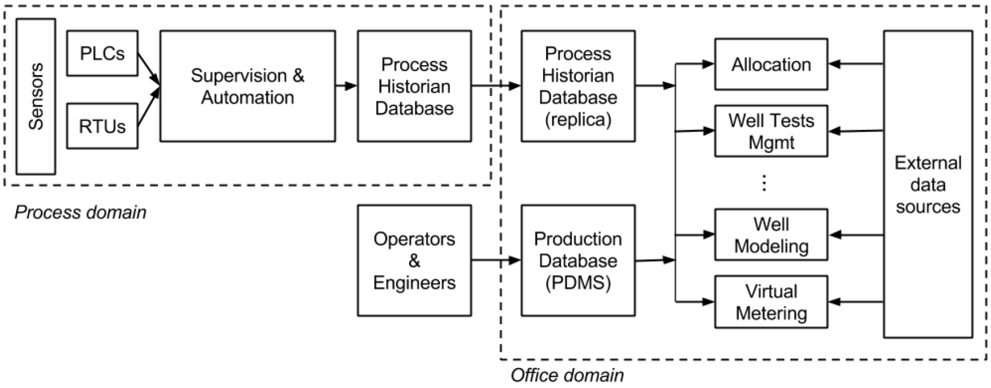

Fig. 1. A typical production IT system.

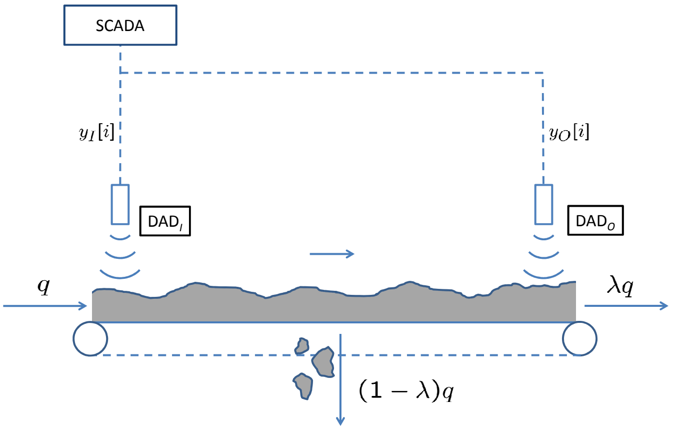

Fig. 3. Bulk-flow.

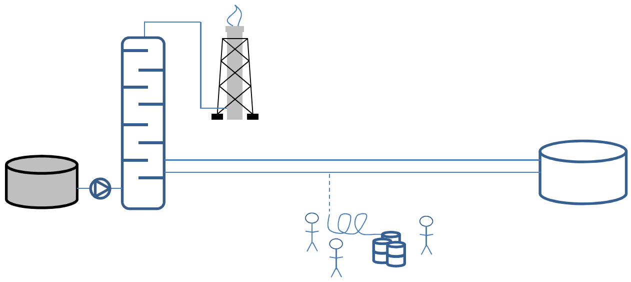

Fig. 2. Detection of leaks on a pipeline.

data to a centralized system (SCADA) which applies a time stamping at the reception. Due to communication delays which vary over time (and may jump by large amounts), the data that is available at the SCADA level is not consistent, because the flow measurements are not synchronous, although they may appear as such in the historian database, if they are received simultaneously. One could imagine that the solution would be to timestamp the data at emission (i.e. at the DAD level). However, there is no guarantee that the internal clocks of the DAD are synchronized. In practice, clocks are not synchronised. We refer the interested reader to the introductory discussion on the flaws of NTP synchronisation algorithm and the weakness of the signals received from atomic clocks (which could appear as an attractive solution when its induced cost does not discard it from applications).

3.1 Model and state-of-the art monitoring system

Mathematically, the usual technique employed for detecting the leaks consists in formulating a mass balance equation (Begovich et al. (2007); Fraden (2010); Geiger (2006)). Note q the inlet flow rate, λq the outlet flow rate, with $\lambda \in (0,1]$ which is unknown. The loss factor is $1-\lambda$ ($\lambda = 1$ corresponds to a no loss situation). The two DADs produce samples of the flow rates (Δt is the sample time), which are corrupted with noises.

$$y_I[i] = q(i\Delta t) + n_i, \qquad y_O[i] = \lambda q(i\Delta t) + n_i'$$

With this modelling, an Imbalance estimator over a time-window $[0,T]$ is

$$\hat{b} = \Delta t \frac{\sum_i y_I[i] - y_O[i]}{\int_0^T q(t)dt}$$

Due to noises, we have:

$$\hat{b} = 1 - \lambda + \text{noise} + \underbrace{\text{numerical integration error}}_{=\ 0 \text{ by assumption}}$$

Then a simple detection algorithm is to define the alarm as follows

$$\hat{b} \geq b^*\text{: loss-alarm}$$
$$\hat{b} < b^*\text{: no loss-alarm}$$

The threshold b^* is the only parameter of this algorithm.

