Transactions on Machine Learning
and Data Mining
Vol. 8, No. 2 (2015) 77-100
ISSN 1864-9734
ISBN: 879-3-942952-39-2

ibai Publishing
www.ibai-publishing.org

Traceable Measurements using Sensor Networks

Alistair B. Forbes

National Physical Laboratory, UK
alistair.forbes@npl.co.uk

Abstract. National Metrology Institutes such as the National Physical Laboratory provide the scientific and technical framework to ensure measurements are traceability to standard units or references. Many of the procedures developed for ensuring traceability involve measurements in laboratory conditions, whereby environmental factors such as temperature are closely controlled, and using instruments whose responses are well-characterised by validated models. However, many societal challenges relating climate and environment, energy and sustainability, health and well-being necessarily involve measurements outside laboratory conditions. In such circumstances, the mathematical and statistical modelling tools need to be strengthened so that concepts of traceability, uncertainty evaluation and calibration can also be applied outside the laboratory. In particular, we need to develop tools that account for uncertainties associated with the models of the system, for we can no longer be certain that we understand all aspects of the response of a potentially complex system. One such challenge is the measurement of air quality. Air quality has a significant impact on quality of life and many regulations now apply to controlling pollution. In order to ensure compliance to regulations, it is necessary to monitor air quality. However, measurement of air quality can only be made at a finite number of sites while the regulations apply to the complete air quality field. In this paper, we consider approaches for accounting for model uncertainty and the use of Gaussian processes to model the temporal and spatial and correlation in order to estimate the air quality field and its associated uncertainty.

Keywords: Traceability, Sensor Networks, Uncertainty, Gaussian Processes

1 Introduction

A primary role of a National Metrology Institute is to ensure that measured data is traceable to standard units. One aspect of establishing traceability is valid uncertainty

evaluation. Standard methodologies such as The Guide to the Expression of Uncertainty in Measurement (the GUM [2]) were developed to provide a probabilistic basis for a coherent and consistent approach to evaluating uncertainty, now adopted worldwide.The current metrology paradigm of standards, calibration and traceability, however, is designed for the measurement of a single quantity using a single, dedicated measuring instrument or system, e.g., measuring the length of an artefact using a laser interferometer. Many societal challenges relating climate and environment, energy and sustainability, health and well-being necessarily involve measurements outside laboratory conditions. In such circumstances, the mathematical and statistical modelling tools need to be strengthened so that concepts of traceability, uncertainty evaluation and calibration can also be applied outside the laboratory.

For example, much of environmental monitoring involves networks of sensors measuring a number of different quantities at several locations and at different times. Whether the characteristic being measured is an air pollutant level, acoustic noise (associated with an airport, for example), or sea water salinity, etc., measurements at particular spatial and time locations are used to make inferences at other individual spatial and time locations or are aggregated to make inferences over a region or time period. The quality of the inferences made will depend on how well the network is designed and how the sensor information is used. Currently, the impact of sensor network data is severely limited by the lack of a methodology for calibration, traceability and uncertainty evaluation applicable to sensor networks. The measurement community needs to develop a much more comprehensive approach to uncertainty quantification, in which uncertainty contributions associated with models and computation are also taken into account.

In section 2, we provide an overview of the concepts of traceable measurement and measurement uncertainty while in section 3 we describe how model selection and model averaging can be used to take account uncertainties associated with models. In section 4, we consider uncertainty evaluation associated with models having spatio-temporal correlation and consider opportunities to calibrate sensor networks using the spatio-temporal correlation. Our concluding remarks are given in section 5.

2 Traceable measurement

Metrology is the science of measurement. Its central aim is to ensure that stated measured values have an unambiguous interpretation. This is achieved by defining standard units and providing procedures that enable a quantity being measured, e.g., the mass of an artefact, to be compared with the appropriate standard unit in a traceable way. “Metrological traceability” is defined in the International Vocabulary of Metrology (VIM [3]) as “the property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty”. In the same document, “measurement uncertainty” is defined as a “non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used”. This definition allows considerable scope for interpretation. Since the publication of the Guide to the Expression of Uncertainty in Measurement (GUM [2]) in the mid 1990’s, measurement uncertainty is defined in terms of probability distributions. The result of a measurement

is a probability distribution $p(a)$ associated with the quantity being measured A, the measurand.

The best estimate of the measurand is taken to be the mean of the probability distribution and the standard uncertainty associated with the measurand (or sometimes said to be associated with the estimate of the measurand) is taken to be its standard deviation, assuming that both the mean and standard deviation exist. Often there is an assumption that the probability distribution associated with the measurand is a Gaussian or at least is approximated well by a Gaussian. The assigned probability distribution allows inferences about the 'true value' a^* of the quantity to be made, e.g., the probability of that the true value lies in the interval $[L,U]$ is estimated by

$$\Pr(L \leq a^* \leq U) = \int_L^U p(a)\mathrm{d}a.$$

The probability distribution is referred to as a 'state of knowledge' distribution. It is usually derived in a deterministic way from data $\mathbf{y}$ and hypothesized model and assumptions which we denote collectively by $\mathscr{H}$. If required, we denote the state of knowledge distribution as $p(a|\mathbf{y},\mathscr{H})$ to reflect the dependence on data and assumptions.

2.1 Traceability chain

A traceability chain is perhaps most easily described in terms of a sequence of Bayesian updates. We associate to the standard unit a_0 the Dirac δ distribution at 1, since by definition the unit has no uncertainty. We perform a comparison of artefact A_1 with the standard unit gathering measurement data y_1 with likelihood $p(y_1|a_1,a_0)$, e.g.,

$$y \sim \mathrm{N}(a_1 - a_0|\sigma_1).$$

Assigning a (usually noninformative) prior $p(a_1)$ for a_1, we determine the posterior distribution

$$p(a_1,a_0|y_1) \propto p(y_1|a_1,a_0)p(a_1)p(a_0),$$

which is marginalised to determine $p(a_1|y_1)$. At the kth stage we record data y_k with associated likelihood $p(y_k|,a_k,a_{k-1})$, leading to joint posterior distribution

$$p(a_k,a_{k-1}|y_k,\mathbf{y}_{k-1}) \propto p(y_k|a_k,a_{k-1})p(a_k)p(a_{k-1}|\mathbf{y}_{k-1}),$$

and associated marginalised distribution

$$p(a_k,|\mathbf{y}_k) = \int_{\mathscr{R}} p(a_k,a_{k-1}|y_k,\mathbf{y}_{k-1})\mathrm{d}a_{k-1}.$$

The chain is traceable if the likelihood $p(y_k|a_k,a_{k-1})$ is assigned appropriately at each stage. If the likelihood is of the form

$$y_k|a_k,a_{k-1} \sim \mathrm{N}(a_k - a_{k-1},\sigma_k^2), \quad k = 1,\ldots,n,$$

corresponding to a simple comparison of artefacts subject to Gaussian noise, then the posterior distribution $p(\mathbf{a}_n|\mathbf{y}_n)$ is the multivariate Gaussian distribution $\mathrm{N}(\hat{\mathbf{a}},V)$ where

$\hat{\mathbf{a}}$ solves the linear set of equations $C\mathbf{a} = \mathbf{z}$, involving the bi-diagonal observation matrix C and data vector $\mathbf{z}$ with

$$C = \begin{bmatrix} w_1 & & & & \\ -w_2 & w_2 & & & \\ & -w_3 & w_3 & & \\ & & \ddots & & \\ & & & -w_n & w_n \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} w_1(a_0 + y_1) \\ w_2 y_2 \\ \vdots \\ w_n y_n \end{bmatrix}, \quad w_k = 1\sigma_k.$$

The variance matrix V associated the estimate $\hat{\mathbf{a}}$ is given by

$$V = (C^{\mathrm{T}}C)^{-1}$$

If $\sigma_k = \sigma$, then the uncertainty associated with the kth artefact is $k^{1/2}\sigma$ and the correlation coefficient associated with artefacts k and $j > k$ is $c_{kj} = (k/j)^{1/2}$. Thus, the best estimates of pairs of quantities that have a common artefact in their traceability chains will be statistically correlated. In practice, calibrations further down the traceability chain are less accurate than those higher up so the strength of the correlations will be much less than for the case $\sigma_k = \sigma$.

