# Probabilistic Leak Detection in Pipelines Using the Mass Imbalance Approach

Jonathan Rougier*

Senior Research Associate

Department of Mathematical Sciences

University of Durham, U.K.

**Abstract**

Typically, models of pipelines and pipe networks are calibrated to metered data by optimising the choice of parameters according to some penalty function. This approach does not provide a natural way to assess the predictive uncertainty when these models are used to infer the presence and description of a leak. This paper describes a fully-probabilistic approach in which the activities of calibration and prediction are unified, using the mass-imbalance approach to leak detection as an example. The resulting probability distribution over leak location and size can be presented graphically, or it can be used within an optimal decision framework to compute an effective response taking uncertainty into account. The approach is generalised to different leak-scenarios, including multiple leaks.

KEYWORDS: BAYESIAN, FRICTION FACTOR, RANDOM FIELD, CALIBRATION, CALIBRATED PREDICTION

*Mailing address: Department of Mathematical Sciences, University of Durham, South Road, Durham DH1 3LE, U.K., tel +44 (0)191 334 3111, fax +44 (0)191 334 3051, e-mail J.C.Rougier@durham.ac.uk.

# 1 Introduction

The development of methods for the detection of leaks in pipelines and pipe networks continues to be a very active research area (see, e.g., Mays, 1989; Pesta and Cassley, 1992; Mears, 1993; Mukherjee and Narasimhan, 1996; Rajtar and Muthiah, 1997, and the references below). Many proposals have the same underlying structure: to solve an inverse problem subject to parametric uncertainty and measurement error. The accepted approach appears to be to formulate the problem as an optimisation. Kapelan et al. (2003a) provides a recent example (see also Purdar and Liggett, 1992; Kapelan et al., 2003b, and the references therein). In this paper a model of a pipe network is parameterised by leak coefficients at each node and by Weisbach friction factors for each leg. Pressure observations are supplied for a subset of the nodes, and the sum of the squared differences between the observations and the physical model output is minimised over the possible values of the parameters. The result is a point estimate for the friction factors and the leak coefficients. In general a great deal of ingenuity is required to make this problem solvable in a practical sense, allowing that the optimand may be a rather irregular function of the parameters. For example, Kapelan et al. (2003a) propose a hybrid method combining the traditional gauss-newton approach with genetic algorithms.

It is not hard to see the deficiency in an optimisation approach, even in the cases where sufficient resources are available to ensure that a solution is found and that that solution is indeed a global optimum in the parameter space. The point estimate conveys no measure of uncertainty. While there are generalisations that use the curvature of the optimand to infer a variance matrix (following a *maximum*

*likelihood* approach), this type of approximation is only valid in the presence of large amounts of effectively-independent data: something that is seldom the case in the case of pipelines or pipe networks.

We need, however, to make a careful appraisal of uncertainty, because the detection of a leak is only the first stage in a complex decision about how to proceed. Decisions of this nature, i.e. potentially costly decisions made under uncertainty, are best approached within a *decision theoretic* framework in which actions are selected based on both probabilities and consequences (see, e.g., Lindley, 1985; Clemen, 1996; Bernado and Smith, 1994). What we would really like to compute in this situation is a joint probability distribution over leak location and size, taking account of parametric uncertainty and measurement errors, and also taking account of expert knowledge about the ways in which the particular pipe network might leak.

It may be argued that the probabilistic approach is simply too difficult, and the formulation of the leak detection problem as an optimisation is a practical compromise although, interestingly, this point is seldom made explicitly. This paper seeks to challenge this view, demonstrating how a simple leak detection approach—the mass imbalance method—can be made fully-probabilistic, and showing the kinds of results that can then be derived. The mass-imbalance approach is used because its simplicity lends itself to a clear exposition of the key features of the probabilistic approach, but it should be clear that the method can be extended immediately to more complex systems such as pipe networks.

The key point of this paper is that the probabilistic approach is a natural generalisation of the optimisation approach, or, to put it another way, the optimisation approach is actually the probabilistic approach but with very strong (and,

one might add, unrealistic) assumptions about the distributions of the uncertain quantities. Therefore methods that are currently formulated as optimisations can be made probabilistic quite straightforwardly, and, if necessary, in stages. Powerful computers and recent advances in inferential calculations, such as Markov chain Monte Carlo sampling (see, e.g., Robert and Casella, 1999), have helped to make fast probabilistic inference possible. These types of stochastic techniques can be improved by a first-stage optimisation in order to select an efficient proposal distribution, which is another way in which the probabilistic approach can be seen as a generalisation of the optimisation one.

The outline of the paper is as follows. Section 2 describes the steady state behaviour of a pipeline with a leak, and the deterministic mass imbalance leak location method; a full list of notation used in this and subsequent sections is given at the end of the paper (page 36 of this manuscript). Section 3 describes the sources of uncertainty, and how they may be accommodated within a probabilistic approach to leak description using the mass imbalance method. This includes a demonstration of two ways in which the probabilistic approach can be used to generalise the deterministic one. Section 4 describes how we might describe the engineer's knowledge about the Weisbach friction factor, which is often the major contributor to parametric uncertainty. This is treated as an uncertain function, or *random field*, along the pipeline. Section 5 provides a simple example, demonstrating one way in which probabilistic information about leak location and leak size may be presented. This section also illustrates that reasonable choices for the distributions of the unknown parameters can give rise to a large amount of uncertainty about the leak. Section 6 generalises the probabilistic model used to describe the engineer's knowledge about the ways in which the pipeline might leak,

to encompass leak detection and leak attribution. Finally, section 7 concludes.

## 2 Steady state without and with a leak

The steady state equation for a non-leaking pipeline is

$$H - \frac{\rho(L)}{D}\,\frac{v^2}{2g} - \kappa\frac{v^2}{2g} = 0 \tag{1a}$$

where $H$ is the upstream piezometric head, $L$ and $D$ the pipeline length and pipe diameter, $v$ the velocity (assumed to be positive), $g$ the gravitational acceleration, $\kappa$ the downstream valve coefficient, and

$$\rho(x) \triangleq \int_0^x f(z)\,dz \tag{1b}$$

where $f(x)$ is the Weisbach friction factor at location $x \in [0, L]$. This friction factor is allowed to be non-constant along the pipeline, according to variations in pipe roughness.

In (1a) the first term $H$ is the total head available, and the second and third terms describe how that head is dispersed by friction along the pipeline, and by the valve at the downstream end of the pipeline. Details may be found in, e.g., Massey (1989) or Wylie and Streeter (1993). Eq. (1) is one of the simplest configurations of a pipeline, in terms of its boundary conditions and its lack of features such as additional valves, or constrictions such as bends or variations in pipe diameter. This simplicity aids presentation, but it is not a restriction on the application of the probabilistic approach described below. The probabilistic approach can be generalised in exactly the same way as the physical model, since it sits 'on top' of

the physical model.

If we introduce a single leak at location $\ell \in (0, L)$, then we can describe the loss of liquid through the leak with an orifice equation

$$v_0 - v_L = \gamma \sqrt{2gh(\ell)} \tag{2a}$$

where $v_0$ and $v_L$ are velocities at the upstream and downstream ends of the pipeline, $\gamma$ is a lumped coefficient representing leak size and discharge coefficient, and $h(\ell)$ is the piezometric head at the leak location. The steady state equation is now

$$\underbrace{H - \frac{\rho(\ell)}{D}\frac{v_0^2}{2g}}_{h(\ell)} - \frac{\rho(L) - \rho(\ell)}{D}\frac{v_L^2}{2g} - \kappa\frac{v_L^2}{2g} = 0 \tag{2b}$$

where the second and third terms represent the head dispersed by friction above and below the leak at $\ell$, and $h(\ell)$ is indicated. The friction factor is taken to be invariant to the changes in fluid velocity that follow the onset of the leak. In other words velocity is sufficiently high or the pipe sufficiently rough that the flow is fully turbulent, to the extent that even a large leak does not affect the Reynold's number sufficiently to impact on the friction factor. Alternatively, we could restrict attention to small leaks, for which $v_0$ and $v_L$ remain close to $v$, the pre-leak velocity.

Taken together, (2) gives us three equations in the three unknowns $v_0$, $v_L$ and $h(\ell)$. If we had precise values for the other parameters, we could solve for these unknowns exactly. For leak detection, however, we need to solve a different problem, because the leak is the unknown. This paper considers the *mass imbalance* approach (see, e.g., Mears, 1993; Wang et al., 1993; Liou, 1994; Rajtar and

Muthiah, 1997), in which we infer the leak location from information about $v_0$ and $v_L$. With these two data we can solve (2b) for $\rho(\ell)$, which we can then invert to find $\ell$, since $\rho(\cdot)$ is continuous and strictly increasing. The key question is how best to proceed when we do not have precise values of the other parameters. In the next section we treat this question in terms of probabilities.