3.2 A priori bound without dating uncertainty

From information theory, it is possible to derive a lower bound on the variance of estimators of the deterministic parameter λ. For this, we consider a particular stochastic setup, but generalizations are possible.

Classically, the Cramér-Rao bound expresses a lower-bound on the variance of any unbiased estimator of λ obtained from a given set of measurements X distributed according to a probability density function $f(X,\lambda)$ (see Frieden (2004)). This bound is

$$\mathrm{Var}\hat{\lambda} \geq \frac{1}{I(\lambda)}$$

where I is the Fischer information defined as

$$I(\lambda) = -\mathrm{E}[\frac{\partial^2 \ell(X,\lambda)}{\partial \lambda^2}], \quad \ell(X,\lambda) = \log f(X,\lambda)$$

Here, the measurements X are the available samples $y_I[]$ and $y_0[]$. For a given flow rate q and centered Gaussian noises n and n', one can express the probability density function f as

$$f(X,\lambda) = C\prod_i G(y_0[i] = X_i, \lambda q(i\Delta t), \sigma')$$

where $G(X, m, \sigma)$ designates a Gaussian distribution of average m and variance σ^2, and C factors all the terms that do not depend on λ. The variance of the discrete processes n (and n') can be obtained from the power spectral density (PSD) R (and R') as

$$\sigma^2 = R/(\Delta t), \quad (\sigma')^2 = R'/(\Delta t)$$

This gives

$$\ell = \log(f(X,\lambda)) = \log C + \sum_i \frac{(\lambda q(i\Delta t) - X_i)^2}{2(\sigma')^2}$$

where c does not depend on λ. Then

$$-\frac{\partial^2 \ell}{\partial\lambda^2} = \sum_i \frac{q(i\Delta t)^2}{(\sigma')^2}$$

which directly give $I(\lambda)$. So, assuming a large number of samples are available and, for ease of reading, considering the limit case $\Delta t \to 0$, one deduces the handy formula

$$\mathrm{Var}(\hat{\lambda}) \geq \frac{R'}{\|q\|_2^2} \tag{1}$$

where $\|q\|_2^2 = \int_0^T q(\tau)^2 d\tau$, for $[0,T]$ covering all the sample times.

3.3 Introducing dating uncertainty

Now, to account for dating uncertainty, we introduce noise in the sampling instants. Straightforwardly, we consider

$$y_I[i] = q(i\Delta t) + n_i, \qquad y_O[i] = \lambda q(i\Delta t + w_i) + n'_i$$

with w_i is a random value.

To quantitatively study the impact of the uncertainty, we introduce the following stochastic modeling (see Magnis and Petit (2013) for extensions and numerical studies). It is considered that λ is an unknown, the dating uncertainties w_i are independent identically distributed centered Gaussian variables, and the overall noise is Gaussian and centered. In principles, this allows us to perform explicit computations for the probability law of accurately detecting losses and generating false alarms.

The uncertainties creates ambiguity which grows with the variance of w.

3.4 A priori bound with dating uncertainty

Note $(\sigma_w)^2 = R_w/\Delta t$ the variance of the centered Gaussian noise w. Developing

$$y_O[i] = \lambda q(i\Delta t) + \lambda w_i \dot{q}(i\Delta t) + n'_i$$

Note

$$s_i(\lambda) = \sqrt{\sigma_i'^2 + \lambda^2\sigma_w^2\dot{q}(i\Delta t)^2}, \quad p_i(\lambda) = s_i(\lambda)^2$$