In the case of mass measurement, the standard unit of mass is (currently) defined by a physical artefact, known as the International Prototype Kilogram (IKP), kept at the Bureau International des Poid et Mesures (BIBM) in Sevres, near Paris. The BIPM has a number of copies of the IPK made of the same material, a 90-10 % alloy of platinum and iridium, and over 30 countries have national copies. The UK has copy no. 18. In theory and, to a large extent, in practice, all mass measurements in the UK are traceable to the calibration of copy no. 18 against the IKP, and therefore, practically all estimates of mass in the UK are statistically correlated.

2.2 Inter-laboratory comparisons (ILCs)

Traceability depends on the reliable assignment of uncertainties associated with the various comparisons that underpin the traceability chain. These uncertainties are assigned on the basis of best practice such as that defined in the GUM and that associated with the scientific domain. However, there is no guarantee that this best practice are based on models that provide an exact characterisation of the physical systems involved. The inter-laboratory comparison is an important tool in demonstrating the validity of the methodologies. Typically, such an inter-comparison involves the measurement of one or more artefacts by a number of laboratories and the results analysed to determine if the spread of estimates is consistent with the stated uncertainties.

3 Model selection and model averaging

In many data-fitting and modelling problems, we often face the challenge to find the best model to best-fit a given data set [5, 20]. This choice is complicated by the fact there may be different and competing models. Loosely speaking, model selection is a process of choosing one appropriate model among many possible models $\mathscr{M}_1, \mathscr{M}_2, ..., \mathscr{M}_K$, with

the aim of balancing goodness-of-fit with minimising complexity of the model. For example, $\mathscr{M}_k$ could represent the space of polynomials of degree at most k and, in general, we would want to choose the polynomial of minimal degree that fits the data reasonably well.

3.1 Model selection according to information criteria

We consider a standard model in which data is generated according to the model

$$y = \phi(\mathbf{x}, \mathbf{a}) + \varepsilon, \quad \varepsilon \in \mathrm{N}(\mathbf{0}, \sigma^2),$$

where the function $\phi(\mathbf{x}, \mathbf{a})$, depending on parameters $\mathbf{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$, models the response of a system. For Gaussian noise, the least squares estimate $\hat{\mathbf{a}}$ of the parameters $\mathbf{a}$ minimises

$$F(\mathbf{a}) = \sum_{i=1}^{m} (y_i - \phi(\mathbf{x_i}, \mathbf{a}))^2,$$

and corresponds to the maximum likelihood estimate, i.e., maximises

$$p(\mathbf{y}|\mathbf{a}) \propto \exp\left\{ -\frac{F(\mathbf{a})}{2\sigma^2} \right\}.$$

Let $\mathrm{RSS} = F(\hat{\mathbf{a}})$.

Suppose there are in fact K competing models defined by functions $\phi_k(\mathbf{x}, \mathbf{a_k})$ involving parameter vectors $\mathbf{a}_k$ of length n_k. For a given data set $\{(\mathbf{x_i}, \mathbf{y_i})\}_{\mathbf{i=1}}^{\mathbf{m}}$, for each model space we can calculate the least squares best estimate $\hat{\mathbf{a}}_k$ of the model parameters and $\mathrm{RSS}_k = F(\hat{\mathbf{a}}_k)$, the residual sum of squares at the least squares estimate. The issue is to select the model that provides an adequate fit to the data, as measured by RSS, while at the same time is not overly complex, as measured by n_k, the number of parameters, for the number of observations m.

There are a number of criteria that are commonly used for this. For example, root mean square residual RMS given by

$$\mathrm{RMS}_k = \sqrt{\frac{\mathrm{RSS}_k}{m - n_k}},$$

is often used. Here, we use a related index $m \log \mathrm{RMS}$ which we write as

$$m \log \mathrm{RMS} = m \log(\mathrm{RSS}_k/m) + m \log\left(\frac{m}{m - n_k} \right). \tag{1}$$

Other criteria used are the Akaike information criterion [1], for this case given by

$$\mathrm{AIC} = m \log(\mathrm{RSS}_k/m) + 2n_k,$$

often with a correction for small number of degrees of freedom [16],

$$\mathrm{AICc} = m \ln(\mathrm{RSS}_k/m) + 2n_k \frac{m}{(m - n_k - 1)},$$

or the Bayes Information Criterion [27].

$$\mathrm{BIC} = m\ln(\mathrm{RSS}_k/m) + n_k \ln m.$$

Written in this way, all the criteria above have the same term representing the goodness of fit but different terms penalising the complexity of the model. The model selected is the one that minimises the criterion value.

3.2 Bayesian Model Averaging

For a given set of data D, the data can arise from one of many possible models $\mathscr{M}_1, \mathscr{M}_2, ..., \mathscr{M}_K$. Loosely speaking, model averaging is the process of estimating some quantity a of interest under each model $\mathscr{M}_i$, and subsequently averaging the estimates according to how likely each model is [15, 25]. The Bayesian approach is to try to calculate the posterior probability $P(\mathscr{M}_i|D)$ for each model, and then $w_i = P(\mathscr{M}_i|D)$ is used as the weights in the model averaging. The aim of the model selection is to select the model with the maximum $P(\mathscr{M}_i|D)$.

For Bayesian model averaging , the posterior distribution of a is

$$P(a|D) = \sum_{i=1}^{K} P(a|\mathscr{M}_i, D)P(\mathscr{M}_i|D), \tag{2}$$

where the posterior probability of model $\mathscr{M}_i$ is given by

$$P(\mathscr{M}_i|D) = \frac{P(D|\mathscr{M}_i)P(\mathscr{M}_i)}{\sum_{j=1}^{K} P(D|\mathscr{M}_j)P(\mathscr{M}_j)}. \tag{3}$$

3.3 Application to inter-laboratory comparisons

As discussed in section 2.2, the inter-laboratory comparison is a major tool in validating the methodologies used to derive measurement results and their associated uncertainties. In addition to this validation role, the ILC can be used more diagnostically to determine systematic laboratory effects similar to that performed in an analysis of variance [6, 13, 17] or to determine a consensus or reference value for a quantity on the basis of a number of different experiments [7]. In any of these applications, it is necessary (or at least extremely advisable) to assess the self-consistency of data (values and associated uncertainties) input into the ILC.

Consider the standard model (after weights have been applied)

$$y_i|a \in \mathrm{N}(a, 1), \qquad i = 1, \ldots, n. \tag{4}$$

Set $\bar{y} = (1/n)\sum_{i=1}^{m} y_i$. If the prior distribution for a is noninformative, $p(a) \propto 1$, then the posterior distribution for a is such that

$$a|\mathbf{y} \sim \mathrm{N}(\bar{y}, 1/n). \tag{5}$$

We note that $\bar{y}$ is the least squares estimate $\hat{a}$ of a. The model (4) also predicts that the residual sum of squares

$$F = \sum_{i=1}^{n} (y_i - \hat{a})^2$$

associated with the least squares fit is drawn for a χ^2_ν distribution with $\nu = n-1$ degrees of freedom. So, for example [7], if $\Pr(\chi^2_\nu > F) \geq 0.05$, then the input data is judged to be consistent with the model and that it safe to make inferences about a on the basis of the posterior distribution (5). In particular the best estimate or reference value for a is $\hat{a} = \bar{y}$, and the associated uncertainty is $u(a) = n^{-1/2}$.

If the χ^2 test fails, then what is to be done? There are many papers that consider approaches to adjusting the input uncertainties in order to bring about consistency. There a number of one parameter adjustment models [8, 30, 31]. Suppose $\mathbf{y} \sim \mathrm{N}(a\mathbf{e}, V_0)$, where $\mathbf{e} = (1,\ldots,1)^{\mathrm{T}}$. The idea is to replace V_0 with a variance matrix $V(\tau)$ depending on a single parameter τ. For the model $\mathbf{y} \sim \mathrm{N}(a\mathbf{e}, V(\tau))$, the least squares estimated is given by

$$\hat{a}(\tau) = \frac{\mathbf{e}^{\mathrm{T}} V(\tau)^{-1} \mathbf{y}}{\mathbf{e}^{\mathrm{T}} V(\tau)^{-1} \mathbf{e}},$$

the observed χ^2 value is given by

$$F(\tau) = (\mathbf{y} - \hat{a}(\tau)\mathbf{e})^{\mathrm{T}} V(\tau)^{-1} (\mathbf{y} - \hat{a}(\tau)\mathbf{e}).$$

The adjustment is made by choosing τ so that $F(\tau) = n-1$, i.e., is chosen so that the observed χ^2 value is the same its expected value. The simplest approach is to scale all the input uncertainties by $(1+\tau)$, i.e., $V(\tau) = (1+\tau)V_0$. This approach is sometimes referred to in the metrology field as the Birge adjustment procedure [4] after Birge who used it in the analysis of data associated with the fundamental constants. A second approach used in metrology is to set $V(\tau) = V_0 + \tau I$, sometimes referred to as the Mandel-Paule method [24]. Bayesian approaches have also been considered [8, 18, 21, 22, 28, 29].

