LETTER FROM THE PRESIDENT

Gas leaks and galacticos

Your beliefs are important, says Fabrizio Ruggeri, who celebrates Italy's World Cup success

Italy are the world champions! I am on holiday in the mid-USA town of Garibaldi, named after my own national hero, but nobody here cares. But **you** all know and understand how I feel: *Italy just won the football World Cup in Germany!*

Looking at the data (the ranking by FIFA, the world football association), Italy were not one of the favourites. The Czech Republic and USA, for example, were among the top teams before the tournament, but both were eliminated in the first round by Italy. For many months, I told my friends that Italy were serious candidates for victory. I told them Italy had excellent forward players. Actually, the forward players scored few goals in the end, with defenders being the most important players, even scoring goals.

Years ago, gas was leaking from a network of a city's old cast iron pipes. We (a mathematician and some engineers) were asked to predict the conditions under which leaks were most likely. The data alone indicated pipes with large diameters, laid deeply under traffic. But the data related to only three escapes in only six kilometres of the full 320km-long network. One escape more, or one less, can make a huge difference to maximum likelihood estimators. When we interviewed 26 people from the gas company we concluded that the worst pipes were those with small diameters that were not laid deeply, and ran under traffic. We reached this conclusion, although experts had different opinions and we had different ideas on how to express the experts' opinions.

Although the two examples are from very different fields, they share an important feature: data, by itself, can lead to conclusions that may be questionable, whereas experts' opinions can be used in estimation and forecasting, although the estimates may be inaccurate and the forecasts controversial.

Therefore, from the saves of Buffon and Cannavaro to the goals by Materazzi and Grosso we move to one of the most serious debates within statistics: **frequentist vs Bayesian**.

I warn you that I am very biased: I am a hard-core Bayesian! If you are unaware of the Bayesian approach, I suggest the books by Bernardo and Smith (1994, Wiley), Robert (2001, Springer) and, for Bayesian data analysis, Gelman, Carlin, Stern and Rubin (2004, Chapman and Hall). As a very basic introduction to Bayesian statistics, think of a random, observable quantity (such as the lifetime of a light bulb) modelled by a distribution depending upon an unknown parameter (such as an exponential with parameter lambda). An opinion (prior distribution) could be expressed about the parameter and, via Bayes's theorem, data would update it into a posterior distribution, to be used for both inference and prediction.

In the case of the leaking cast-iron pipes, engineers were used to Gaussian distributions and the lognormal distribution was their natural choice as the prior on the parameter of a Poisson model. The mathematician was more inclined to the Gamma distribution, which would lead to simpler computations (being conjugate with respect to the model, in our specialist language).

We asked the experts to compare different sections of the gas network defined by location, depth and diameter, taking two values each so there were eight classes. The 26 experts (from different areas of the company) compared them pairwise, saying which one was more likely to have gas escapes before the other. The experts could express their opinions as 'Class A is more likely than class B to fail', 'definitely more likely' and other different levels of certainty. At each qualitative judgement a number between one and nine was assigned for each of such statements and the inverse, between one and 1/9, for statement about minor probability. This produces an eight-square matrix for all comparisons.

Saaty proposed this method, called *Analytic Hierarchy Process*. He takes the eigenvector associated with the largest eigenvalue of the matrix as the vector of the probabilities of occurrence of each event (in our case the probability that a gas escape occurs in each class). The experts had different probabilities for each class. We considered the 26 probabilities for each class as a 'sample' and we computed sample mean and variance and used them to choose the prior distribution on the parameter of interest.

In another study of gas escapes, but in steel pipes this time, we asked experts to choose a length of network and assign to it a probability of gas escape from a list of values. From this information, we were able to estimate the parameters of the prior distribution.

We asked unit managers of an oil company for their opinions on the costs of the parts of a project they were in charge of, to prepare a bid for the construction of a plant. What did each manager think was the most likely value and what were their estimates of the minimum and maximum values? We expressed these opinions as prior, triangular, distributions.