The density function of the measurement vector X is

$$f(X,\lambda) = C\prod_i G(y_0[i] = X_i, \lambda q(i\Delta t), s(\lambda))$$

where, again, C gathers all the terms that do not depend on λ. Here, $\ell = \log(f(X,\lambda))$ has the form

$$\ell = \log C + \sum_i -\frac{1}{2}\log p_i(\lambda) - \frac{(\lambda q(i\Delta t) - X_i)^2}{2p_i(\lambda)}$$

Thus,

$$\begin{aligned}\frac{\partial\ell}{\partial\lambda} = \sum_i &- \frac{\lambda\sigma_w^2\dot{q}(i\Delta t)^2}{p_i(\lambda)} + \frac{(\lambda q(i\Delta t) - X_i)^2\lambda\sigma_w^2\dot{q}(i\Delta t)^2}{p_i(\lambda)^2}\\ &- \frac{q(i\Delta t)(\lambda q(i\Delta t) - X_i)}{p_i(\lambda)}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2\ell}{\partial\lambda^2} = \sum_i &- \frac{\sigma_w^2\dot{q}(i\Delta t)^2}{p_i(\lambda)} + \frac{2\lambda^2\sigma_w^4\dot{q}(i\Delta t)^4}{p_i(\lambda)^2}\\ &+ \frac{2q(i\Delta t)(\lambda q(i\Delta t) - X_i)\lambda\sigma_w^2\dot{q}(i\Delta t)^2}{p_i(\lambda)^2}\\ &+ \frac{(\lambda q(i\Delta t) - X_i)^2\sigma_w^2\dot{q}(i\Delta t)^2}{p_i(\lambda)^2}\\ &- \frac{4(\lambda q(i\Delta t) - X_i)^2\lambda^2\sigma_w^4\dot{q}(i\Delta t)^4}{p_i(\lambda)^3}\\ &- \frac{q(i\Delta t)^2}{p_i(\lambda)}\\ &+ \frac{2\lambda\sigma_w^2\dot{q}(i\Delta t)^2 q(i\Delta t)(\lambda q(i\Delta t) - X_i)}{p_i(\lambda)^2}\end{aligned}$$

Then,

$$\begin{aligned}-\mathrm{E}[\frac{\partial^2\ell}{\partial\lambda^2}] = \sum_i &\frac{\sigma_w^2\dot{q}(i\Delta t)^2}{p_i(\lambda)} - \frac{2\lambda^2\sigma_w^4\dot{q}(i\Delta t)^4}{p_i(\lambda)^2}\\ &- \frac{\sigma_w^2\dot{q}(i\Delta t)^2}{p_i(\lambda)} + \frac{4\lambda^2\sigma_w^4\dot{q}(i\Delta t)^4}{p_i(\lambda)^2}\\ &+ \frac{q(i\Delta t)^2}{p_i(\lambda)}\end{aligned}$$

After simplification,

$$I(\lambda) = -\mathrm{E}[\frac{\partial^2\ell}{\partial\lambda^2}] = \sum_i \frac{2\lambda^2\sigma_w^4\dot{q}(i\Delta t)^4}{p_i(\lambda)^2} + \frac{q(i\Delta t)^2}{p_i(\lambda)}$$

For which we derive the following Cramér-Rao inequality

$$\mathrm{Var}(\hat{\lambda}) \geq \frac{1}{I(\lambda)}$$

For small values of σ_w the dominant term in the expansion of this expression (in powers of σ_w) is, using a convenient limit $\Delta t \to 0$

$$\mathrm{Var}(\hat{\lambda}) \geq \frac{R'}{\|q\|_2^2} + R_w\lambda^2\frac{\|q\dot{q}\|_2^2}{\|q\|_2^4} \tag{2}$$

where $\|q\|_2^2 = \int_0^T q(\tau)^2 d\tau$, $\|q\dot{q}\|_2^2 = \int_0^T q(\tau)^2\dot{q}(\tau)^2 d\tau$, for $[0,T]$ covering all the sample times. Interestingly, (2) can be compared to (1), stressing the additional error due to dating uncertainty.

3.5 Application example

To avoid having too many data interlacing, i.e. data arriving in the wrong order, it is natural to impose that $\sigma_w < \Delta t/\gamma$, where $\gamma > 1$. Then, this gives $R_w < \Delta t^3/\gamma^2$. We now consider a periodic signal $q(t) = 1 + \frac{1}{2}\sin(2\pi Nt)$ where N is a given frequency. We wish to determine a critical frequency above which the error due to noise is overwhelmed by the error due to the mis-synchronisation as computed in (2).