The approach described by Cox in [9] to the analysis of inconsistent ILC data is akin to a model selection approach. If the complete set of data is inconsistent then participants are removed from the exercise until a consistent subset is determined. The algorithmic approach efficiently determines a subset of the participants which is self-consistent according to the χ^2 criterion and no other subset with the same or greater number of participants has a smaller observed χ^2 value. (Exceptionally, this subset might not be unique.) Each subset of the participants can be thought of defining a model in which the uncertainties provided by the selected participants is regarded as reliable and those associated with the excluded participants are not. The largest consistent subset (LCS) defines the selected model.

Here we describe a model averaging approach to the analysis of inconsistent ILC data; see also [11]. We assume that there is a prior possibility that one or more participants have underestimated their uncertainty by a factor of three, say, and that the fraction of such participants follows a binomial distribution defined by parameter $0 < \lambda < 1$. The hyper-parameter λ is assigned a prior Beta distribution $B(\alpha, \beta)$. Thus, the prior

expected value of λ is $\alpha/(\alpha+\beta)$ and prior variance associated with λ is

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

We assume a prior that all participants have the same probability of supplying unreliable uncertainty estimates. Let $\mathscr{M}_0$ be the model in which all reported uncertainties are reliable, $\mathscr{M}_i$ than in which participant i only has an unreliable uncertainty estimate, $\mathscr{M}_{ij}$, where participants i and j are unreliable, etc., leading to 2^n models in all, labelled by an index q, say.

The assignment of a prior to λ assigns a joint prior distribution $p(q,\lambda) = p(q|\lambda)p(\lambda)$. The likelihood $p(\mathbf{y}|a,q,\lambda)$ is easy to calculate, so that

$$p(a,q,\lambda|\mathbf{y}) \propto p(\mathbf{y}|a,q,\lambda)p(q|\lambda)p(\lambda)p(a).$$

The posterior distributions are given by marginalisation, e.g.,

$$p(a|\mathbf{y}) = \sum_{q=1}^{2^n} \left\{ \int_0^1 p(\mathbf{y}|a,q,\lambda)p(q|\lambda)p(\lambda)p(a)\mathrm{d}\lambda \right\},$$

and

$$p(q|\mathbf{y}) = \int_{-\infty}^{\infty}\int_0^1 p(\mathbf{y}|a,q,\lambda)p(q|\lambda)p(\lambda)p(a)\mathrm{d}\lambda\,\mathrm{d}a.$$

We illustrate the behaviour of these approaches on three simulated data sets $\mathbf{y}_k$, $k = 1,2,3$. Figure 1 shows simulated data $\mathbf{y}_1$ involving 10 participants. The uncertainty bars represent $\pm$ two standard deviations. The result from participant 6 seems outlying and is deemed so according to the χ^2 test. The largest consisted subset [9] is determined by the remaining nine participants. Figure 4 shows the posterior distributions $p(a|\mathbf{y})$ for various adjustment procedures: 'mixture' denotes the model averaging approach, 'input' denotes the case where all input uncertainty estimates are regarded as reliable, 'Birge' the Birge adjustment procedure and 'LCS(9)' the distribution associated with the largest consistent subset. Figure 7 shows the prior and posterior distributions for λ determined using the model averaging approach.

Table 1 shows the posterior probabilities $p(q|\mathbf{y}_1)$ for the most likely models indexed by q. All other models are associated with the probabilities less than 0.01. The model best supported by the data by far is the model in which participant 6 alone is regarded as unreliable. This accounts for the moderately good agreement between the model averaging approach and the LCS approach. Table 2 shows the prior and posterior probabilities of k participants being considered unreliable.

Figure 2 shows a second set of simulated data $\mathbf{y}_2$ involving 10 participants. The data is similar to that in figure 1, only that both participants 6 and 10 seem potentially outlying. The largest consistent subset is judged to have nine participants with participant 6 excluded as before. The observed χ^2 value is 15.0 compared with a test value of 16.9. If participant 10 is excluded instead, the corresponding χ^2 value is 18.9. From this point of view there is a case for participant 6 or 10 or both participants 6 and 10 being excluded. Figure 5 shows the posterior distributions $p(a|\mathbf{y}_2)$ for various adjustment procedures as

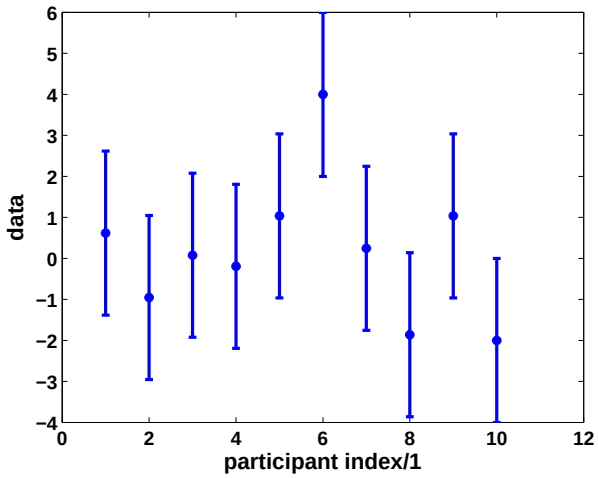

Fig. 1. Simulated data $\mathbf{y}_1$ involving 10 ILC participants. The result from participant 6 seems outlying.

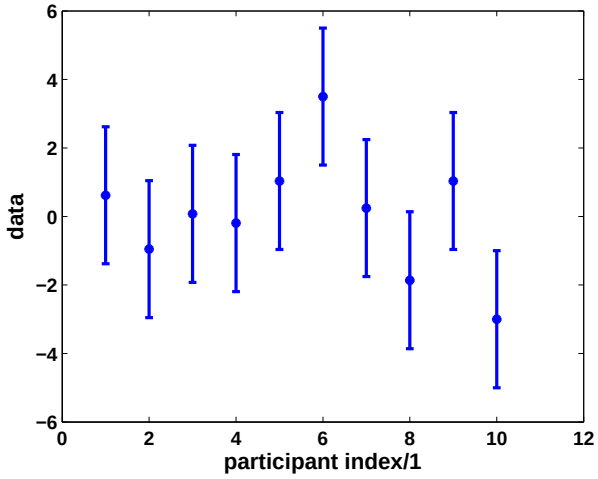

Fig. 2. Simulated data $\mathbf{y}_2$ involving 10 ILC participants. The result from participant 6 and 10 are potentially outlying.

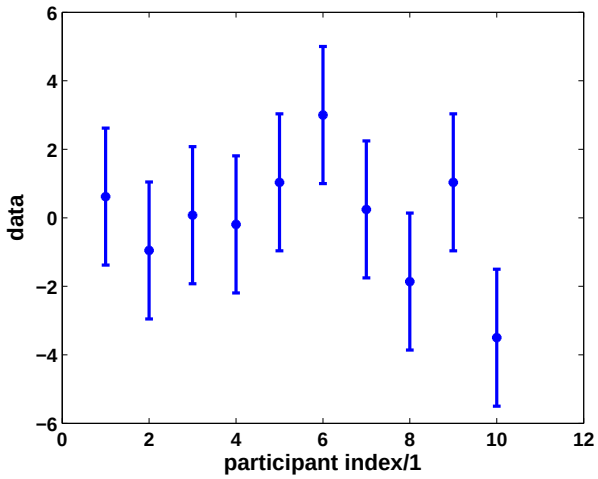

Fig. 3. Simulated data $\mathbf{y}_3$ involving 10 ILC participants. The result from participants 10 and 6 are potentially outlying.