# 3 Inferential calculations

## 3.1 Sources of uncertainty

There are three sources of uncertainty when applying the mass imbalance approach. First, *measurement uncertainty*. The values $v_0$ and $v_L$ are not necessarily the true steady-state velocity values, owing to transient fluctuations in the fluid, and to measurement errors. The same could be said for the piezometric head $H$, but in this paper we are treating this as known, for simplicity. We can remove the effect of transient fluctuations by time-averaging, because transient behaviour in an incompressible fluid tends to die away rapidly. This assumption is common in steady-state methods (Mukherjee and Narasimhan, 1996). However, the restriction to steady state is not a necessary condition for a probabilistic treatment: Rougier and Goldstein (2001) derive a fully-probabilistic treatment of transient fluid behaviour, useful in situations where the boundary conditions are constantly changing, or where the transients reflect extreme events like rapid valve closure. Removing transient effects leaves systematic meter bias as the main source of measurement uncertainty, for example mis-calibration or externally-driven bias such as the effects of changes in ambient temperature. These tend to vary slowly in time to the extent that they can be treated as unknown constants for real-time

monitoring.

The second source of uncertainty is *parametric uncertainty*. At any point in time we may not know the exact setting of, say, the downstream valve. More generally, we cannot ever know the function $\rho(\cdot)$, which is effectively an infinite-dimensional parameter (i.e. one for which no finite set of measurements will ever be sufficient). Furthermore, we seldom know, before the event, the location and size of the leak, should one occur.

The third source of uncertainty is *model uncertainty*, also termed *model inadequacy*. Our model, described by (1) in the absence of a leak, is not a perfect analogue for the underlying pipeline. To put this more formally, we cannot imagine making precise measurements of the parameters with the property that the model evaluated at those parameters would give an exact match with precise measurements of pressure and flow. To give one example, we do not know how to make a precise measurement of the 'relative roughness' of a pipe, which is an abstraction of a much more complex phenomenon. This is discussed in more detail in section 4.

The treatment of model inadequacy is a rapidly developing area in statistics (see, e.g., Kennedy and O'Hagan, 2001; Craig et al., 2001; Goldstein and Rougier, 2004; Higdon et al., 2004), with direct implications for both model calibration and prediction. The simplest approach, which we will adopt here for convenience, is to accept the *strong constraint*. This states that although we know our model to be imperfect, we still believe that somewhere in the parameter space there is a combination of values that would make our model a perfect match for the pipeline itself. This is consistent with optimisation approaches to calibration which are prevalent in leak detection, as discussed in the Introduction. The less

restrictive model, sometimes referred to as the *best input* model (Rougier, 2004), is a straightforward generalisation that can be added into the probabilistic approach below, when the strong constraint is deemed inappropriate.

### 3.2 The probabilistic mass imbalance approach

We recast the mass imbalance approach in probabilistic terms. The task is to infer the distribution of leak location $\ell$ and leak coefficient $\gamma$ conditional on imperfect observations of the upstream and downstream velocities, and taking account of our uncertainty about the valve coefficient $\kappa$ and uncertainty about the friction factor, which leads to uncertainty about $\rho(\cdot)$.

The first point to note is that we if we have two velocity meters we should have more than two velocity observations: potentially we have four. Prior to the leak occurring we have the measurements $v_0^0$ and $v_L^0$, where the superscript '0' indicates that no leak has occurred. Since both velocities should be equal to $v$, any difference between them must be on account of systematic meter bias. Denote the biases as $\epsilon_0$ and $\epsilon_L$, so that $v_0^0 \equiv v + \epsilon_0$ and $v_L^0 \equiv v + \epsilon_L$. Then after the leak occurs we have two different measurements, *but the same systematic biases*. In other words, $v_0^1 \equiv v_0 + \epsilon_0$ and $v_L^1 \equiv v_L + \epsilon_L$, where the superscript '1' indicates the presence of a single leak. We can choose to disregard the measurements of $v_0^0$ and $v_L^0$, just as the deterministic mass imbalance approach does, but the probabilistic approach allows us to include these data, and we would expect them to be informative because they can help us to 'correct' our prediction to take account of meter bias.

In this subsection, however, we use only the post-leak values $v_0^1$ and $v_L^1$, to stay as close as possible to the deterministic mass-imbalance method, and to give the clearest possible explanation of the general probabilistic approach. In

subsection 3.3 we generalise to include pre-leak values as well.

Our objective is to compute $\mathsf{Pr}\big(\ell, \gamma \mid \bar{v}_0^1, \bar{v}_L^1\big)$, where the observed values of random variables are denoted with an overbar. However, in order to make use of the physics in (2), we need to introduce additional uncertain quantities, namely $\rho(\ell)$, $\rho(L)$ and $\kappa$. For simplicity we eliminate $h(\ell)$ using (2a) and (2b), to give

$$v_0 - v_L = \gamma \sqrt{2g \left( H - \frac{\rho(\ell)}{D} \frac{v_0^2}{2g} \right)}. \tag{3}$$

Writing

$$\theta \triangleq \big(\ell, \gamma, \rho(\ell), \rho(L), \kappa\big), \quad \epsilon \triangleq \big(\epsilon_0, \epsilon_L\big), \quad \text{and } V \triangleq \big(v_0^1, v_L^1\big) \tag{4}$$

(for reasons that will be explained immediately below) the *predictive* Probability Density Function (PDF) becomes

$$\mathsf{Pr}\big(\theta, \epsilon \mid \bar{V}\big) \equiv c\, \mathsf{Pr}\big(\theta, \epsilon, V = \bar{V}\big) = c\, \mathsf{Pr}\big(V = \bar{V} \mid \theta, \epsilon\big)\, \mathsf{Pr}\big(\theta, \epsilon\big) \tag{5}$$

where $c \triangleq \mathsf{Pr}\big(V = \bar{V}\big)^{-1}$, and $\bar{V}$ are the observed values of $V$. The equivalence follows from the definition of conditional probability, and the following equality is a factorisation of the joint distribution that is often referred to as *Bayes's Theorem*, the two PDFs being the *likelihood function* and the *prior distribution*, respectively. In practise we seldom need to compute the normalising constant $c$ explicitly, and so we can ignore it in what follows.

The problem is that (5) is not easily computable, because the likelihood function is degenerate: it is zero almost surely for reasonable choices of $\big(\theta, \epsilon\big)$, because when we put such a choice into (2b) and (3) along with the two values in $\bar{V}$ these

two equations will not hold simultaneously (technically they will hold only on a set of measure zero). This requires quite careful handling. We must choose two of our uncertain parameters to be constrained by the physical model and the data combined. The natural choices are the two values in $\epsilon$, because for given $\theta$ the two equations (2b) and (3) define a bijective mapping,

$$\epsilon \longleftrightarrow V. \tag{6}$$

That is to say, if we are given $\theta$ and $V$ we can solve for $\epsilon$, and if we are given $\theta$ and $\epsilon$ we can solve for $V$. This may not be obvious. The reason is that using $\theta$ we can solve solve (2b) and (3) for $\big(v_0, v_L\big)$. Then we can either compute $\epsilon$ using $V$, or $V$ using $\epsilon$, since

$$V \equiv \big(v_0, v_L\big) + \epsilon. \tag{7}$$

Denote by $\bar{\epsilon}$ the solution values for $\epsilon$ using $\bar{V}$ at the given $\theta$. The predictive distribution is not affected if we remove from the prior PDF regions of the parameter space with zero likelihood. Therefore we can express (5) as

$$\mathsf{Pr}\big(\theta \mid \bar{V}\big) = c\,\mathsf{Pr}\big(V = \bar{V} \mid \theta, \bar{\epsilon}\big)\,\mathsf{Pr}\big(\theta\big) \tag{8a}$$

where we have removed all non-compatible values of $\epsilon$ by making $\epsilon$ a deterministic function of $\theta$ and $\bar{V}$.