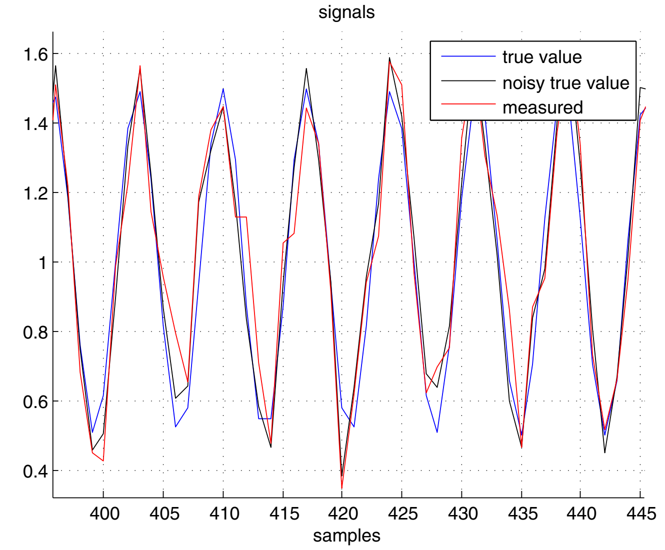

Fig. 5. Signals N=14 for which dating uncertainty is causing as much error on the loss detection as measurement noise.

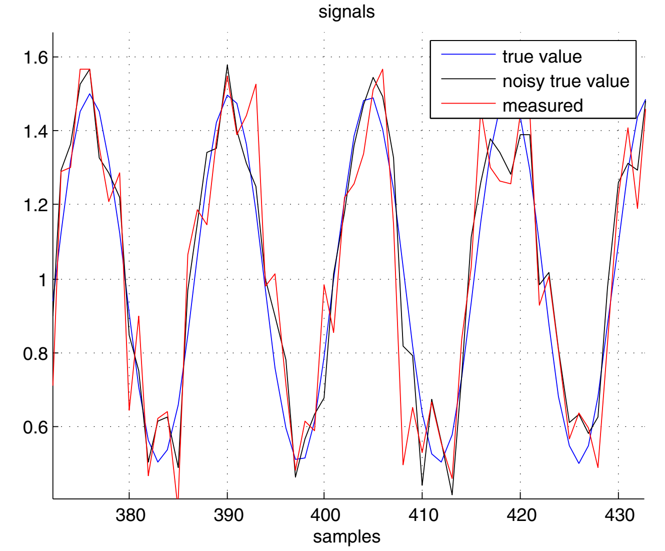

Fig. 4. Signals N=7 for which dating uncertainty is causing as much error on the loss detection as measurement noise.

For a time horizon $[0, T]$, one has

$$\dot{q}(t) = \pi N \cos(2\pi N t)$$

which shows that $\|q\dot{q}\|_2^2$ is an increasing function of N.

When the variance of w is increased, the contribution of dating uncertainty to the variance in the Cramér-Rao bound is dominant. This is clear from (2). This can be observed in Table 1 and Table 2. Typically, when approx. 10% of the dating uncertainty is larger than Δt ($= 0.01$ in this simulation) in absolute value, this is the case for $N = 7$. When the frequency of the signal is doubled to $N = 14$, this percentage drops to the very demanding value of 0.2%. Signals are reported in Figure 4 and Figure 5.

Further for large values of N, i.e. when the measured signal is high frequency, the variance due to dating uncertainty gets large. Eventually, it gets larger than the variance due to noise. This can be observed by comparing the value in Table 1 and Table 2 respectively.

4. ALLOCATION FACTOR: DATA VALIDATION AND RECONCILIATION (DVR)

The second application that we wish to discuss stresses the difficulties related to geographically distributed IT systems (see Préveral et al. (2014) for more details).

4.1 Data Validation and Reconciliation (DVR)

A daily task that production engineers must perform in the oil industry is Data Validation and Reconciliation (DVR). After having gathered production data consisting in redundant real-time measurements of flow rates (and possibly pressures, temperatures) from sensors placed in various locations in the production networks, the task consists in producing best estimates of the production of each well, by an analysis of the data and their comparison at the light of mass balance equations. Determining each well production level is important to perform hydrocarbon accounting, detect possible production network leaks, and very importantly, validate geophysics studies modeling the mid-term or long-term behavior or the producing reservoir. In theory, this task boils down to a clear linear data analysis problem, as is shown below. In practice, up to 60% of production engineers monitoring time is spent gathering data. This is due to the fact that the various IT subsystems (PDMS (well tests results, shuts-ins, production events) and historian databases (sensor data)) needed are not properly interconnected, produce data with inconsistent formats, and are not synchronized (and sometimes missing). This tedious task is performed daily.