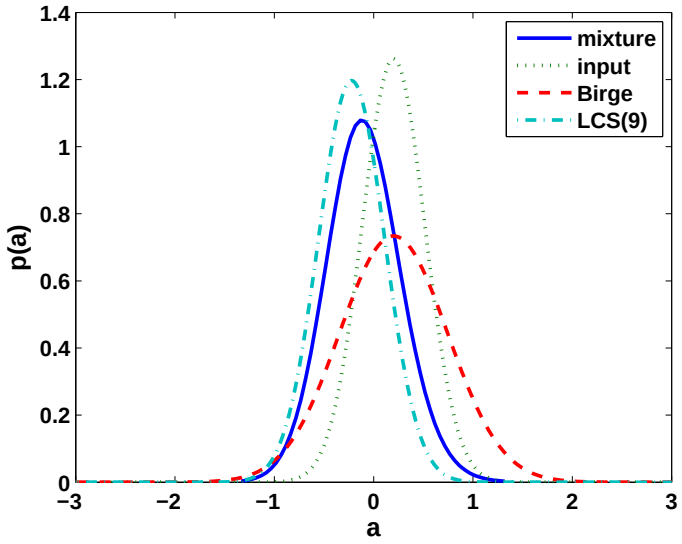

Fig. 4. Posterior distributions $p(a|\mathbf{y}_1)$ for a according to different adjustment procedures determined from the data in figure 1. The label 'mixture' denotes the model averaging approach, 'input' denotes the case where all input uncertainty estimates are regarded as reliable, 'Birge' the Birge adjustment procedure and 'LCS(9)' the distribution associated with the largest consistent subset.

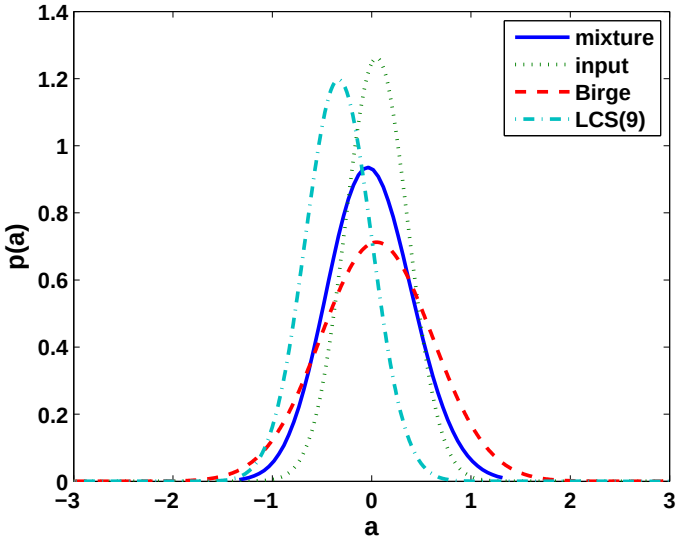

Fig. 5. Posterior distributions $p(a|\mathbf{y}_2)$ for a as in figure 4 but for data in figure 2.

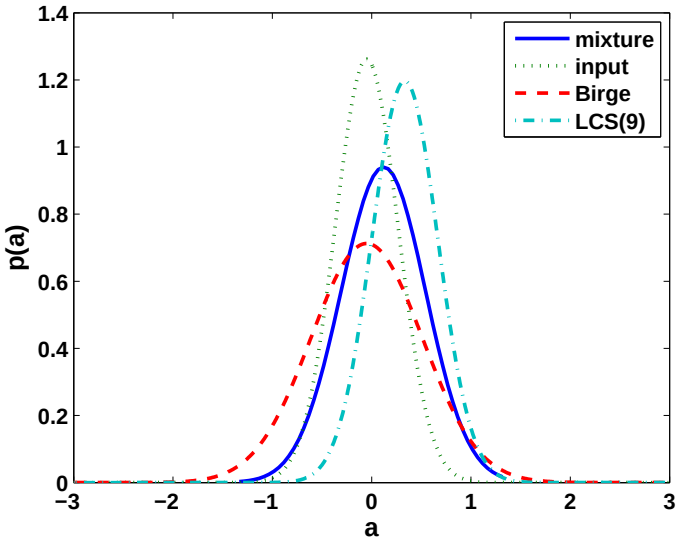

Fig. 6. Posterior distributions $p(a|\mathbf{y}_3)$ for a as in figure 4 but for data in figure 3.

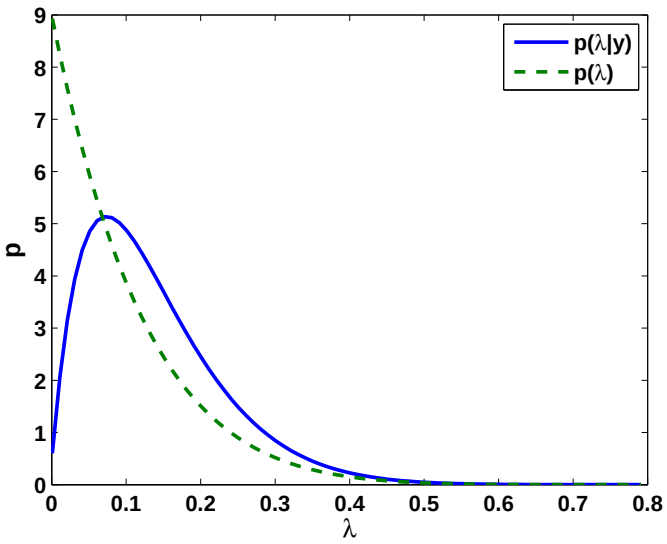

Fig. 7. Posterior and prior distributions for λ for a determined from the data $\mathbf{y}_1$ in figure 1 .

in figure 4 but for the data $\mathbf{y}_2$. The LCS approach in this case is significantly different from the model averaging approach. The LCS posterior distribution is constructed on the basis the selected model, participant 6 is unreliable, is certain, while the model averaging approach takes into account other possibilities. Table 1 shows the posterior probabilities $p(q|\mathbf{y}_2)$ for the most likely models indexed by q for data $\mathbf{y}_2$. It is seen that the three models most supported by the data are those that relate to 6, 10, or both being assessed as unreliable. Table 2 shows the posterior probabilities of k participants being considered unreliable for this dataset.

Figure 3 shows simulated data $\mathbf{y}_3$ similar to data $\mathbf{y}_2$ in that both participants 6 and 10 seem potentially outlying but in this case the largest consistent subset is judged to have nine participants with participant 10 excluded. Figure 6 shows the posterior distributions $p(a|\mathbf{y}_3)$ for various adjustment procedures as in figure 4 but for the data $\mathbf{y}_3$. Again, the LCS approach is significantly different from the model averaging approach. Comparing figures 5 and 6, we see that the LCS approach give rise to significantly different distributions for the two data sets $\mathbf{y}_2$ and $\mathbf{y}_3$, although the two data sets are similar. This is because of the discrete nature of the model selection process. The model averaging approach gives a smooth response to the changes in the data sets. Table 1 shows the posterior probabilities $p(q|\mathbf{y}_3)$ for the most likely models indexed by q for data $\mathbf{y}_2$. The results are similar to those for data $\mathbf{y}_2$ but with the roles of participants 6 and 10 interchanged. Table 2 shows the posterior probabilities of k participants being considered unreliable for this dataset.