Following on from (6), the event $\big(V = \bar{V} \mid \theta, \bar{\epsilon}\big)$ is equivalent to the event $\big(\epsilon = \bar{\epsilon} \mid \theta, \bar{V}\big)$. Making this substitution, the resulting form for the predictive distribution is

$$\mathsf{Pr}\big(\theta \mid \bar{V}\big) = c\,\mathsf{Pr}\big(\epsilon = \bar{\epsilon} \mid \theta, \bar{V}\big)\,\mathsf{Pr}\big(\theta\big)\,. \tag{8b}$$

If we want to simplify this further we have to make some additional choices about the marginal distribution of $\big(\theta, \epsilon\big)$. An obvious simplification presents itself: the engineer's knowledge about the meter biases is independent of his or her knowledge regarding the other uncertain quantities. Taking this to be acceptable, we can simplify the predictive distribution to

$$\mathsf{Pr}\big(\theta \mid \bar{V}\big) = c\, \mathsf{Pr}\big(\epsilon = \bar{\epsilon}\big)\, \mathsf{Pr}\big(\theta\big) \tag{8c}$$

where it is understood that the value $\bar{\epsilon}$ is a function of $\theta$ and $\bar{V}$. In other words, the likelihood function for $\theta$ is constructed from the engineer's knowledge about the meter biases. Informally, the larger the biases are thought to be, the less informative the data in $\bar{V}$ will be about $\theta$, where $\theta$ includes the leak's location $\ell$ and size coefficient $\gamma$.

Like the likelihood function, the prior distribution $\mathsf{Pr}\big(\theta\big)$ can also benefit from some quite natural simplifications. First, the engineer's knowledge about the friction function $\rho(\cdot)$ is very likely to be independent of his or her beliefs about the other uncertain quantities. Similarly, the engineer's knowledge about the valve coefficient $\kappa$ is very likely to be independent of the other uncertain quantities. Finally, the engineer's knowledge about the leak coefficient $\gamma$ is very likely to be conditionally independent of the other uncertain quantities given the leak location, $\ell$. If these are acceptable, we can write the prior distribution in (8c) as

$$\begin{aligned} \mathsf{Pr}\big(\theta\big) &\equiv \mathsf{Pr}\big(\ell, \gamma, \rho(\ell), \rho(L), \kappa\big) \\ &= \mathsf{Pr}\big(\rho(\ell), \rho(L) \mid \ell\big)\, \mathsf{Pr}\big(\gamma \mid \ell\big)\, \mathsf{Pr}\big(\ell\big)\, \mathsf{Pr}\big(\kappa\big)\,. \end{aligned} \tag{9}$$

We are taking the pipeline length $L$ as known throughout, so the first PDF repre-

sents the engineer's knowledge about $\rho(\cdot)$ at two known locations $\ell$ and $L$. Note that the one dependence that has been preserved in (9) is that of leak size on leak location. This allows the engineer to incorporate the knowledge that leaks in certain locations are likely to be bigger than others, for example those that occur in regions of seismic activity. More details of this kind of statistical modelling are given in section 6.

Using (8c) and (9), and specifications for the PDFs therein, we can turn the data $\bar{V}$ into a prediction for $\big(\ell, \gamma\big)$, using standard probabilistic tools such as Monte Carlo sampling. Examples will be given in sections 5 and 6.

Note that the probabilistic approach can go where the deterministic one cannot: leak location with a single meter. Suppose that we have just the downstream meter. Given $\theta$, we still have a bijective mapping between $\epsilon_L$ and $v_L^1$, and consequently we can compute $\bar{\epsilon}_L$, and the predictive distribution for $\theta$ becomes

$$\mathsf{Pr}\big(\theta \mid \bar{v}_L^1\big) = c\, \mathsf{Pr}\big(\epsilon_L = \bar{\epsilon}_L\big)\, \mathsf{Pr}\big(\theta\big) \tag{10}$$

where $c \triangleq \mathsf{Pr}\big(v_L^1 = \bar{v}_L^1\big)^{-1}$, and $\mathsf{Pr}\big(\theta\big)$ is as given in (9). The amount of information that the observation $\bar{v}_L^1$ can contribute will depend strongly on the prior. For example, if $\mathsf{Pr}\big(\ell\big)$ consisted of two widely-separated narrow peaks then we would expect $\bar{v}_L^1$ to be highly informative, as it would typically select one peak or the other. On the other hand, if $\mathsf{Pr}\big(\ell\big)$ was flat over the interval $(0, L)$ then we would expect that $\bar{v}_L^1$ would not do much more than put a bulge or a tilt into the predictive distribution of $\ell$.

### 3.3 Using pre-leak measurements

We can follow the steps outlined in the previous section to incorporate the extra information available in the observations $\left(v_0^0, v_L^0\right)$. We use the difference $v_0^0 - v_L^0 \equiv \epsilon_0 - \epsilon_L$, and redefine the collection $V$ as

$$V \triangleq \left(v_0^0 - v_L^0, v_0^1, v_L^1\right). \tag{11}$$

We can write the relation between the observations $V$ and the biases $\epsilon$ as

$$\begin{pmatrix} v_0^0 - v_L^0 \\ v_0^1 \\ v_L^1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v_0 \\ v_L \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \epsilon_0 \\ \epsilon_L \end{pmatrix}. \tag{12}$$

As we can solve (2b) and (3) for $\left(v_0, v_L\right)$ using $\theta$, we can map from $\epsilon$ to $V$ given $\theta$, using (12). This mapping is not invertible, but we can use the structure of (2) to show that is it still possible to map back from $V$ to $\epsilon$ given $\theta$. We can write the lefthand side of (2a) as

$$\begin{aligned} v_0 - v_L &\equiv (v_0^1 - \epsilon_0) - (v_L^1 - \epsilon_L) \\ &= v_0^1 - v_L^1 - (\epsilon_0 - \epsilon_L) \\ &\equiv v_0^1 - v_L^1 - (v_0^0 - v_L^0) \end{aligned} \tag{13}$$

which is a known value given $V$. Therefore we can solve (2a) directly for the value of $h(\ell)$, which we can substitute into (2b) to solve for $v_L$. Once we have solved for $v_L$ we can compute $\epsilon_L \equiv v_L^1 - v_L$, and then we can find $\epsilon_0 \equiv (v_0^0 - v_L^0) + \epsilon_L$.

Therefore we have shown once again that given $\theta$ we have a bijective relationship

$$\epsilon \longleftrightarrow V, \tag{14}$$

where $V$ is now the collection given in (11).

Therefore (8c) still holds under the new definition of $\bar{V}$ (being the observed values of the new $V$). The collection $\bar{\epsilon}$ has the same meaning, namely those values that are implied by $\theta$ and $\bar{V}$, but the procedure for computing $\bar{\epsilon}$ is different, and, in fact, simpler. No extra input is required of the engineer, so the incorporation of the extra datum $\bar{v}_0^0 - \bar{v}_L^0$ is genuinely a 'free upgrade'.

### 3.4 Likelihood-based approaches to inference

It may be helpful to outline the statistical justification for the optimisation approach so that we can see how it could be extended to include uncertainty assessments regarding the leak, and how this contrasts with probabilistic methods.

Optimisation through minimising a least-squares penalty function can be justified in inferential terms using *maximum likelihood* (see, e.g., Pawitan, 2001). In general, the Maximum Likelihood Estimator (MLE) $\hat{\theta}$ satisfies

$$\hat{\theta} \triangleq \max_{\theta} \mathsf{Pr}\big(\epsilon = \bar{\epsilon}\big) \tag{15}$$

remembering that $\bar{\epsilon}$ is a function of $\theta$ and $\bar{V}$. We must assume that the meter biases are gaussian, with known means and variances. Taking the means to be

zero and denoting the variance matrix as $\Sigma$, the MLE is then

$$\hat{\theta} = \max_{\theta} \left\{ -\bar{\epsilon}^T \Sigma^{-1} \bar{\epsilon} \right\} \tag{16}$$

where $\bar{\epsilon}^T$ is the transpose of the vector $\bar{\epsilon}$. This is the (generalised) least-squares formulation. However, in order to make inferences about the leak, i.e. to have more than a point estimate for $\theta$, we ought to compute not the MLE but the *profile likelihood*

$$\mathsf{L}\big(\ell, \gamma\big) \triangleq \max_{\theta \setminus \{\ell, \gamma\}} \mathsf{Pr}\big(\epsilon = \bar{\epsilon}\big) \tag{17}$$

(Pawitan, 2001, pp. 61–64). Using this profile likelihood we can calculate a value for the variance of our estimates for $\big(\ell, \gamma\big)$, although this calculation is not at all straightforward. The calculated value is asymptotically correct in the limit as the length of $\epsilon$ becomes large, and subject to conditions on the approximate independence of the components of $\epsilon$. The problem is, however, that for leak detection the measurements are seldom abundant and often dependent. For the simple mass-imbalance approach, for example, we cannot feel confident about an asymptotically-justified measure of uncertainty based on only two data. Therefore likelihood-based approaches would appear not to be particularly useful in determining our uncertainty about $\theta$.