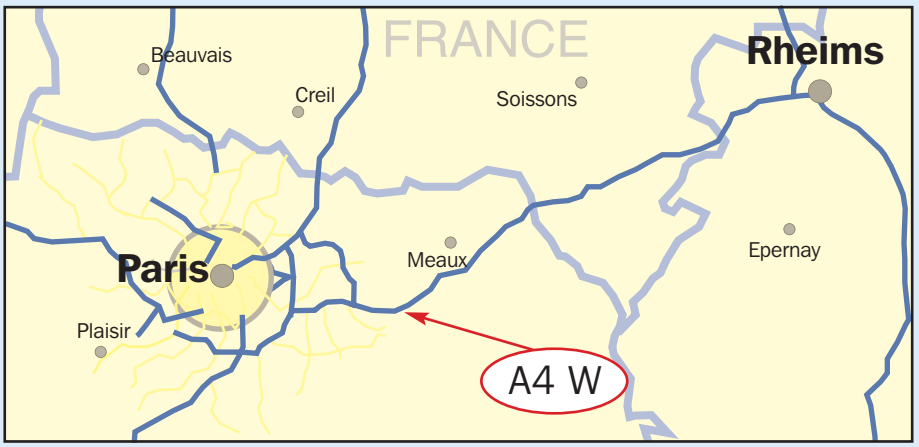
These are just a few possible ways to use experts' opinions, among many presented in literature. Statisticians can waste knowledge if they do not use prior information. But in the Bayesian approach, arbitrary information may be used and the prior may be badly chosen. Robustness of the method has been thoroughly studied in the past (see Rios Insua and Ruggeri, 2000, Springer).

The formalised use of prior information was thoroughly discussed in a workshop organised by the ENBIS Special Interest Group on Reliability at the 2005 ENBIS meeting in Newcastle. At the same time, a workshop on six-sigma was organised. A mix of more theoretical and more applied papers and workshops, combined with a mix of people from industry and academia, are the characteristics of the ENBIS events. The first opportunity to appreciate it is in the 2006 ENBIS meeting in Wroclaw, Poland, on 18 to 20 September. Further opportunities are at the 2007 Spring workshop on computer experiments in Torino, Italy, and a workshop on reliability just before a major event in the field.

Beating the Paris snarl-ups

Fabrice Gamboa, Jean-Michel Loubes and Elie Maza

show how cluster analysis enables them to make short-term forecasts of travel times. They generalise their method and suggest more applications



Traffic builds up as you drive into Paris from Rheims or Meaux in the east. As more commuters funnel in from small towns and villages, traffic slows, time and money are wasted and some drivers lose patience. The highways authority has built a road network to ease the congestion, but this is not enough. Drivers and traffic controllers also need good forecasting of flow density and velocity. Our task is to produce short-term (about one hour) forecasts of travel time on the Parisian highway network.

Long-term road traffic forecasting was developed a long time ago but short-term forecasting is recent. New technologies enable us to obtain precise quantitative data, such as speed, flow and occupancy, as well as qualitative information such as whether a car stream is moving, blocking or stopped.

The ultimate aim is to forecast, at any time H, the time needed at time H+h, to travel from any point on the network to any another. To do this, we must predict speed at every point of the observation grid: all the measurement stations of the network. We limited our study to an axis of the highway network (named A4W) that is known to represent Parisian road traffic behaviour and where it is difficult to forecast travel times. This road section is 21.82 kilometres long and has 38 counting stations about 500 metres apart. We also chose h to be four hours.

Each station recorded, every three minutes of the day, the flow, the occupancy, and the velocity of the vehicle flow estimated as the mean over the previous six minutes. The whole database represents three years of measures. To deal with this high-dimensional data, we developed the following method.

We relied on two common assumptions. First, short-term prediction of traffic mostly depends on what just happened. Second, there is a fixed number of traffic patterns, and every new observation day can be compared with these patterns. Our data confirmed these assumptions. Thus, we can divide the forecasting of travel time into three steps.

First, we estimate the representative behaviours or patterns of speeds. We used cluster analysis to define different patterns. A cluster analysis (or classification) is a statistical method that aims to gather similar individuals into a cluster. Subsequently, the issue is to find a good mathematical definition of the term 'similar'. Also, much of this work was to find the optimum number of clusters and measures of distance between them.

Second, we identify the specific behaviour of each pattern. This is important. In Figure 2 (a), we can see that each speed curve seems to be related to each other curve by a single shift. We observe that speed curves with the same traffic jam or speed reduction have the same form but different starting times. So, because of these shifts, the mean curve is not representative enough and does not convey the true information. Hence, we choose any one curve as a base and calculate a shift for each other curve from that base. When we offset the shifts, we obtain the mean curve with blue line on Figure 2 (d). So, to find the best representative profile, the structural estimator, we estimate the shifts, move the curves in the cluster by the shifted amounts and then take the mean to get a good estimate of the road traffic behaviour corresponding to that pattern.