$p(q\|\mathbf{y}_1)/0.01$	1	2	3	4	5	6	7	8	9	10
53	0	0	0	0	0	1	0	0	0	0
11	0	0	0	0	0	1	0	0	0	1
8	0	0	0	0	0	1	0	1	0	0
4	0	0	0	0	0	1	0	0	1	0
4	0	0	0	0	1	1	0	0	0	0
4	0	0	0	0	0	1	0	1	0	1
3	0	1	0	0	0	1	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
$p(q\|\mathbf{y}_2)/0.01$	1	2	3	4	5	6	7	8	9	10
31	0	0	0	0	0	1	0	0	0	1
21	0	0	0	0	0	1	0	0	0	0
10	0	0	0	0	0	1	0	1	0	1
6	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	1	0	0	1	1
3	0	0	0	0	1	1	0	0	0	1
3	0	0	0	0	0	1	0	1	0	0
$p(q\|\mathbf{y}_3)/0.01$	1	2	3	4	5	6	7	8	9	10
30	0	0	0	0	0	1	0	0	0	1
21	0	0	0	0	0	0	0	0	0	1
9	0	0	0	0	0	1	0	1	0	1
8	0	0	0	0	0	0	0	0	1	0
6	0	0	0	0	0	1	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	1	0	0	1	1
3	0	0	0	0	1	1	0	0	0	1

Table 1. Posterior percentage probabilities associated with the most likely models determined from the data $\mathbf{y}_k$, $k = 1,2,3$, in figures 1–3. A '1' in a column indicates the corresponding participant's uncertainty statement is considered unreliable.

k	$p(k)$	$p(k\|\mathbf{y}_1)$	$p(k\|\mathbf{y}_2)$	$p(k\|\mathbf{y}_3)$
0	39	3	4	3
1	39	54	28	27
2	17	31	42	43
3	5	10	21	21
4	1	2	5	5
6	0	0	1	0
7	0	0	0	0

Table 2. Percentage probabilities of observing k participants judged to be unreliable. The second column is the prior assignment, the remaining three columns give the posterior assignments determined from the data $\mathbf{y}_1$, $\mathbf{y}_2$ and $\mathbf{y}_3$, in figures 1–3, respectively.

4 Sensor networks and Gaussian processes

Model selection and model averaging go some way to accounting for the fact that the model of the underlying physical system may be only partially known. However, both still rely on defining a set of models to select from or to average over. If the selected set of models, e.g., polynomials, splines, etc., does not match well the actual behaviour of the physical system, then inferences based on the selected set of models could well be unreliable. Often, some of the aspects of the model are well understood and captured in a physical model but that there other systematic effects that are present that are not well understood but are expected to vary smoothly, i.e., the response is correlated with the stimulus variables [19, 12]. Often the response is correlated over space and/or over time. The idea of a Gaussian process model is to model the correlation behaviour in a flexible way and to let the measurement data define the actual form of the response. Often, the data is gathered by a network of sensors distributed spatially or temporally.

Sensor networks represent a new measurement paradigm with applications across environmental monitoring, earth observation, structural health monitoring, etc. The paradigm involves multiple sensors, acting collaboratively through wireless communications and internet services, to provide raw data that is converted to information-based products, e.g., a map of air quality in a region. To convert these information products into metrology products, it is necessary to provide an acceptable calibration, traceability and uncertainty framework. Gaussian processes present a modelling approach that can address this requirement. We first give a summary of the basic approach, sometimes known as universal Kriging.

4.1 Universal Kriging

Universal Kriging can be developed in a classical estimation or Bayesian framework; see e.g., [10, 26]. Suppose

$$\begin{bmatrix}\eta\\ \zeta\end{bmatrix} | \alpha \sim \mathrm{N}\left(\begin{bmatrix}C\\ D\end{bmatrix}\alpha, V\right),$$

with

$$V = \begin{bmatrix}V_{11} & V_{21}^{\mathrm{T}}\\ V_{21} & V_{22}\end{bmatrix} = LL^{\mathrm{T}}, \quad L = \begin{bmatrix}L_{11} & \\ L_{21} & L_{22}\end{bmatrix},$$

so that $L_{21} = V_{21}L_{11}^{-\mathrm{T}}$. We also write this model as

$$\eta|\alpha = C\alpha + L_{11}\varepsilon_1, \quad \zeta|\alpha = D\alpha + L_{21}\varepsilon_1 + L_{22}\varepsilon_2 \quad \varepsilon_1, \varepsilon_2 \sim \mathrm{N}(\mathbf{0}, I).$$

In an example application, η represents the response of the system at a set of locations that depends on model parameters α but also on systematic effects ε_1. Similarly ζ represents the response at a different set of locations that depends on α, ε_1 and ε_2. The common dependence on ε_1 characterises the spatial correlation between the two sets of locations. Suppose observations $\mathbf{y}$ of η are made. What can be said about α and ζ on the basis on the information supplied by $\mathbf{y}$?

The best estimate $\mathbf{a}$ of α is given the solution of the Gauss-Markov problem, namely

$$\mathbf{a} = \left(C^{\mathrm{T}} V_{11}^{-1} C\right)^{-1} C^{\mathrm{T}} V_{11}^{-1} \mathbf{y}, \quad V_{\mathbf{a}} = \left(C^{\mathrm{T}} V_{11}^{-1} C\right)^{-1}.$$

In addition, the best estimate of ε_1 is $\mathbf{e} = L_{11}^{-1}(\mathbf{y} - C\mathbf{a})$ and the best estimate of ε_2 is $\mathbf{e}_2 = \mathbf{0}$. Therefore, the best estimate $\mathbf{z}$ of ζ is $\mathbf{z} = D\mathbf{a} + L_{21}\mathbf{e}$. From a Bayesian point of view, assuming noninformative priors for α, etc.,

$$\zeta | \mathbf{y} \sim \mathrm{N}(\mathbf{z}, V_{\mathbf{z}}),$$

where terms $\mathbf{z}$ and $V_{\mathbf{z}}$ can be evaluated as

$$\mathbf{z} = V_{21} V_{11}^{-1} \mathbf{y} + E\mathbf{a}, \quad V_{\mathbf{z}} = V_{22} - V_{21} V_{11}^{-1} V_{21}^{\mathrm{T}} + E V_{\mathbf{a}} E^{\mathrm{T}}.$$

with

$$E = D - V_{21} V_{11}^{-1} C.$$

The Cholesky factor [14] of $V_{11} = L_{11} L_{11}^{\mathrm{T}}$ can be used to evaluate these expressions. For spatial applications, for example, the observed responses at one set of locations allows us to estimate the responses at other locations.

4.2 Spatio-temporal correlation models

The principle of the Gaussian process approach used here is that the readings from a group of sensors reflect a signal that is correlated spatially, temporally or both. The general formulation is as follows. Let $\mathbf{y} = (y_1 \ldots, y_m)^{\mathrm{T}}$ be a set of measured values. Associated to each measured value y_i are spatio-temporal coordinates $(\mathbf{x_i}, \mathbf{t_i})$ representing the spatial location $\mathbf{x_i}$ of the sensor that produced the measured value and the time t_i the measurement was taken. We model the system that gave rise to these data as $\mathbf{y} = C\mathbf{a} + \mathbf{e}$ where $\mathbf{a}$ are parameters specifying the systematic behaviour according to a known and validated model and $\mathbf{e}$ represents random effects which we assume are drawn from a multivariate Gaussian distribution, $\mathbf{e} \in \mathrm{N}(0, V_{\sigma})$ with mean zero and $m \times m$ variance matrix V_{σ} that encodes the spatio-temporal correlation. The correlation behaviour is determined by the second set of parameters σ, known as hyper-parameters that, together with the spatio-temporal coordinates $(\mathbf{x_i}, \mathbf{t_i})$, determine the variance matrix V_{σ}.

We give an example of how we can characterise this correlation relating to two spatial dimensions and one temporal dimension. For any two measurements y_i and y_j, the covariance $V_{\sigma}(i, j)$, $\sigma = (\sigma, \sigma_0, \lambda, \tau)^{\mathrm{T}}$, associated with the corresponding random effects e_i and e_j is given by

$$V_{\sigma}(i, j) = k(\mathbf{x_i}, \mathbf{x_j}, \mathbf{t_i}, \mathbf{t_j} | \lambda, \tau) = \sigma^{\mathbf{2}} \mathbf{k}(\mathbf{x_i}, \mathbf{x_j} | \lambda) \mathbf{k}(\mathbf{t_i}, \mathbf{t_j} | \tau), \quad \mathbf{i} \neq \mathbf{j}, \tag{6}$$

and $V(\sigma)(i, i) = \sigma^2 + \sigma_0^2$. Here,

$$k(\mathbf{x}, \mathbf{x}' | \lambda) = \exp\left\{ -\frac{\mathbf{1}}{\mathbf{2}\lambda^{\mathbf{2}}} (\mathbf{x} - \mathbf{x}')^{\mathrm{T}} (\mathbf{x} - \mathbf{x}') \right\} \tag{7}$$

is a kernel that encodes the spatial correlation and depends on a length scale parameter λ, and, similarly,

$$k(t,t'|\tau) = \exp\left\{-\frac{1}{2\tau^2}(t-t')^2\right\} \tag{8}$$

encodes the temporal correlation and depends on a time scale parameter τ. The term σ^2 gives the variance of the correlated effects and σ_0^2 the variance associated with repeatability, i.e., random effects that occur over very short spatial and temporal scales. These effects could be associated with either the signal, the measuring sensor or both.