Once we move away from likelihood-based inference in our search for measures of uncertainty about the leak, it is natural first of all to consider the probabilistic case with a flat prior distribution for $\theta$, subject to limits on the values of each of the components. In this case we can treat the likelihood function as the predictive distribution, subject to renormalisation. No new choices have to be made (except for the limits), but the treatment of the data is quite different. Rather than

maximise the likelihood function for a point estimate of $\theta$, we would now sample from it (as described in section 5), or summarise it directly by integrating over $\theta$ to find the mean vector and variance matrix.

But at this point we may well ask whether a flat prior for $\theta$ is really an adequate reflection of the engineer's knowledge about the pipeline. The fully-probabilistic approach provides the engineer with an opportunity to incorporate his or her knowledge about the pipeline into the inferential calculation. The engineer can adopt or feign ignorance and stick with a flat prior, but this is a choice and not a mandatory part of the inferential process. For example, if some leaks are believed to be more likely than others, then this information can go into $\mathsf{Pr}\big(\theta\big)$, likewise if some parts of the pipeline are believed likely to be rougher than others, or small values for the valve coefficient are believed to be more likely than large ones.

To summarise the argument in this subsection, the optimisation-based approach to choosing a point value for $\theta$ does not generalise easily, if at all, to computing measures of uncertainty. If we want such measures then we have to adopt a fully-probabilistic approach. Once we have made this step, the opportunity arises to incorporate the engineer's knowledge into the inferential process, which we expect to improve our assessment of $\theta$ to the extent that the engineer's knowledge about the pipeline is well-founded.

## 4 Modelling knowledge about the friction factor

The previous calculations involve the PDF of $\rho(\cdot)$ at given locations along the pipeline. The function $\rho(\cdot)$, defined in (1b) is uncertain because the Weisbach friction factor $f(\cdot)$ is uncertain. This section explains why $f$ might be treated as

uncertain, and how $f$ can be modelled in a flexible manner so that the resulting solution remains tractable.

### 4.1 Sources of uncertainty

The standard method to assess $f(\cdot)$ is to determine from experimental data the equivalent sand-roughness of the pipe material, and then to determine, by further experiment, the relation between equivalent sand-roughness and the friction coefficient. In practise it is uneconomical to perform these experiments, and so the results of standard experiments are used instead, in the form of tabulated values or mathematical formulae (see, e.g., Colebrook, 1939). The consensus, as summarised by Massey (1989, p. 204) is that "Accurate prediction of friction losses is thus difficult to achieve." As for the scale of the uncertainty, Haaland (1983, p. 90) notes that the Colebrook-White formula "... may be 3-5 percent, if not more, in error as compared to experimental data", on top of which should be added uncertainty about equivalent sand-roughness. As losses to friction will dominate the other losses in many pipelines, this seems to justify a careful stochastic analysis.

### 4.2 Beliefs about the friction coefficient

The two requirements of a stochastic approach are (a) to find a way to model knowledge about the friction coefficient along the pipe, and then (b) to turn this into

$$\mathsf{Pr}\big(\rho(x_1), \ldots, \rho(x_n) \mid x_1, \ldots, x_n\big)\,, \tag{18}$$

where $\big(x_1, \ldots, x_n\big)$ are any finite set of locations along the pipeline. This latter PDF is required in the the inferential calculations described in the previous sec-

tion, for the special case $\big(x_1, x_2\big) = \big(\ell, L\big)$. But the more general treatment is useful in case we have access to pressure readings at known locations along the pipeline, which we would like to incorporate into our predictive distribution using the methods previously described.

The technical difficulties of part (b) will restrict the choices for part (a). In fact the only tractable and non-trivial model for the friction coefficient is a *gaussian random field*. This model is parametrised by a mean function $\mu_1(\cdot)$ and a covariance kernel $\mu_2(\cdot, \cdot)$:

$$\mu_1(x) \triangleq \mathsf{E}\big(f(x)\big)\,, \quad \text{and} \quad \mu_2(x, x') \triangleq \mathsf{Cov}\big(f(x), f(x')\big)\,. \tag{19}$$

The mean function is the best guess for the pointwise value of the friction factor along the pipeline (it may, for example, be a constant). The covariance kernel $\mu_2$ describes two aspects of uncertainty about the difference between the actual friction factor and the mean value. First there is the pointwise uncertainty, i.e. how far the true value might deviate from the mean value at each point. This is represented by the standard deviation $\sqrt{\mu_2(x, x)}$. Second, there is the degree to which deviations at different points are related, represented by the correlation $\mu_2(x, x')/\sqrt{\mu_2(x, x)\,\mu_2(x', x')}$.

As we assume that $f$ is a *gaussian* random field the joint distribution of any finite collection is gaussian. This is not ideal, as the gaussian distribution does not respect the constraint that $f(x) > 0$. However, it will usually be the case that $\mathsf{Pr}\big(f(x) \leq 0\big)$ is tiny, in which case the implicit truncation of $f(x)$ at zero will have no practical effect. The great practical advantage of using a gaussian random field is that it follows that (18) is *also* gaussian. The mean and covariance

of designated points on $\rho(\cdot)$ are then given by

$$\begin{aligned} \mathsf{E}\big(\rho(x) \mid x\big) &= \int_0^x \mu_1(z)\, dz, \text{ and} \\ \mathsf{Cov}\big(\rho(x), \rho(x') \mid x, x'\big) &= \int_0^x \int_0^{x'} \mu_2(z, z')\, dz'\, dz, \end{aligned} \tag{20}$$

see, e.g., Parzen (1962, sec. 3-3), or Bartlett (1978, sec. 5.11).

### 4.3 Choices for the covariance kernel

The simplest choice for $\mu_2(x, x')$ is $\mu_2(x, x') = 0$. This would be appropriate if there was complete certainty about the friction coefficient at every location along the pipe. The next simplest choice is $\mu_2(x, x') = \sigma^2$ for some known quantity $\sigma$. This would be appropriate if a single error-free measurement at any location in the pipe would be sufficient for complete knowledge at every point in the pipe. In other words, there is perfect correlation between $f(x)$ and $f(x')$ for all $x$ and $x'$. These two choices would appear to be inappropriate for real-world pipes.

If knowing the friction coefficient at $x$ was not sufficient for knowing the coefficient at some other location $x'$ then the correlation between $f(x)$ and $f(x')$ would have to be less than one. If the pointwise uncertainty was the same along the pipe and the correlation depended only on the separation between $x$ and $x'$ then the appropriate model would be a *stationary* kernel of the general form $\mu_2(x, x') = \sigma^2 r(|x - x'|)$, where the function $r(\cdot)$ must be positive definite (see, e.g., Bartlett, 1978, sec. 6.1). Here $\sigma$ describes the pointwise uncertainty about the friction coefficient, while $r(\cdot)$ controls how the correlation between coefficients at two different locations drops according to their separation. Essentially, $r(\cdot)$ controls the predictability of $f(x')$ from knowing the value $f(x)$.

More general specifications for $\mu_2(x, x')$ are possible, in which the variance is not required to be the same at all points but depends upon the location, and in which correlation as a function of distance also depends upon the location. These would be useful in the presence of detailed beliefs about the behaviour of the friction coefficient along the pipe. The stationary case with appropriate selection of the function $r(\cdot)$ should be sufficiently flexible to model general uncertainty.

### 4.4 Practical issues

In general the friction factor depends on both the equivalent sand-roughness and on the Reynolds number. In this paper, however, we restrict attention to the case where the pre-leak velocity is sufficiently large or the leak size sufficiently small that the friction factor both before and after the leak can be modelled with a constant known Reynolds number. Writing $K(x)$ for the equivalent sand-roughness at location $x$ and 'Re' for the Reynolds number, we can use the Colebrook-White formula (Haaland, 1983, p. 89)

$$\frac{1}{\sqrt{f(x)}} = -2\log_{10}\left(\frac{2.51}{\mathrm{Re}\sqrt{f(x)}} + \frac{K(x)}{3.7D}\right) \tag{21}$$

to induce uncertainty about $f(x)$ from uncertainty about $K(x)$. Unfortunately $f(x)$ is not a linear transformation of $K(x)$, and so we cannot assign a gaussian random field to $K(\cdot)$ and then infer the mean function and variance function for $f(\cdot)$ directly. However, we can use our pointwise uncertainty about $K(x)$ to induce pointwise uncertainty about $f(x)$. The simplest way to do this is by sampling $K(x)$ and then summarising the transformed values for $f(x)$ in terms of the mean and standard deviation.