4.2 A case study: production network

DVR solution Mathematically, the DVR problem can be simply modeled as follows (in its simples form). First, one shall partition the variables as

$$x = \begin{pmatrix} x_m \\ x_u \end{pmatrix}$$

where x_m are measured flow rates, and x_u are unmeasured flow rates. The equation relating these variables are, from conservation principles (material balance)

$$Ax = 0$$

where A is a network representation matrix given in Figure 6. Then, by separating the material balance as

$$A_m x_m + A_u x_u = 0$$

and by introducing some uncertainty to account for measurement noise

$$y = x_m + e$$

where e is zero-mean (un-correlated) Gaussian noise of standard deviation σ, the DVR can be reformulated as the following constrained optimization problem

% outside $[-\Delta t, \Delta t]$	var. due to noise	var. due to dating	Cramér-Rao bound
31.7	9.42e-06	2.15e-05	3.09e-05
21.1	9.42e-06	1.37e-05	2.32e-05
13.3	9.46e-06	9.56e-06	1.89e-05
8.0	9.42e-06	7.03e-06	1.64e-05
4.5	9.42e-06	5.38e-06	1.48e-05
2.4	9.42e-06	4.25e-06	1.36e-05
1.2	9.42e-06	3.44e-06	1.28e-05
0.6	9.42e-06	2.84e-06	1.22e-05
0.3	9.42e-06	2.39e-06	1.18e-05
0.0	9.42e-06	1.34e-06	1.07e-05

Table 1. Error Variance due to dating uncertainty can overwhelm noise (N=7)

% outside $[-\Delta t, \Delta t]$	var. due to noise	var. due to dating	Cramér-Rao bound
31.7	9.42e-06	8.61e-05	9.55e-05
21.1	9.42e-06	5.51e-05	6.45e-05
13.3	9.42e-06	3.82e-05	4.77e-05
8	9.42e-06	2.8e-05	3.75e-05
4.5	9.42e-06	2.1e-05	3.09e-05
2.4	9.42e-06	1.70e-05	2.64e-05
1.2	9.42e-06	1.37e-05	2.32e-05
0.6	9.42e-06	1.13e-05	2.08e-05
0.3	9.42e-06	9.56e-06	1.89e-05
0.04	9.42e-06	7.03e-06	1.64e-05
0.0	9.42e-06	5.38e-06	1.48e-05

Table 2. Error Variance due to dating uncertainty can overwhelm noise (N=14)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
Well 1	1	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Well 2	0	1	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Well 3	0	0	1	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0
Sep. 1	0	0	0	1	1	1	0	0	0	0	-1	0	0	0	0	0	0	0	0
Well 4	0	0	0	0	0	0	1	0	-1	0	0	0	0	0	0	0	0	0	0
Well 5	0	0	0	0	0	0	0	1	0	-1	0	0	0	0	0	0	0	0	0
...																			
Export 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1

Fig. 6. Network matrix representation.

$$\begin{aligned} \min_{x} \quad & \sum_i \left(\frac{y_i - x_i}{\sigma_i} \right)^2 \\ \text{s.t. } & Ax = 0 \end{aligned} \tag{3}$$

In the so-called *observable case*, the solution to the DVR (3) produces, using the QR decomposition

$$A_u = QR = (\, Q_1 \;\, Q_2 \,) \begin{pmatrix} R \\ 0 \end{pmatrix} \Pi$$

and noting $S = Q_2^T A_m$

$$\Sigma = \text{diag}(\sigma)$$

then

$$\hat{x}_m = y - \Sigma S^T (S \Sigma S^T)^{-1} S y$$
$$\hat{x}_u = -\Pi^{-1} R^{-1} Q_1^T A_m \hat{x}_m$$

which are unbiased normally distributed estimates. Above, $\hat{x}_m$ is called the reconciliation of measured variables, and $\hat{x}_u$ is the coaptation for unmeasured variables.