Third, we compare the incoming observations with these patterns, and allocate each set of observations to the most probable pattern.

Our method relies on a probabilistic study of real data. This enables us to predict with more accuracy than with standard methods based on deterministic modelling without probability components. The strength of this project lies in combining the knowledge of high level researchers and engineers. Starting from a practical issue, we use statistical theory to improve the results, keeping in mind the practical goal of our project, which is the main purpose of applied mathematics.

Our results are better than those from standard forecasting procedures. For example, Figure 1 shows the evolution of the forecast travel times throughout a particular day on the A4W trunk road. We predicted travel times for a traveller starting his trip at each time of the day.

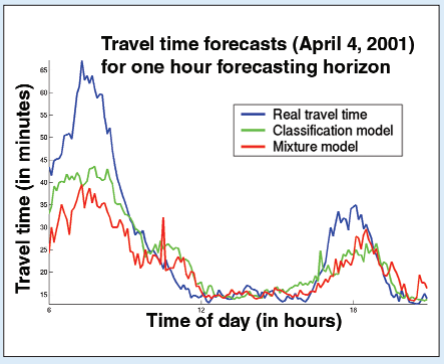


Figure 1. Evolution of the forecast travel times for a specific journey, from one point on the route to another. Real travel times are plotted with a blue line. Travel times predicted with a standard forecasting procedure are plotted with a red line. Travel times predicted with our method are plotted with a green line.

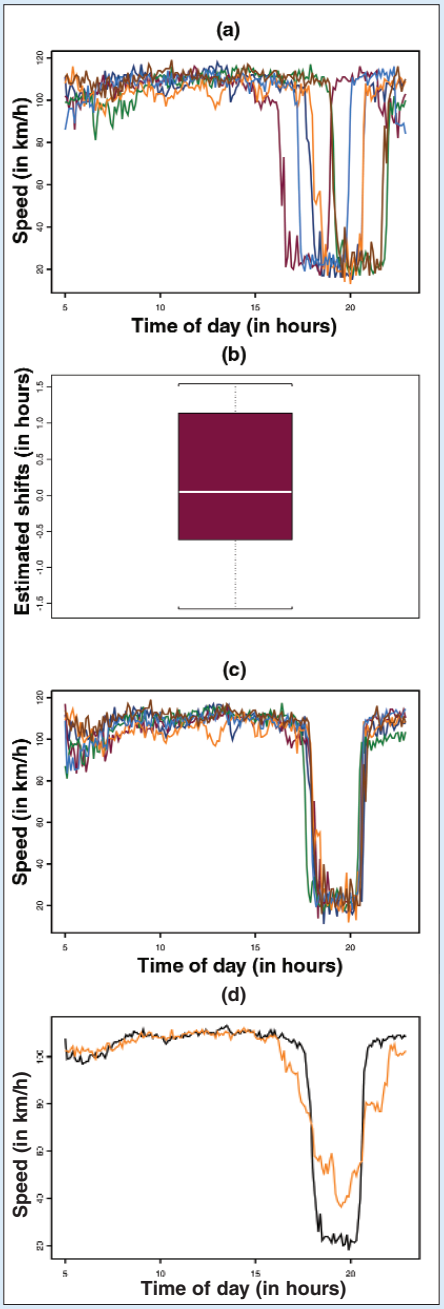


Figure 2. Traffic speeds recorded at a single counting station with estimates of shift values and of the mean of the shifted curves.

Structural estimation with shifts estimation

Our method can be used to analyse experimental data when the outcome is a noisy sample of curves instead of a random variable. Examples of such data might be growth curves, longitudinal data in medicine, speech signals, traffic data or expenditure curves for some goods in the econometric domain. The individuals usually experience similar events that are described by a pattern but the starting time of the event occurs sooner or later. Hence, computing a classical representative curve for this sample severely distorts the data analysis. The average curve (usually the mean or the median) over-smoothes the studied phenomenon and distorts the reality. Hence, the solution we propose is by two steps:

- Estimate the transformation between the curves; and
- Invert the transformations so as to align the data and estimate the feature of the observed phenomenon by taking the mean of the rescaled curves.

