THE USE OF STATISTICAL TECHNIQUES TO JUSTIFY THE RECOVERY OF CORRUPTED POLLUTION DATA

EurIng Andrew Newman BEng(Hons) CEng MIEE

ABSTRACT:
Data from continuous pollution monitoring can be corrupted due to many reasons, for example, by instrumentation failure, calibration faults or operator error. Although distorted, this data often exhibits an underlying relationship to the correct measurements. Operators intuitively feel that such data can be recovered by minor manipulation, and thus included in their compliance reports. However this can be hard to justify, and regulatory authorities have understandably preferred to regard data as either "good" or "bad".
Borrowing terminology from self validating (SEVA) instruments, data may usefully be regarded as "blurred" where the underlying measurements can still be recovered.
The justification of such data manipulation has been aided by three developments:

i) Much wider adoption of statistical techniques in industry (e.g. six sigma) has improved the skills base;
ii) There is greater availability of software tools; and,
iii) An increase in logger memory capacity means more system and analyser health diagnostics are now recorded.

As a contribution to the debate about "best practice" two detailed industrial case studies are presented to illustrate how several months of blurred data have been recovered, and how statistical techniques and diagnostic logs have been used to justify the data manipulation to the regulatory authorities. Case study one concerns calibration error. Case study two concerns instrument malfunction. To allow this paper to be used as a training aid, all techniques used are explained with examples.

1 INTRODUCTION:

Operators of continuous gas analysis systems, face many common problems, whether they are measuring emissions (CEM applications) or ambient air quality (AQM applications). Confidence in the data can only be maintained if the data availability is high (typically around 95%), however the measurement systems are complex and require considerable effort to achieve such high capture rates.

Indeed with regard to CEM applications the Environment Agency has commented, "*given the rather extreme nature of stack gases…,… it is not surprising that specialised instruments rarely operate for long periods without trouble"*.**[1]** The network management requirement for AQM monitoring is also onerous, and daily checks of the system integrity are recommended.**[2]** This will typically be achieved by automatic data collection and a software generated daily report.

Where instrumentation errors or other faults are not speedily identified and corrected, then the integrity of the measurements may be compromised. Regrettably, it is a not uncommon experience for operators to find that several months of data are questionable. Inspection of such data often intuitively suggests that minor manipulation may recover meaningful measurements. This paper addresses the issue of whether such adjustment can be justified.

For example, figure 1(a) shows corrupt NO2 data; figure 1(b) shows the same data which has been recovered; and figure 1(c) shows reliable data from the same site during a subsequent month. Comparison of these graphs by the human eye suggests that the recovered data, figure 1(b), has a similar morphology to the reference data, figure 1(c). However standard statistical techniques fail to support this observation.