4.3 Analysis of the results

Considering the network pictured in Figure 7, simulation data have been treated in a DVR scenario. The sensors have various level of uncertainty. Well 5 has a Multiphase Flow Meter. Two export stations are measured accurately. Other sensors are deployed, providing mid-to-large uncertainty measurements.

Several mis-synchronisation has been introduced in the data. Typical synchronisation errors due to heterogeneous databases are introduced. Some flow rates are down-sampled and averaged to daily values, which also introduces some lag. Delays are present on flow rate measurements.

Individual well flow rate estimates are reproduced in Figure 8 Averaged values are consistent with true means, but poorly-synchronized estimates fail to warn of well 4 flow rate reduction (at time=180). Further a false warning at time=400 as production is smooth, while DVR detects a non-existing anomaly.

On the other hand, the key production indicator which is the allocation factor (ratio of total well flow estimate to total measured bulk/fiscal flow) is wrongly computed. The histories reproduced in Figure 9 stress this fact. Instead of the true value which kindly oscillates about 1, spurious oscillations appear in the reconstructed value. Reality is much smoother than DVR estimates suggest. Spurious oscillations are due to mis-synchronisation of data produced by the IT system.

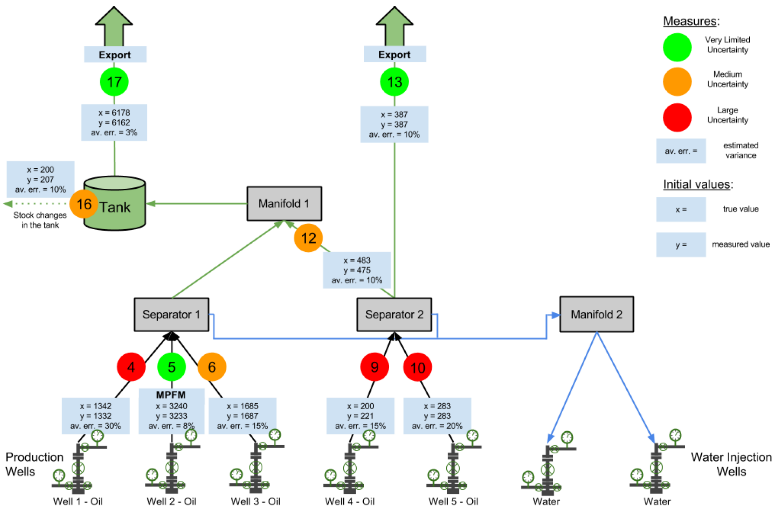

Fig. 7. A production network.

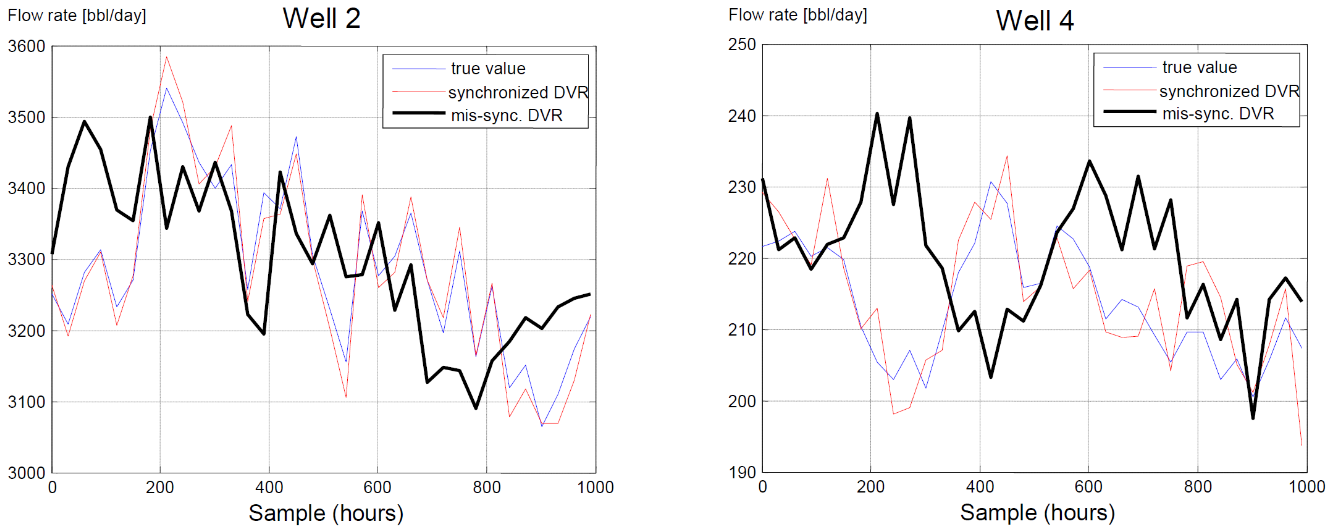

Fig. 8. Individual well flow rate estimates.