Mathematical modelling

A sample of observed functions may be modelled as follows. We observe noisy data that corresponds to the values of the functions observed at discrete observation times and blurred by a stochastic noise. This noise is a random variable that models all the uncertainty and errors in the measurements. We assume that these functions are close to each other in the sense that there exists an unknown archetype with unknown parameters such that the observed curves differ from the archetype by the values of the parameters. So, to recover the archetype pattern, these parameters must be estimated so as to align the shifted curves and then build an estimator of the archetype by taking the average of the rescaled curves.

Methodology

The main difficulty of this method is that the estimation of the shift parameters cannot rely on the pattern, which is unknown, but these quantities are deeply linked. For this reason we use an estimator built on the discrete Fourier coefficients of the observations and consider the estimation problem in the frequency domain. Under identifiability assumptions, we provide a consistent method to estimate the unknown translation parameters by using a construction that relies on semi-parametric estimation theory. It is defined as the solution of a minimisation program, which can be solved using quadratic methods.

We prove that these estimators are close to the real shifts and that fluctuations of our estimates are asymptotically Gaussian. This theoretical result enables us to test the validity of our assumptions. We also provide an efficient algorithm to compute the shifts and finally build the structural estimator of the archetype.

Numerical results for road traffic data

For our traffic data example, the results are as follows. Figure 2 (a) represents a particular cluster at a particular counting station. Figure 2 (b) shows the box-plot of estimated shifts. Shifted curves are plotted on Figure 2 (c). So, in this homogeneous cluster, where only a shift phenomenon appears, difference are obvious between the mean curves in Figure 2 (d) of shifted curves (blue line) and of primary curves (red line). Hence, the shift estimated mean is clearly more representative of the individual behaviour.

The forecasting method is specific to our traffic problem, but shift estimation is quite general (see the forthcoming papers: 'Road trafficking description and short term travel time forecasting, with a classification method' by J-M. Loubes, E. Maza, M. Lavielle & L. Rodriguez; and 'Semiparametric estimation of shifts between curves' by F. Gamboa, J-M. Loubes and E. Maza).

The national research project, MIST-R (Modélisation Informatique et Statistique du Trafic Routier), was set up by the CNRS to forecast road traffic travel times. The director is Jean-Michel Loubes (CNRS and Université Montpellier 2). Local directors are Mehdi Danech-Pajouh (INRETS), Fabrice Gamboa (Université Toulouse 3) and Michèle Sebag (CNRS and Université Paris Sud). There are 16 people in four research teams comprising statisticians, computer scientists and road traffic professionals. Their first aim is to understand the different phenomena appearing in road traffic, mixing both periodic and random effects. Then they will build models of road traffic short-term travel time forecasting.

Jean-Michel Loubes is at CNRS and at Laboratoire de Mathématiques, Equipe de Probabilités et Statistique, Bâtiment 9, Université Montpellier 2, 34000 Montpellier
Jean-Michel.Loubes@math.univ-montp2.fr

Fabrice Gamboa and Elie Maza are at Laboratoire de Statistique et Probabilités, UMR C5583, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse
Fabrice.Gamboa@math.ups-tlse.fr

Autocorrelation is all around us

John Logsdon shows you how to recognise autocorrelation when you meet it

The April/May issue of *Scientific Computing World* included my article on serial correlation. An updated version is on the ENBIS website together with the missing graphs. In this article, I explain a little more about serial correlation and how to detect it. Data that are serially correlated, you will recall, are those where each measurement is influenced by earlier observations.

We considered protective oxidation where oxidation that had already occurred would impede further oxidation. We expect the sequential measurements to be interdependent in some way. An excess of corrosion in one step is expected to lead to a correspondingly smaller increase in the next step and vice versa. We would therefore expect the serial correlation to be negative.

But what does autocorrelated data look like? Some coding is contained in www.enbis.org/SCW5.R but, to use this program, you need to download the R Statistics and Graphing program from www.r-project.org. Installation of R will put an icon on your desktop.