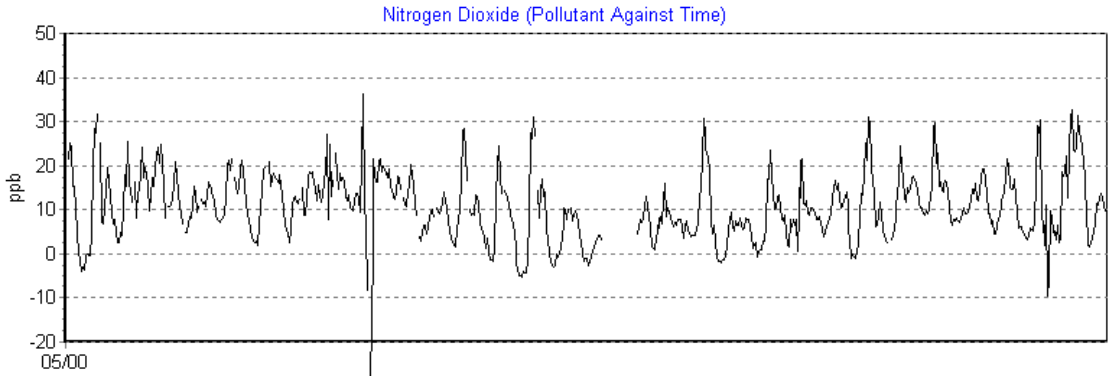


figure 1(a): corrupt (blurred) NO2 data

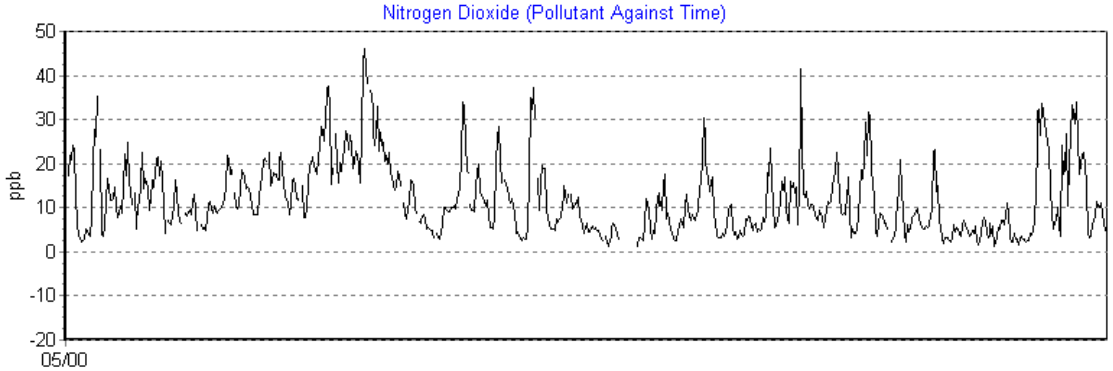


figure1(b): corrected NO2 data

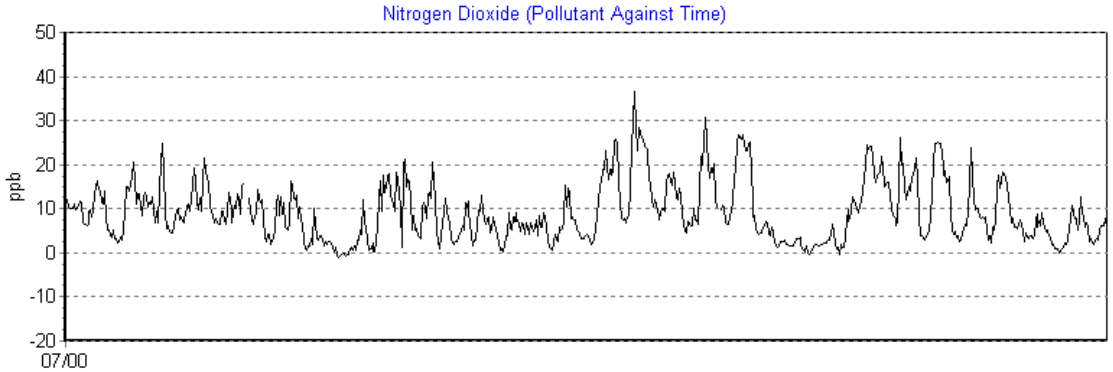


figure 1(c): reference (clear) NO2 data from the same site during a different month

The recovery of data corrupted by instrument failure is the subject of increasing research interest, as automated control becomes more widespread.**[3]** Such developments are particularly weighted towards high volume actuators, such as thermocouples, where development costs are spread over many thousands of units. Such self validating (SEVA) instruments are on the brink of full commercial release by Foxboro.

Béla Lipták observes in his authoritative handbook on Process Control that technology transfer is slow into areas of industry dominated by small companies, and this observation applies to gas analyser manufacturers.**[4]** SEVA technology is therefore unlikely to be directly incorporated into gas analysis systems in the foreseeable future. However, the classification of measurement validity (MV) status can usefully be borrowed from SEVA. These are:

"CLEAR	*The raw data is fine*
DAZZLED	*A possible transient abnormality (e.g. a spike or outlier is possible in the raw data, but has been corrected for and eliminated by the validation algorithm)*
BLURRED	*Abnormal (e.g. noisy) raw data. But believed to have some correspondence to the real measure.*
BLIND	*The raw data is completely untrustworthy, such as with a confirmed and persistent dazzled condition."* **[5]**

The management of air quality data requires skilled and specialist staff, who appreciate the issues of measurement uncertainty, and who can apply statistical techniques.**[5]**. The recovery of "blurred" data is achievable by trained staff, provided that the measurement system provides enough supporting, diagnostic information. Historically, the high cost of data storage has limited the availability of such information, however the falling cost of memory, and particularly the employment of PC loggers has considerably widened the scope of data management. Furthermore, the widespread availability and ease of use of PC spreadsheet packages has contributed to much wider use of statistical techniques in industry. (such as 6σ).

Case studies are presented here which demonstrate how "blurred" data can be recovered, and the data recovery can be justified by the simple application of statistical techniques, accessible to the non-specialist.

Signal Ambitech manufacture integrated systems for both CEM applications (Emiraks) and AQM applications (Ambiraks). These employ PC loggers and monitor system and analyser diagnostics. The Ambirak systems are designed to operate unmanned in remote locations, and system maintenance is aided by the measuring and logging of up to 39 diagnostic parameters (such as temperatures, pressure and flow measurements). These diagnostic parameters may be accessed by telephone.

The availability of such rich information of the analyser operating conditions, means that the air quality monitoring is being conducted in virtually laboratory conditions. Furthermore, the gas analysis techniques are, *mutatis mutandis,* frequently the same as those used in extractive CEM systems. The lessons learned in one application domain can therefore be employed in the other.

The following examples both concern chemiluminescent NOx analysers, which are used in AQM and CEM applications. These analysers measure the concentration of NO. NOx measurement is achieved by passing the gas through a catalytic converter which reduces NO2 to NO. NO2 measurement is thus achieved by cycling the converter in and out of the gas path, and comparing the resulting difference. The analysers drift with time, and the compensation for this drift is achieved by regular referencing to bottled span gas.

CASE STUDY ONE – RESCALING DATA

This case study illustrates how data corrupted by erroneous calibrations was recovered.