It is also possible to use kernels that encode more complex behaviour. For example, correlation involving diurnal or other periodical cycles can be encoded in a kernel of the form

$$k(t,t'|\tau_1,\tau_2) = \exp\left\{-\frac{1}{2\tau_1^2}[t-t']_1^2 - \frac{1}{2\tau_2^2}[t-t']_2^2\right\}$$

where $[t-t']_1$ is the time difference modulo a cycle and $[t-t']_2$ is time difference in terms of number of cycles.

4.3 The Kalman filter

The Kalman filter can be seen as an example of a Gaussian process operating over the temporal domain. We consider here the simplest example governed by

$$a_k|a_{k-1} \sim \mathrm{N}(a_{k-1},\sigma_P^2), \quad y_k|a_k \sim \mathrm{N}(0,\sigma_M^2), \quad p(a_0) \propto 1. \tag{9}$$

The first relationship shows how the quantity a_k is predicted with uncertainty σ_P on the basis on knowing a_{k-1}. The second relationship defines the likelihood associated with a measurement of a_k. The posterior distribution for $\mathbf{a} = (a_1,\ldots,a_n)^{\mathrm{T}}$ depends on the measurement information and the predictive capability. Estimates of the system parameters are determined by solving the linear least squares system $C\mathbf{a} \approx \mathbf{z}$ with

$$C = \begin{bmatrix} w & -w & & & & \\ & v & & & & \\ & w & -w & & & \\ & & v & & & \\ & & \ddots & & & \\ & & & w & -w & \\ & & & & v & \\ & & & & w & -w \\ & & & & & v \end{bmatrix}, \quad \mathbf{z} = v\begin{bmatrix} 0 \\ y_1 \\ 0 \\ y_2 \\ \vdots \\ 0 \\ y_n \end{bmatrix},$$

with $w = 1/\sigma_P$ and $v = 1/\sigma_M$. The QR factor [14] of C can be determined sequentially with the kth update step having the form

$$\begin{bmatrix} r_{22} & 0 \\ w & -w \\ 0 & v \end{bmatrix} \mapsto \begin{bmatrix} \tilde{r}_{11} & \tilde{r}_{12} \\ 0 & \tilde{r}_{22} \end{bmatrix}.$$

This Kalman filter converges in the sense that eventually the bottom 2×2 upper triangle is unchanged. The requirement that

$$\begin{bmatrix} s_{11} & s_{12} & 0 \\ 0 & s_{22} & 0 \\ 0 & w & -w \\ 0 & 0 & v \end{bmatrix} \mapsto \begin{bmatrix} s_{11} & s_{12} & 0 \\ 0 & s_{11} & s_{12} \\ 0 & 0 & s_{22} \end{bmatrix},$$

leads to the relations

$$s_{11}s_{12} = -w^2, \quad (s_{11}+s_{12})^2 = v^2, \quad s_{22}^2 = s_{11}^2 - w^2,$$

from which we derive

$$s_{11} = -\frac{1}{2^{1/2}\sigma_P}\left(2+\frac{\sigma_P^2}{\sigma_M^2}+\frac{\sigma_P}{\sigma_M}\left(4+\frac{\sigma_P^2}{\sigma_M^2}\right)^{1/2}\right)^{1/2},$$

and

$$s_{12} = -\frac{1}{s_{11}}\frac{\sigma_P^2}{\sigma_M^2}, \quad s_{22} = \left(s_{11}^2 - \frac{\sigma_P^2}{\sigma_M^2}\right)^{1/2}.$$

The upper triangular factor R of the observation matrix is a bi-diagonal matrix whose diagonal elements r_{kk} converge to s_{11} and above diagonal elements converge to s_{12} and r_{nn} converges to s_{22} as n increases. If the filter has converged then, at the kth stage, the uncertainty associated with the estimate of a_k is $1/s_{11}$. However, subsequent steps will help to improve this estimate since the measurement at the $(k+1)$th and subsequent steps provide additional information about a_k.

The variance matrix associated the fitted parameters is given by $(R^{\mathrm{T}}R)^{-1}$ and the uncertainty associated with the kth parameter is given by the $\|\mathbf{g}_k\|$ where $\mathbf{g}_k$ is such that $R^{\mathrm{T}}\mathbf{g}_k = \mathbf{e}_k$ where $\mathbf{e}_k$ is the kth column of the $n\times n$ identity matrix. If the filter has converged, then $\mathbf{g}_k$ solves

$$\begin{bmatrix} \ddots & & & & & & \\ s_{12} & s_{11} & & & & & \\ & s_{12} & s_{11} & & & & \\ & & s_{12} & s_{11} & & & \\ & & & s_{12} & s_{11} & & \\ & & & & s_{12} & s_{11} & \\ & & & & & & \ddots \end{bmatrix} \mathbf{g}_k = \begin{bmatrix} \vdots \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}.$$

We have $\mathbf{g}_k(j) = 0$ if $j < k$, $\mathbf{g}_k(k) = 1/s_{11}$, $\mathbf{g}_k(k+1) = -s_{12}/s_{11}^2$, $\mathbf{g}_k(k+2) = s_{12}^2/s_{11}^3$, etc., so that the elements of $\mathbf{g}_k$ form a geometric series. From this fact, we know that the norm of $\mathbf{g}_k$ is such that

$$u^2(a_k) = \|\mathbf{g}_k\|^2 = \frac{1}{s_{11}^2 - s_{12}^2}. \tag{10}$$

Thus, the asymptotic behaviour of the filter can be determined analytically in terms of σ_P and σ_M. If σ_P is sufficiently smaller that σ_M so that $\sigma_P^4/\sigma_M^4 \approx 0$, then the uncertainty $u(a_k)$ associated with the kth parameter is such that

$$u^2(a_k) \approx \frac{\sigma_P \sigma_M}{2}.$$

For $\sigma_P \gg \sigma_M$, $u(a_k) \approx \sigma_M$, as would be expected. The term

$$n_{\text{eff}} = \frac{\sigma_M^2}{u^2(a_k)} \tag{11}$$

with $u(a_k)$ given by (10) represents how many independent repeated measurements would be required to determine the same uncertainty associated with a_k as that achieved by exploiting the predicative capability in the filter.

Figure 8 shows the uncertainty $u(a_k)$ as a function of k for the case $\sigma_M = 1$ and (a) $\sigma_P = 0.2$, top, and (b) $\sigma_P = 0.1$, bottom. For $\sigma_P = 0.2$, corresponding to weaker predictive capability, the asymptotic value of the uncertainty $u(a_k)$ is larger but is attained in few time steps. Figure 9 shows the asymptotic value of $u(a_k)$ given by (10) as a function of σ_P.

The analysis for this simple Kalman filter model shows the extent to which correlation over time or the spatial domain can be used to improve the uncertainties associated with the fitted parameter estimates.

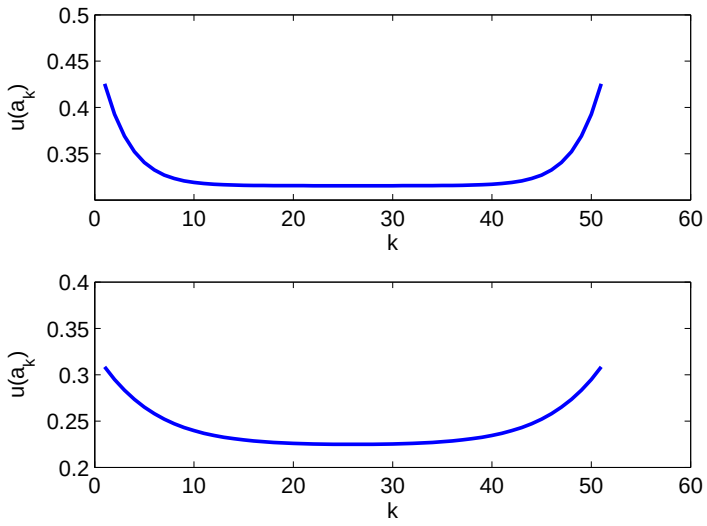

Fig. 8. Uncertainty $u(a_k)$ as a function of k associated with the simple Kalman filter (9) for the case $\sigma_M = 1$ and (a) $\sigma_P = 0.2$, top, and (b) $\sigma_P = 0.1$, bottom.