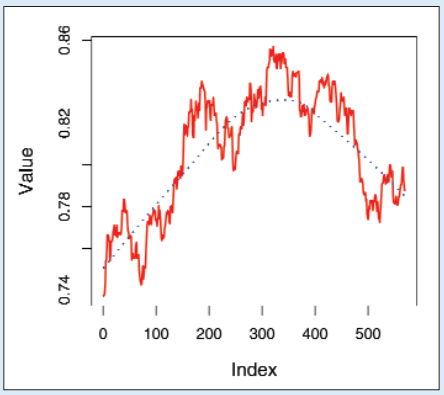


Figure 1 - Mystery data

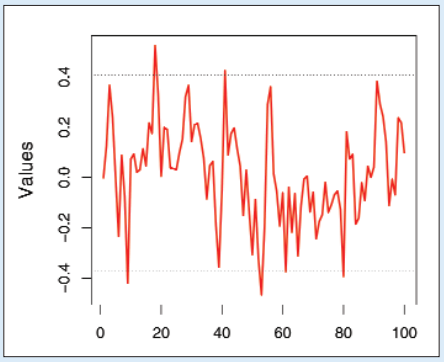


Figure 2 - Is it random?

Visit www.enbis.org/SCW5.R to view the coding. Select and copy all the text into the clipboard. Now open R. You will see a window R Console with an entry point in red (> ■). Above the window is a menu bar. Click on File, then on New Script. Paste text from clipboard into the New script window. Save Script as SCW5.R. In the R Console window, type

```
>       source("SCW5.R")
```

Let's look at some real data:

```
>       plot.surprise()
```

will plot Figure 1. Note how the data has some bigger bumps and troughs with some strong peaks between them. There is a trend upwards and then downward but perhaps my blue smoothed line has misled you. Does this data look random to you? And if not, how non-random is it?

Generate your own autocorrelated sequence by:

```
>       gen.auto()
```

Type for example:

```
>       gen.auto(N=1000,corr=-0.99,sd=0.5)
```

Explore by setting corr, the autocorrelation, in turn to 0.7, -0.5, 0.3, 0, 1, -1, 0.99, and -0.99. Do any of the resulting plots look like our mystery data? Does the trace in Figure 2 look random? There are 100 points of which five are outside the 95 per cent confidence interval but it has an autocorrelation of 0.5. The data are not random.

Do your traces remind you of any of your data? If they do – and it is highly unlikely that any scientist or engineer has never seen such things – then that data will have serial correlation. Examples that spring to mind are transducer traces such as temperatures and pressures. We would spend hours looking at these in the past.

Note how large negative correlation leads to bunching that looks almost like an audio profile. Data with large positive correlation looks like the stock market; is this any surprise? Can you explain these to yourself?

While it is easy to recognise very large autocorrelations, it is generally difficult to spot data with smaller, but still substantial, positive or negative autocorrelation. Try for yourself. Generate some plots and pass them round the office. Serial correlation is frequently there but not easy to see. But you can't miss it if you sniff around, which is why it should be taken into consideration. Fun isn't it?

So what is the mystery data of Figure 1? It is the inter-bank dollar-euro exchange rates from the beginning of 2005 to mid 2006: 569 trading days. These data have an autocorrelation of about 0.993. When the euro is high, the next day it is also likely to be high and vice versa. This is so even if the trend line is removed – the autocorrelation drops slightly to 0.97. It is no surprise that financial data has high autocorrelation.

The data we have been exploring has what is known as lag 1 autocorrelation and the shorthand we use is AR(1) or AR1. That means that the relationship is between immediately adjacent measurements. By recurrence, more distant measurements are related but not directly.

Suppose you have some data which you now realise could be auto correlated. What do you do? There are several approaches, but here is by far the simplest. Forget the autocorrelation to start with, analyse the data using some standard regression model and examine the residuals. Any autocorrelation of lag 1 will show up and be reasonably well estimated.

So if your model is just the mean, we can use the proc.auto function that is in the program:

```
>       proc.auto(lm(gen.auto()~1))
```

which generates data, plots it, fits it using a standard linear model and estimates the autocorrelation. But suppose your model is rather more complex. For example, you may have a number of serially-correlated observations with different run lengths. You are now meeting the problems of structured data and the best calculations can take rather longer.

Load up www.enbis.org/SCW4.R file as well as the present program and generate some data using

```
>       run.simple.corr()
```

We are not interested in the plot so much as a by-product. Type:

```
>       proc.corr(my.corr.data)
```

to estimate the autocorrelation from all the data and this will give about the same value as the plot reports.

In this short article I have tried to give you a taste of the world of repeated measurements, time series and the like. It is an enormous field and people spend their lives devoted to it. But, for practical purposes, it suffices to know that serially correlated data is all around and needs to be treated with particular respect.