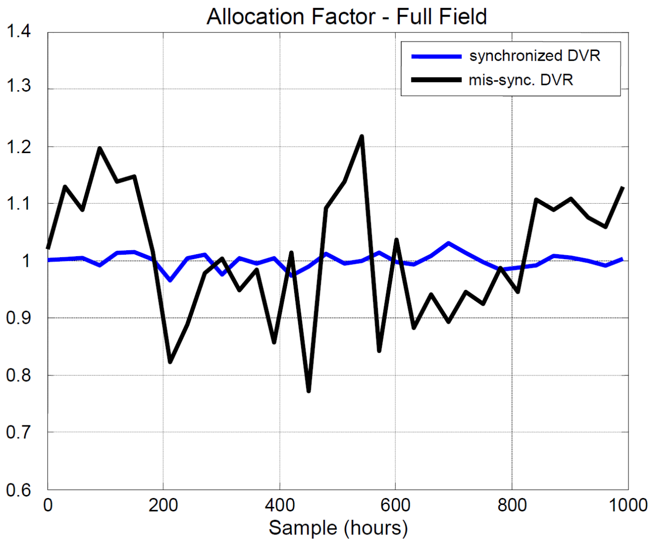

Fig. 9. Allocation factor.

5. CONCLUSIONS

Based on two simple but representative problems, this article wish to stress the negative impact of measurements dating errors on monitoring algorithms. In fact, mis-synchronisation can be more important than noises, which is not a well-known fact. The analytical study conducted on the bulk-flow monitoring problem proves this statement. It could be generalized to other cases of applications (including the second example presented here), and to the vast question of data assimilation.

Acknowledgments This work was partially funded by Institut Carnot M.I.N.E.S. The author is indebted to Lionel Magnis, Aymeric Préveral-Etcheverry and Antoine Trihoreau.

REFERENCES

API (1995). *1155: Evaluation Methodology for Software Based Leak Detection Systems.* American Petroleum Institute.

Beaulieu, A. (2005). *Learning SQL.* O'Reilly Media.

Begovich, O., Navarro, A., Sánchez, E.N., and Besançon, G. (2007). Comparison of two detection algorithms for pipeline leaks. 16^{th} *IEEE International Conference on Control Applications*, 777–782.

Cesarini, F. and Thompson, S. (2009). *Erlang programming, A concurrent approach to software development.* O'Reilly Media.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W.C., Wallach, D.A., Burrows, M., Chandra, T., Fikes, A., and Gruber, R.E. (2008). Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2), 4.

DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., and Vogels, W. (2007). Dynamo: Amazon's highly available key-value store. In *in Proc. SOSP*, 205–220.

Dudek, M. (2005). Liquid leak detection focused on theft protection. In *Proceedings of Pipeline Simulation Interest Group PSIG annual meeting.*

Fraden, J. (2010). *Handbook of modern sensors: physics, designs and applications.* Springer Science and Business Media.

Frieden, B.R. (2004). *Science from Fischer Information.* Cambridge Univ. Press.

Geiger, G. (2006). State-of-the-art in leak detection and localisation. 1^{st} *PipeLine Technology Conference.*

Luyben, W.L., Tyreus, B.D., and Luyben, M.L. (1998). *Plantwide Process Control.* McGraw-Hill.

Magnis, L. and Petit, N. (2013). Impact of imprecise dating of measurements for bulk material flow network monitoring. In *Programmable Devices and Embedded Systems*, volume 12, 274–279.

Noble, S. (2012). Loss, latency, and speed - TN0021. Technical report, Data Expedition, Inc.

Préveral, A., Trihoreau, A., and Petit, N. (2014). Geographically-distributed databases: A big data technology for production analysis in the oil & gas industry. In *Proceedings of SPE Intelligent Energy Conference and Exhibition*, volume SPE167844.

White, T. (2012). *Hadoop: the definitive guide.* O'Reilly Media, 3rd edition.