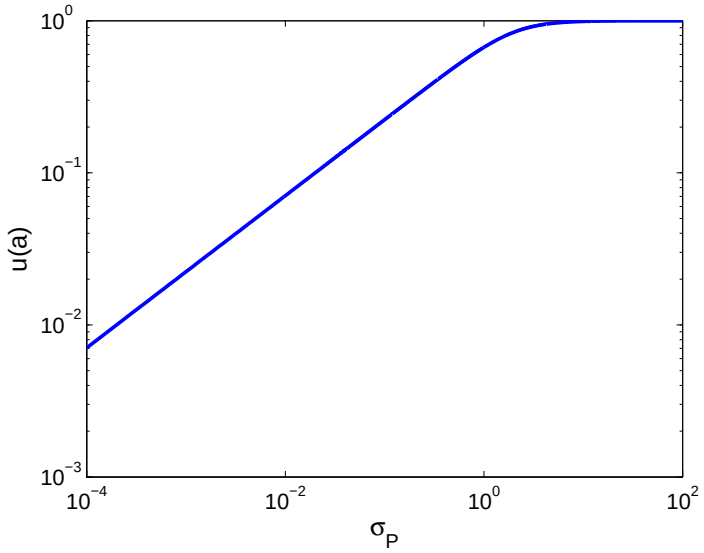

Fig. 9. Asymptotic value of $u(a_k)$ given by (10) as a function of σ_P for the case $\sigma_M = 1$.

4.4 Estimating the hyper parameters

In general, the hyperparameters associated with the spatio-temporal correlation have to be estimated from the data and we use a Bayesian formulation to define this computational approach. We assume that some prior information about the hyperparameters σ is available and encoded in a density $p(\sigma)$. Given the observed data $\mathbf{y}$, the posterior joint distribution $p(\mathbf{a},\sigma|\mathbf{y})$ for $\mathbf{a}$ and σ is such that

$$p(\mathbf{a},\sigma|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{a},\sigma)p(\mathbf{a})p(\sigma)$$

involving the likelihood $p(\mathbf{y}|\mathbf{a},\sigma)$ and the priors. The fact that the likelihood is Gaussian and that the model parameters occur linearly in the model means that the maginalised posterior distribution $p(\sigma|\mathbf{y})$ for σ, given by

$$p(\sigma|\mathbf{y}) = \int_A p(\mathbf{a},\sigma|\mathbf{y})d\mathbf{a},$$

can be evaluated analytically. If $\hat{\mathbf{a}} = \mathbf{a}_\sigma$ is the least squares solution that minimises

$$(\mathbf{y} - C\mathbf{a})^{\mathrm{T}} V_\sigma^{-1} (\mathbf{y} - C\mathbf{a}),$$

then the marginalised distribution $p(\sigma|\mathbf{y})$ is such that

$$p(\sigma|\mathbf{y}) \propto p(\sigma)|V_\sigma|^{-1/2}|C^{\mathrm{T}}V_\sigma^{-1}C|^{-1/2}\exp\left\{-\frac{1}{2}F(\sigma)\right\},$$

where $F(\sigma)$ is the residual sum of squares

$$F(\sigma) = (\mathbf{y} - C\hat{\mathbf{a}}_\sigma)V_\sigma^{-1}(\mathbf{y} - C\hat{\mathbf{a}}_\sigma). \tag{12}$$

The σ that maximises $p(\sigma|\mathbf{y})$ can be found be minimising $-\log p(\sigma|\mathbf{y})$ or equivalently, minimising

$$E(\sigma|\mathbf{y}) = \frac{1}{2}\left(F(\sigma) + \log|V_\sigma| + \log|C^{\mathrm{T}}V_\sigma^{-1}C|\right) - \log p(\sigma). \tag{13}$$

4.5 Analysis of air quality data

We have applied a spatio-temporal Gaussian process model (GP) defined by (6–8) to air quality data extracted from the London Air Quality Network (LAQN) [23] over an eight week period starting on 28th February 2011. The results reported here relate to the measurement of NO_2 gathered from five sites in central london within an area of approximately 5 km radius. The measurements were taken hourly. Estimates of the four hyperparameters were determined by minimising $E(\sigma|\mathbf{y})$ in (13) and the fitted values were $\sigma = 15.1\ \mu\mathrm{g\ m}^{-3}$, $\sigma_0 = 4.5\ \mu\mathrm{g\ m}^{-3}$, $\lambda = 6.5$ km, and $\tau = 2.5$ hr. The value of the length scale parameter suggests that there is significant spatial correlation in the data. Figure 10 shows the Gaussian process model fitted to the data from week 1; the spatial correlation is reflected in the similarity in the data series. The figure also shows the $\pm$ two standard deviations uncertainty band associated with the fitted model.

The spatial correlation enables predictions to be made for missing data. Figure 11 shows on the top the GP model fitted to a complete set of 7 days for a site (top) and the predicted results on the basis of the first three days and the results from neighbouring sites for all seven days (middle). It is seen that the prediction for the last four days is almost the same as model fit to all the data (top).

The spatial correlation also enables the inter-calibration of sensors. We can simulate an experiment in which, after a number of days, a calibrated sensor is replaced by an uncalibrated sensor with an unknown offset a_0 to be determined as part of the model fitting process. The bottom graph in figure 11 shows the fitted GP model for one urban background site for days 5 to 7 on basis of days 1 to 3 and data from days 5 to 7 subject to an unknown calibration offset compared to the actual measured data. The actual offset applied to the data for days 5 to 7 was 20 $\mu\mathrm{g\ m}^{-3}$ with an associated uncertainty of 1.0 $\mu\mathrm{g\ m}^{-3}$, demonstrating that an accurate cross-calibration is possible.

5 Summary and concluding remarks

The impact of metrology in addressing societal challenges depends on being able to develop appropriate concepts of traceability, uncertainty and calibration for complex systems, in particular, accounting for the uncertainty associated with the modelling of such systems. In this paper, we have provided an overview of the concept of traceability from a statistical modelling point of view, and looked at how tools such as model selection and model average can provide a coherent approach to assessing the uncertainty associated with models. Finally, we have used a Gaussian process model to help develop estimates of the air quality field on the basis of sensor network data and implement a cross-calibration scheme of sensors.

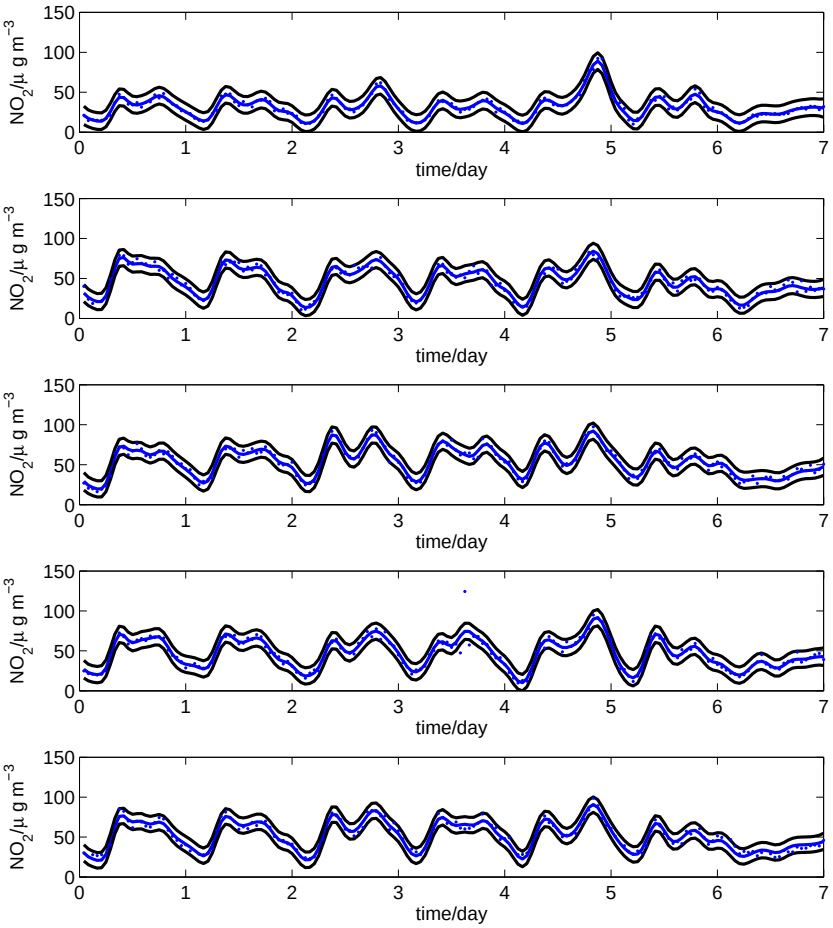

Fig. 10. Gaussian process model fitted to NO_2 data for week 1.