If we accept that a stationary random field is appropriate for $f(\cdot)$, then there is a natural choice for $r(\cdot)$, namely the *Ornstein-Uhlenbeck* correlation function

$$r(d) = \exp\big(-\tau d\big) \tag{22}$$

for $\tau > 0$. This correlation function generates quite 'rough' sample paths (no pun intended), that are continuous but not differentiable in the mean-squared sense (see, e.g. Bartlett, 1978, sec. 5.11). It has the *markov* property that, for three locations $x < x' < x''$,

$$\mathsf{Cov}\big(f(x), f(x'') \mid f(x')\big) = 0.$$

In other words if we knew the friction coefficient at $x'$ then also knowing the coefficient at $x$ would be of no additional help in predicting the coefficient at $x''$. This seems appropriate for pipelines, in which both the original pipe roughness and the subsequent internal surface degradation might be driven primarily by local factors. This correlation function also has the attractive feature that we can compute the covariance of $\rho(\cdot)$ directly:

$$\mathsf{Cov}\big(\rho(x), \rho(x') \mid x, x'\big) = (\sigma/\tau)^2(1 - e^{-\tau x'})(e^{\tau x} - 1) \tag{23}$$

for $x' \geq x$. This explicit form for the covariance helps to speed up the inferential calculations because we need to compute $\mathsf{Cov}\big(\rho(\ell), \rho(L) \mid \ell\big)$ for every candidate value for $\ell$.

# 5 A simple example

Consider the following example. Oil with kinematic viscosity $10^{-5}\,\mathrm{m}^2/\mathrm{s}$ flows through a cast-iron pipe $D = 0.1\,\mathrm{m}$ in diameter and $L = 1000\,\mathrm{m}$ long under a piezometric head difference of $H = 50\,\mathrm{m}$, terminating at a valve.

## 5.1 Specifying prior knowledge

In order to treat the leak detection probabilistically, we need to specify the engineer's prior knowledge about (a) the friction factor and the valve coefficient; (b) the likely location and size of the leak; and (c) the velocity meter accuracy. In this subsection we specify 'reasonable' choices that may be applied across a range of actual problems.

The PDF of the friction factor is chosen using the approach described in section 4.4, based on knowledge about the relative roughness of cast iron. Before we do this we need a typical value for the velocity, to compute the Reynolds number. This can be found from (1) and (21), setting the valve coefficient $\kappa$ and the equivalent sand-roughness $K(\cdot)$ to their best guess values, for which we choose 49 and $0.25 \times 10^{-3}\,\mathrm{m}$, respectively. This gives a velocity of $1.64\,\mathrm{m/s}$, and a Reynolds number of $16.4 \times 10^3$. This combination of Reynolds number and relative roughness put the pipeline in the transition zone between laminar flow and complete turbulence.

Using this Reynolds number we can sample a large number of values for $f(\cdot)$ based on uncertainty about $K(\cdot)$. We treat the marginal distribution of $K(x)$ as invariant to $x$, and assign it a Gamma distribution. The Gamma distribution is a simple but flexible choice for strictly positive quantities; its PDF is given below

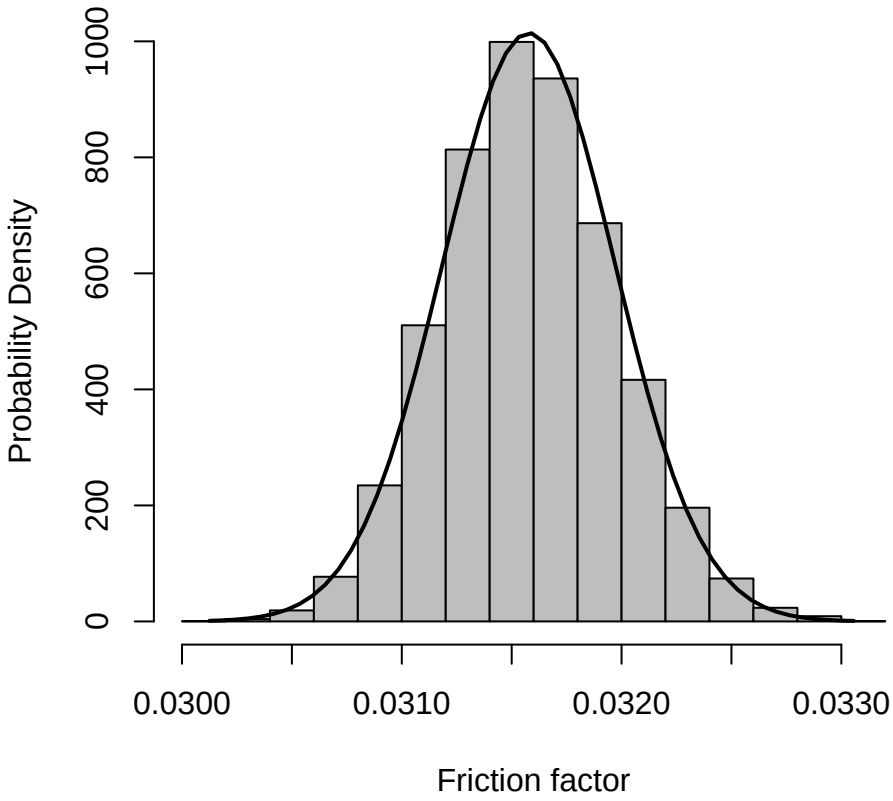


Figure 1: The marginal distribution of the Weisbach friction factor $f(x)$, based on uncertainty about the equivalent sand-roughness of cast iron. The overlaid curve shows a gaussian density function.

in eqrefeq:Gamma. For $K(\cdot)$ we choose a mean of $0.25 \times 10^{-3}$ and a standard deviation one tenth as large as the mean. The resulting sample for the marginal distribution of $f(x)$ at any $x$ is shown in Figure 1. A gaussian PDF is overlaid, showing that modelling $f(\cdot)$ with a gaussian random field is quite reasonable in this case. The mean and standard deviation of $f(x)$ are chosen to be 0.0316 and $0.396 \times 10^{-3}$. The Ornstein-Uhlenbeck correlation function, given in (22), is used to describe the spatial structure of $f(\cdot)$. The value $\tau = 0.513 \times 10^{-3}$ is chosen, which gives a correlation of 0.95 at a separation of 100 m and 0.60 at 1000 m.

The PDF for the valve coefficient is chosen to be Gamma, with mean 49 and standard deviation 7. This choice for $\kappa$ corresponds to the belief that most of the head loss in the pipeline is due to friction.

Initially our beliefs about the leak location and size are taken to be very simple. The PDF for $\ell$ is chosen to be uniform on the interval $(0, L)$, and the PDF for $\gamma$ is chosen to be Gamma with mean 0.003 and standard deviation 0.003. This choice for $\gamma$ corresponds to the belief that the leak will be small, of the order a few percent of total velocity. In section 6 we show how more careful modelling of the leak can be implemented.

Finally, for our beliefs about meter accuracy we choose independent gaussian PDFs for $\epsilon_0$ and $\epsilon_L$, with mean zero and standard deviation equal to 1 percent of typical velocity.

### 5.2 Results

The predictive distribution of $\big(\ell, \gamma\big)$ is summarised using *rejection sampling*. Simple descriptions of the principles of this kind of sampling are given in Ripley (1987) and Smith and Gelfand (1992) while Robert and Casella (1999) provides further details, including extensions to the Markov chain Monte Carlo (MCMC) approach.

The simplest form of rejection sampling is used. First, a point is sampled from $\mathsf{Pr}\big(\theta\big)$. Using this value and the data $\bar{V}$ we can compute $\bar{\epsilon}$, as described in section 3, and then the likelihood value $v \stackrel{\triangle}{=} \mathsf{Pr}\big(\epsilon = \bar{\epsilon}\big)$. The point is then kept or rejected with probability $v/M$, where

$$M \geq \sup_{\theta} \mathsf{Pr}\big(\epsilon = \bar{\epsilon} \mid \theta, \bar{V}\big)\,. \tag{24}$$

For $M$ we can use $\mathsf{Pr}\big(\epsilon = \mathbf{0}\big)$, which we know is at least as large as the supremum because $\epsilon$ has a gaussian distribution with zero mean. All the calculations were carried out in the R statistical programming environment (R Development Core Team, 2004).

The predictive distribution of $\big(\ell, \gamma\big)$ is summarised in terms of High Density Regions (HDR), e.g. as described in Tanner (1993). These are presented as iso-density contours bounding an area containing a designated amount of probability, in our case 95%, 50% and 20%. As these are computed from a finite sample they are estimates of the true HDR regions, and will show some sampling variation. Samples of size $4 \times 10^4$ were used, as this seemed sufficient to give smooth sampled contours.

To generate some measurement data we set all uncertain values to their $70^{\text{th}}$ percentile values: this is rather arbitrary but it avoids any suspicion that the results have been tuned in any way. The resulting leak descriptors are $\ell = 700$ and $\gamma = 0.0361$, and the measurements are

$$\big(\bar{v}_0^0, \bar{v}_L^0, \bar{v}_0^1, \bar{v}_L^1\big) = \big(1.63, 1.63, 1.66, 1.59\big). \tag{25}$$

When analysing the results the focus should not be on whether the probabilistic approach gets the prediction 'right' or not, as this depends in part on how consistent the measured data are with the engineer's prior knowledge: the choice of the $70^{\text{th}}$ percentile is meant to be representative of 'moderate consistency'. What is more important is to get a feeling for the amount of uncertainty that remains in the predictive distribution, which is proxied by the area of the HDRs. It is this uncertainty that the deterministic mass imbalance approach leaves out.

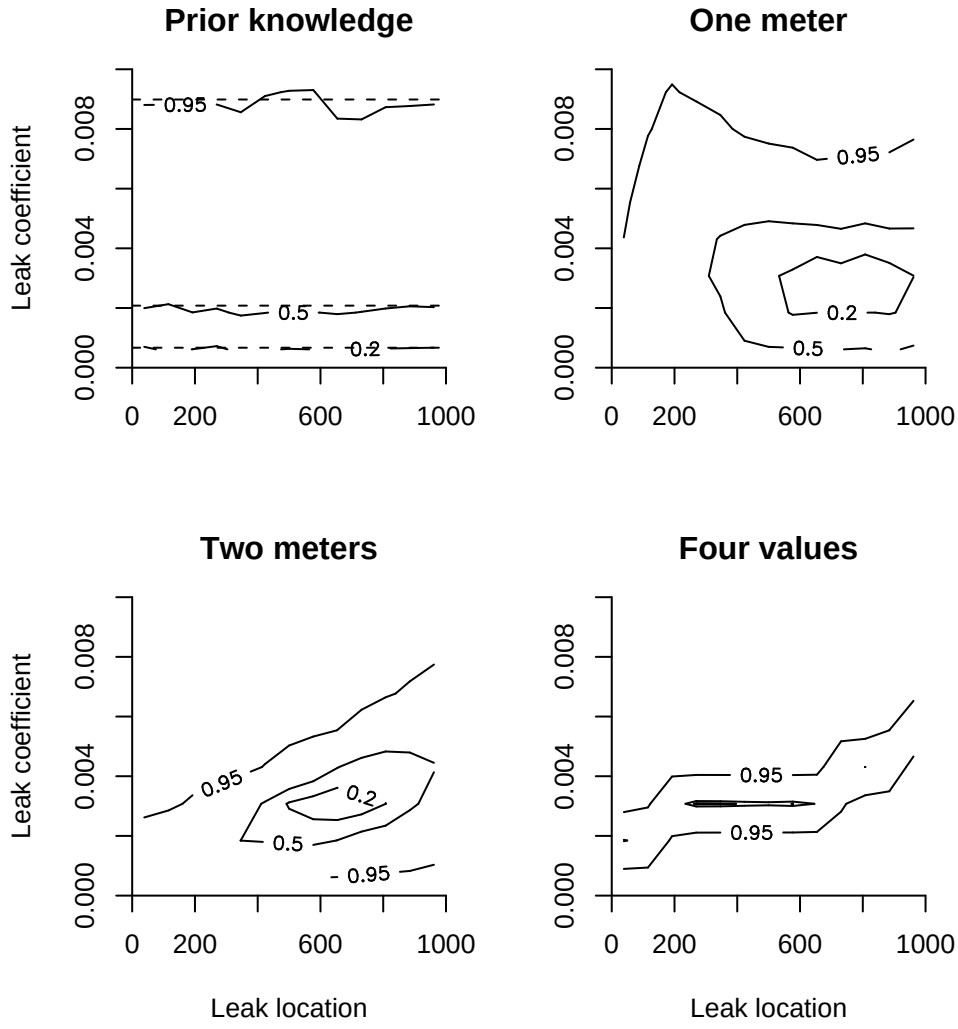


Figure 2: High density regions (95%, 50% and 20%) for the predictive distribution of leak location $\ell$ and leak coefficient $\gamma$ for four different sets of measurements. The top lefthand panel is prior to the data (the true values are shown as dashed lines, to show the amount of variation induced by the finite sample). The data in the remaining panels are $(\bar{v}_0^1, \bar{v}_L^1)$ for 'Two meters'; $(\bar{v}_L^1)$ for 'One meter'; and $(\bar{v}_0^0 - \bar{v}_L^0, \bar{v}_0^1, \bar{v}_L^1)$ for 'Four values'.

The results for four different types of inference are shown in Figure 2. In the top lefthand panel we have the prior distribution of $\big(\ell, \gamma\big)$. The iso-density contours from the true joint distribution are directly computable in this case, and they are shown using dashed lines, allowing us to get some feeling for the amount of variation induced by the finite sample. In the other three panels we have the three variations discussed in section 3.

Taking the standard case first, in the bottom lefthand panel, we see that a large amount of uncertainty remains after using the data $\big(\bar{v}_0^1, \bar{v}_L^1\big)$. Going by the 50% HDR, the leak location is restricted to the range $(300, 1000)$, and the leak coefficient to the range $(0.002, 0.005)$. While this represents a large decrease on the prior ranges, it indicates that the kinds of uncertainties we have chosen, which do not seem extreme, have a very substantial impact on our ability to describe the leak. The key uncertainty in terms of its impact on our predictive variance is the meter error $\mathsf{Var}\big(\epsilon\big)$, for which it could be argued that our choice of a standard deviation of only 1% is rather conservative.

The upward slope of the iso-density contours in this panel indicate that the predictive distribution of $\big(\ell, \gamma\big)$ is positively-correlated. This is a consequence of the orifice equation (2a). A given amount of leakage arises as a trade-off between $h(\ell)$ and $\gamma$, where $h(\ell)$ is itself closely and positively related to $L - \ell$. If our data are not strong enough to pin down either $\ell$ or $\gamma$, then the result will be a range of predictive values in which increased values for $\ell$ may be offset by increased values for $\gamma$.

In the top righthand panel we have the one meter case, in which the datum is the single observation $\bar{v}_L^1$. This is the case for which the mass imbalance method cannot give an answer. Contrasting this panel to the two meter case, we can see

that the uncertainty, e.g. as measured by the extent of the 50% HDR, is much larger, although still much less than the prior, particularly with regard to leak location.

In the bottom righthand panel we have the two meter case with pre- and post-leak observations, i.e. the dataset $\left(\bar{v}_0^0 - \bar{v}_L^0, \bar{v}_0^1, \bar{v}_L^1\right)$. As mentioned in section 3.3 this is effectively a 'free upgrade' to the standard case in the bottom lefthand panel. What is surprising is how much difference it makes: the 90% HDR is perhaps one fifth the size of that in the standard case. The reason is that, as mentioned above, the meter error is a key factor in determining the predictive variance. The extra observation allows us to incorporate information directly relevant to that error, since $v_0^0 - v_L^0 \equiv \epsilon_0 - \epsilon_L$. With the effect of meter error reduced, the positive correlation between $\ell$ and $\gamma$ that arises from the orifice equation is even more clearly visible.

## 6 Leak detection and attribution

In the previous section the engineer's knowledge about the leak was modelled very crudely. In this section we show how more complicated knowledge can be represented, within a more general framework in which we classify the leak according to a type. In this case the probabilistic approach encompasses leak *attribution* as well as leak description, i.e. we can compute the probability that the leak is of each of the specified types. If we include among our types a 'no leak', then attribution extends to leak *detection*, i.e. we can compute the probability that a leak has occurred.

We define a classifying variable $\mathcal{T} \in \left\{0, 1, \ldots, T\right\}$, which denotes whether the

pipeline has a leak, and what type of leak it is. Let $\mathcal{T} = 0$ denote no leak. All of our calculations so far have been predicated on the notion that there has been a leak, even if only a very small one, i.e. they have been conditioned on $\mathcal{T} > 0$. Let $\mathcal{T} = 1$ denote the leak described in section 5, which we will call an 'Ordinary Leak'. For the ordinary leak we have

$$\mathsf{Pr}\big(\ell, \gamma \mid \mathsf{ordinary}\big) \propto 1\big(\ell \in (0, L)\big) \times \mathsf{Gamma}\big(\gamma\,;\, a = 1, s = 0.003\big) \tag{26}$$

where $1(P)$ is the indicator function of the proposition $P$. The first term is for the uniform distribution on $(0, L)$, while the second is the gamma PDF with shape parameter $a$ and scale parameter $s$, which satisfies

$$\mathsf{Gamma}\big(\gamma\,;\, a, s\big) \propto \gamma^{a-1}\, e^{-\gamma/s}. \tag{27}$$

The mean and variance of a Gamma random variable are $as$ and $as^2$, respectively. For ordinary leaks the values $a = 1$ and $s = 0.003$ correspond to both the mean and the standard deviation being 0.003 (the special case with $a = 1$ is known as an *exponential* distribution). Note that users of computer software that can generate Gamma random variables should be alert to the fact that the two parameters can be specified in different ways; the version given in (27) corresponds to the R statistical programming environment (R Development Core Team, 2004).