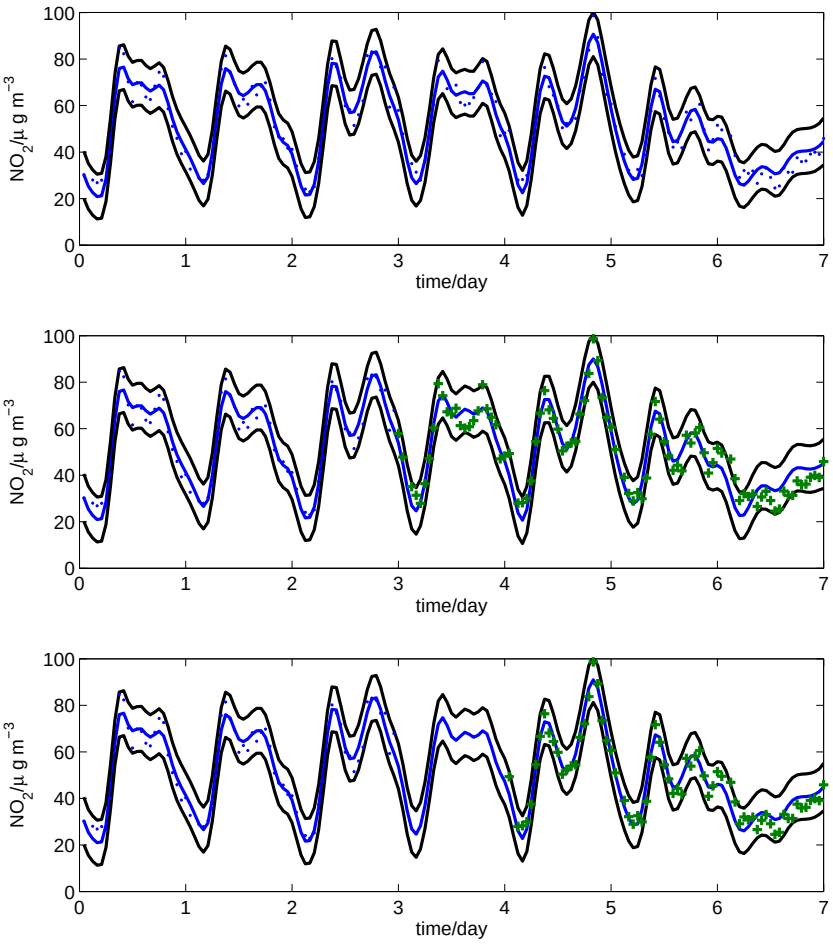

Fig. 11. Prediction and calibration for NO_2 data for week 1.

Acknowledgements

This work has been supported by the UK's National Measurement System programme for Mathematics and Modelling and through a UK NERC Network of Sensors project on air quality led by the University of Cambridge. I thank my colleagues Elena Barton, Peter Harris, Minh Hoang and Martin Milton for many helpful discussions.

References

1. H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.
2. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. Joint Committee for Guides in Metrology, JCGM 100:2008.
3. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. International vocabulary of metrology — basic and general concepts and associated terms,. Joint Committee for Guides in Metrology, JCGM 200, 2008.
4. R. T. Birge. Probable values of the general physical constants. *Rev. Mod. Phys*, 1:1–73, 1929.
5. H. Chipman, E. I. George, and R. E. McCulloch. *The practical implementation of Bayesian model selection*. Institute of Mathematical Statistics, Beachwood, Ohio, 2001.
6. A. G. Chunovkina, C. Elster, I. Lira, and W. Wöger. Analysis of key comparison data and laboratory biases. *Metrologia*, 45(2):211–216, 2008.
7. M. G. Cox. The evaluation of key comparison data. *Metrologia*, 39:589–595, 2002.
8. M. G. Cox, A. B. Forbes, J. Flowers, and P. M. Harris. Least squares adjustment in the presence of discrepant data. In P. Ciarlini, M. G. Cox, F. Pavese, and G. B. Rossi, editors, *Advanced Mathematical and Computational Tools in Metrology VI*, pages 37–51, Singapore, 2004. World Scientific.
9. M.G. Cox. The evaluation of key comparison data: determining the largest consistent subset. *Metrologia*, 44(3):187–200, 2007.
10. N. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, New Jersey, 2011.
11. C. Elster and B. Toman. Analysis of key comparisons: estimating laboratories' biases by a fixed effects model using bayesian model averaging. *Metrologia*, 47(3):113–119, 2010.
12. A. B. Forbes. Empirical functions with pre-assigned correlation behaviour. In F. Pavese, W. Bremser, A. Chunovkina, N. Fischer, and A. B. Forbes, editors, *Advanced Mathematical and Computational Tools for Metrology X*, pages 17–28, Singapore, 2015. World Scientific.
13. A. B. Forbes and C. Perruchet. Measurement systems analysis: concepts and computational approaches. In *IMEKO World Congress, September 18–22, 2006, Rio de Janeiro*, 2006.
14. G. H. Golub and C. F. Van Loan. *Matrix Computations*. John Hopkins University Press, Baltimore, third edition, 1996.
15. J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14:382–401, 1999.
16. C. M. Hurvich and C. Tsai. Regression and time series model selection in sample samples. *Biometrika*, 76:297–307, 1989.
17. International Organization for Standardization, Geneva. *ISO 5725-1: Accuracy (trueness and precision) of measurement methods and results – Part 1: General principles and definitions*, 1994.

18. R. N. Kacker, A. Forbes, R. Kessel, and K.-D. Sommer. Classical and Bayesian interpretation of the Birge test of consistency and its generalized version for correlated results from interlaboratory evaluations. *Metrologia*, 45(3):257–264, 2008.
19. M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *J. Roy. Sat. Soc. B*, 64(3):425–464, 2001.
20. H. Linhart and W. Zucchini. *Model Selection*. Wiley, New York, 1986.
21. I. Lira. Bayesian evaluation of comparison data. *Metrologia*, 43:S231–S234, 2006.
22. I. Lira. Combining inconsistent data from interlaboratory comparisons. *Metrologia*, 44(5):415–422, 2007.
23. London Air Quality Network. `www.londonair.org.uk`.
24. R. C. Paule and J. Mandel. Consensus values and weighting factors. *J Res. Natl. Bur. Stand.*, 87:377–385, 1982.
25. A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression. *Journal of the American Statistical Association*, 92:179–191, 1997.
26. C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Mass., 2006.
27. G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.
28. K. Shirono and K. Tanaka, H.and Ehara. Bayesian statistics for determination of the reference value and degree of equivalence of inconsistent comparison data. *Metrologia*, 47(4):444–452, 2010.
29. B. Toman. Statistical interpretation of key comparison degrees of equivalence based on distributions of belief. *Metrologia*, 44(2):L14–L17, 2007.
30. K. Weise and W. Woeger. Removing model and data non-conformity in measurement evaluation. *Measurement Science and Technology*, 11:1649–1658, 2000.
31. R. Willink. Statistical determination of a comparison reference value using hidden errors. *Metrologia*, 39:343–354, 2002.