Now we introduce a second type of leak. Suppose that the engineer was worried about sabotage along a section of the pipeline that ran alongside an access road. For leaks of this type we must have $\ell$ restricted to the section accessible by road, and the leak size might be much larger: say 5 times as large. Therefore the

Sabotage Leak has

$$\mathsf{Pr}\big(\ell, \gamma \mid \mathsf{sabotage}\big) \propto 1(\ell \in S) \times \mathsf{Gamma}\big(\gamma\,;\, a = 1, s = 0.015\big) \tag{28}$$

where $S \subset (0, L)$ is the accessible section of the pipeline. The choices $a = 1$ and $s = 0.015$ in (28) follow from our previous choice, where the scale parameter is 5 times larger than in the ordinary leak case. For the accessible section we choose $S = (300, 800)$ for our example.

We complete this more general model by assigning a probability to each of the types. In section 5 we have $\mathcal{T} = 1$ with probability 1, so that $\mathsf{Pr}\big(\mathcal{T} = 0\big) = 0$. Now if the engineer believed that the pipeline was very leak-prone then setting $\mathsf{Pr}\big(\mathcal{T} = 0\big)$ close to zero might indeed be appropriate. But if the pipeline was thought to be tight, then we would be more inclined to treat data suggesting a small difference between $v_0$ and $v_L$ as indicative of measurement error. In other words, for a tight pipeline $\mathcal{T} = 0$ would have a probability close to 1. For our example we suppose that the pipeline is quite leak-prone, and set

$$\mathsf{Pr}\big(\mathcal{T} = i\big) = \begin{cases} 0.5 & i = 0 \text{ (no leak)} \\ 0.4 & i = 1 \text{ (ordinary leak)} \\ 0.1 & i = 2 \text{ (sabotage leak)} \end{cases} \tag{29}$$

although as we have restricted our modelling to three leak-types, we should re-name the 'ordinary' leak to the catch-all 'other unspecified leak'.

For the inference we can incorporate $\mathcal{T}$ among the uncertain parameters $\theta$. Then when we compute the predictive distribution of $\theta$ we are also computing the

probability distribution of $\mathcal{T}$ conditional on the velocity data. In other words, we can use the probabilistic approach to turn the data into an assessment of whether there has been a leak, and—on the condition that there has been a leak—what kind of leak it is.

We illustrate this with the example from section 5. In that example we had $\gamma = 0.0361$. We reconsider this leak within the more general framework, and also consider two other leak coefficients, one twice as large, and one four times as large. These three scenarios are denoted as $A$, $B$ ($2\times$) and $C$ ($4\times$). The velocity data from the three scenarios are

$$\begin{array}{lcc} & \bar{v}_0^1 & \bar{v}_L^1 \\ \text{Scenario } A & 1.66 & 1.59 \\ \text{Scenario } B & 1.69 & 1.55 \\ \text{Scenario } C & 1.73 & 1.47 \end{array} \tag{30}$$

Figure 3 shows the detection and attribution probabilities for each of the three scenarios, in the top lefthand panel, along with the HDR for leak location and leak coefficient in each case. The probabilities show that in scenario $A$ there remains a small probability that $\mathcal{T} = 0$, i.e. that no leak has occurred, and that the difference between measured upstream and downstream velocities is attributable to meter bias. In the second and third scenarios, for which the difference between the upstream and downstream velocity measurements is larger, the probability that $\mathcal{T} = 0$ is effectively zero. The other three panels show that as the leak type probability shifts from $\mathcal{T} = 1$ to $\mathcal{T} = 2$ so the predictive distribution of the leak location and coefficient also shifts. In scenario $C$ the probability strongly favours a sabotage leak, and we see that most of the leak location probability is contained

in the sabotage leak section of the pipeline. The more that different types of leak have different 'signatures' in location and size, the more sensitive and effective will be the leak attribution.

The approach outlined in this section can easily be extended to consider multiple leaks arising from the same or different causes.

## 7 Conclusion

The activity of leak detection in the presence of uncertainty is an example of a statistical 'calibrated prediction'. That is, we want to predict certain quantities such as the leak location and leak size, but we need at the same time to calibrate our model to imperfect data in order to reduce our uncertainty about quantities such as the pipe roughness and the valve settings. Within a fully-probabilistic setting we are able to do this in a way that ensures that our predictive uncertainty accounts for our residual calibration uncertainty. Other methods, in particular those based on a first-stage optimisation of the uncertain parameters in the model, are not able to provide an adequate measure of uncertainty. Insofar as the optimisation approach can be considered to be inference by maximum likelihood, the basic condition that justifies an uncertainty analysis using the profile likelihood is that there is a reasonable amount of independent data. This is often not the case with pipelines and pipe networks, where there are typically only as many data as there are meters, and often there are fewer meters than there are uncertain parameters.

This paper has outlined the fully-probabilistic treatment of the problem of leak detection in a single pipe, using the mass-imbalance approach. However, the treatment is perfectly general, and may be extended straightforwardly to

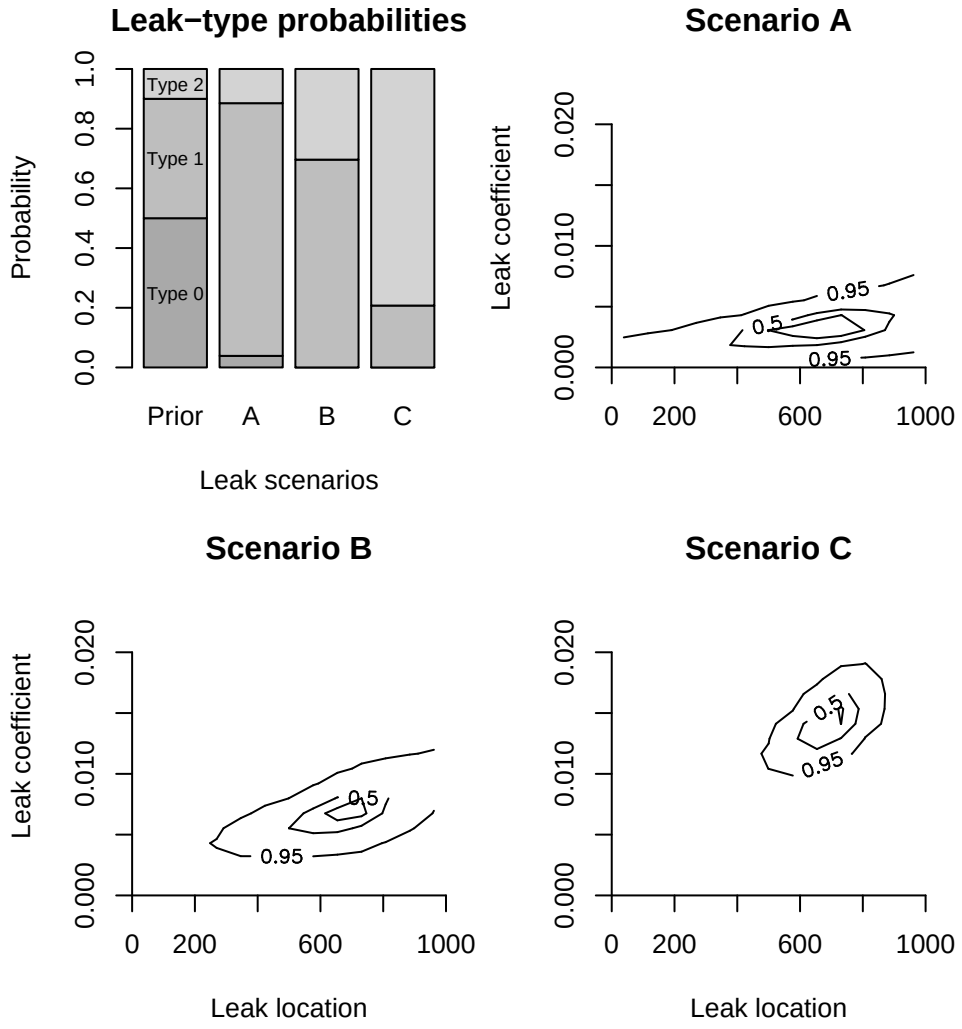


Figure 3: Leak detection and classification. The top lefthand panel shows the prior and predictive probabilities on the three types of leak (0: no leak; 1: ordinary leak; 2: sabotage leak). The three scenarios correspond to three different sets of velocity measurements, given in (30). The remaining panels show the high density regions for the predictive distribution of leak location and coefficient in the 'Two Meter' case, conditional on a leak having taken place, for each of the three scenarios.

models that cover pipe networks, providing that each evaluation of the model at a particular choice of the model parameters is quick. This is certainly the case for simulators of pipe networks. One of the problems that can arise with a probabilistic treatment is how to present the results. In this paper the predictive distribution describing the leak location and size has been computed by sampling. The results can then be presented in an intuitive manner as approximate high density regions. Summary statistics such as mean and variances can also be computed, but the graphical presentation is probably easier to interpret if a rapid response is required.

The paper also shows how we may generalise the problem of leak description, premised on a single 'typical' leak, to leak detection and leak attribution based on a number of scenarios describing possible events, and the types of leak that might follow. In this case the probabilistic approach allows us to assign a probability to each scenario and to update that probability using data from meters. A probabilistic description of our uncertainty about the pipeline feeds naturally into a decision-theoretic framework for establishing the most effective response.

## Notation

### Quantities and functions

$D$ = Pipe diameter (assumed constant)
$\epsilon_0$, $\epsilon_L$ = Velocity meter measurement errors (collectively $\epsilon$)
$f(\cdot)$ = Weisbach friction factor along the pipeline
$g$ = Gravitational acceleration
$\gamma$ = Leak coefficient
$H$, $h(\cdot)$ = Piezometric head at the upstream end of the pipeline, piezometric head along the pipeline
$K(\cdot)$ = Equivalent sand-roughness along the pipeline
$\kappa$ = Downstream valve coefficient
$L$ = Pipeline length
$\rho(\cdot)$ = Definite integral of $f(\cdot)$, see eq. (1b)
$\mathsf{Re}$ = Reynolds number
$S$ = Region of the pipeline in which a sabotage leak might occur
$\mathcal{T}$ = Leak type indicator
$\theta$ = Collection of parameter values
$v$ = Fluid velocity without a leak
$v$, $v_0$, $v_L$ = Fluid velocity, upstream and downstream fluid velocity (collectively $V$); a superscript 0 denotes prior to the leak, and 1 denotes post-leak
$x$, $z$, $d$ = Index variables for distance along pipeline

### Random field for the Weisbach friction factor

$\mu_1(\cdot)$, $\mu_2(\cdot,\cdot)$ = Mean function and covariance function of $f(\cdot)$
$\sigma^2$ = Variance of the stationary random field for $f(\cdot)$
$r(\cdot)$ = Correlation function of the stationary random field for $f(\cdot)$
$\tau$ = Decay rate of the Ornstein-Uhlenbeck stationary random field for $f(\cdot)$

### Probabilistic operators and related quantities

$\mathsf{Pr}\big(\cdot\big)$ = Probability density function
$\mathsf{E}\big(\cdot\big)$, $\mathsf{Cov}\big(\cdot,\cdot\big)$ = Expectation and covariance
$a$, $s$ = Shape and scale parameters for the gamma distribution
$c$ = Constant of proportionality
'$|$' = 'Conditional upon', in probability distribution functions and covariances

### Other symbols

'$\stackrel{\triangle}{=}$', '$\equiv$' = 'Defined as', 'Equivalent by definition'

Measured values of previously uncertain quantities are denoted with an over-

bar. For example, the measured value of the upstream velocity after the leak has occurred is $\bar{v}_0^1$.

## References


Bartlett, M.: 1978, *An Introduction to Stochastic Processes*. Cambridge UK: Cambridge University Press, 3rd edition.

Bernado, J. and A. Smith: 1994, *Bayesian Theory*. John Wiley & Sons.

Clemen, R. T.: 1996, *Making Hard Decisions*. Pacific Grove, CA: Duxbury Press, 2nd edition.

Colebrook, C. F.: 1939, 'Turbulent Flow in Pipes, with particular reference to the Transition Region between the Smooth and Rough Pipe Laws'. *Journal of the Institution of Civil Engineers* **11**, 133–156.

Craig, P., M. Goldstein, J. Rougier, and A. Seheult: 2001, 'Bayesian Forecasting for Complex Systems Using Computer Simulators'. *Journal of the American Statistical Association* **96**, 717–729.

Goldstein, M. and J. Rougier: 2004, 'Probabilistic formulations for transferring inferences from mathematical models to physical systems'. *SIAM Journal on Scientific Computing*. Forthcoming.

Haaland, S.: 1983, 'Simple and Explicit Formulas for the Friction Factor in Turbulent Pipe Flow'. *Journal of Fluids Engineering* **105**, 89–90.

Higdon, D., M. Kennedy, J. Cavendish, J. Cafeo, and R. D. Ryne: 2004, 'Combining Field Data and Computer Simulations for Calibration and Prediction'. *SIAM Journal on Scientific Computing*. Forthcoming.

Kapelan, Z. S., D. A. Savic, and G. A. Walters: 2003a, 'A Hybrid Inverse Transienty Model for Leakage Detection and Roughness Calibration in Pipe Networks'. *Journal of Hydraulic Research* **41**(5), 481–492.

Kapelan, Z. S., D. A. Savic, and G. A. Walters: 2003b, 'Multiobjective Sampling Design for Water Distribution Model Calibration'. *Journal of Water Resources Planning and Management* **129**, 466–479.

Kennedy, M. and A. O'Hagan: 2001, 'Bayesian calibration of computer models'. *Journal of the Royal Statistical Society, Series B* **63**, 425–464. With discussion.

Lindley, D.: 1985, *Making Decisions*. London: John Wiley & Sons, 2nd edition.

Liou, J.: 1994, 'Mass Imbalance Error of Waterhammer Equations and Leak Detection'. *AMSE Journal of Fluids Engineering* **116**, 103–9.

Massey, B.: 1989, *Mechanics of Fluids*. London: Chapman and Hall, sixth edition.

Mays, L.: 1989, *Reliability Analysis of Water Distribution Systems*. New York: ASCE Press.

Mears, M.: 1993, 'Real World Applications of Pipeline Leak Detection'. In: *Proceedings of the International Conference on Pipeline Infrastructure*. pp. 189–209.

Mukherjee, J. and S. Narasimhan: 1996, 'Leak Detection in Networks of Pipelines by the Generalized Likelihood Ratio Method'. *Ind. Eng. Chem. Res.* **35**, 1886–1893.

Parzen, E.: 1962, *Stochastic Processes*. San Francisco: Holden-Day, Inc.

Pawitan, Y.: 2001, *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Clarendon Press.

Pesta, T. and A. Cassley: 1992, 'Leak Detection—A Regulatory Perspective'. In: *Proceedings of the Offshore Mechanics and Arctic Engineering Conference*. pp. 493–500.

Purdar, R. S. and J. A. Liggett: 1992, 'Leaks in Pipe Networks'. *Journal of Hydraulic Engineering* **118**, 1031–1046.

R Development Core Team: 2004, 'R: A language and environment for statistical computing'. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, `http://www.R-project.org`.

Rajtar, J. and R. Muthiah: 1997, 'Pipeline Leak Detection System for Oil and Gas Flowlines'. *Journal of Manufacturing Science and Engineering* **119**, 105–9.

Ripley, B.: 1987, *Stochastic Simulation*. New York: John Wiley & Sons.

Robert, C. and G. Casella: 1999, *Monte Carlo Statistical Methods*. New York: Springer.

Rougier, J.: 2004, 'Prediction of Future Climate Using an Ensemble of Computer Simulator Evaluations'. Available at `http://www.maths.dur.ac.uk/stats/people/jcr/EnsemblesA4.pdf`.

Rougier, J. and M. Goldstein: 2001, 'A Bayesian analysis of fluid fow in pipelines'. *Applied Statistics* **50**(1), 77–93.

Smith, A. and A. Gelfand: 1992, 'Bayesian Statistics Without Tears: A Sampling-Resampling Perspective'. *The American Statistician* **46**, 84–88.

Tanner, M.: 1993, *Tools for Statistical Inference : Methods for the Exploration of Posterior Distributions and Likelihood Functions*. New York: Springer-Verlag, 2nd edition.

Wang, G., D. Dong, and C. Fang: 1993, 'Leak Detection for Transport Pipelines Based on Autoregressive Modeling'. *IEEE Transactions on Instrumentation and Measurement* **42**, 68–71.

Wylie, E. and V. Streeter: 1993, *Fluid Transients in Systems*. New Jersey: Prentice-Hall, Inc.