2.1 DIAGNOSING THE PROBLEM:

The operator reported repeated negative NO2 readings over a period of two months. Investigation showed that the problem was only observable with the data scaled from mV to ppb (logged in the processed data log), in contrast the raw mV outputs of the analyser bench were consistently positive (logged in the raw data log). Operator error during calibrations was at first suspected, however examination of the calibration trace revealed that NO2 was present during the calibration. The calibration gas should have been NO in N2 with trace amounts of NO2. (See figure 1)

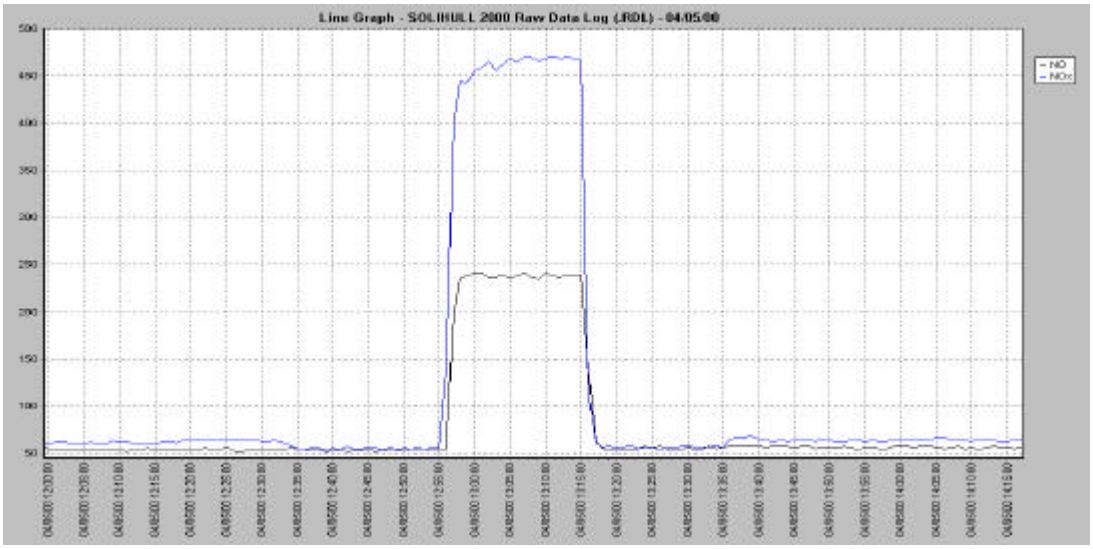


figure 1: calibration trace showing NO2 in the cylinder

A more typical calibration response is shown in figure 2. This is a calibration on the same system after the faulty calibration cylinder was exchanged for a good, new one.

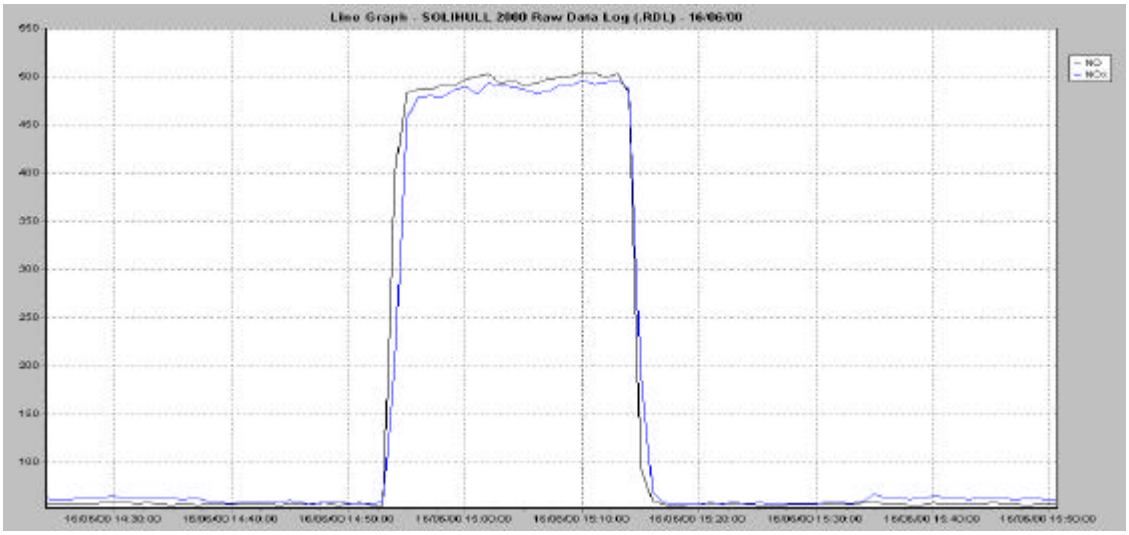


figure 2: calibration response showing only trace amounts of NO2 in the cylinder

The PC logger records all details of the calibrations, so the relative NO and NOx responses were readily available. Plotting these against time (see figure 3) shows that the overall analyser response exhibited linear drift, as would be expected. However, the NO response declined exponentially, which is the characteristic response when there is oxygen in the gas cylinder.

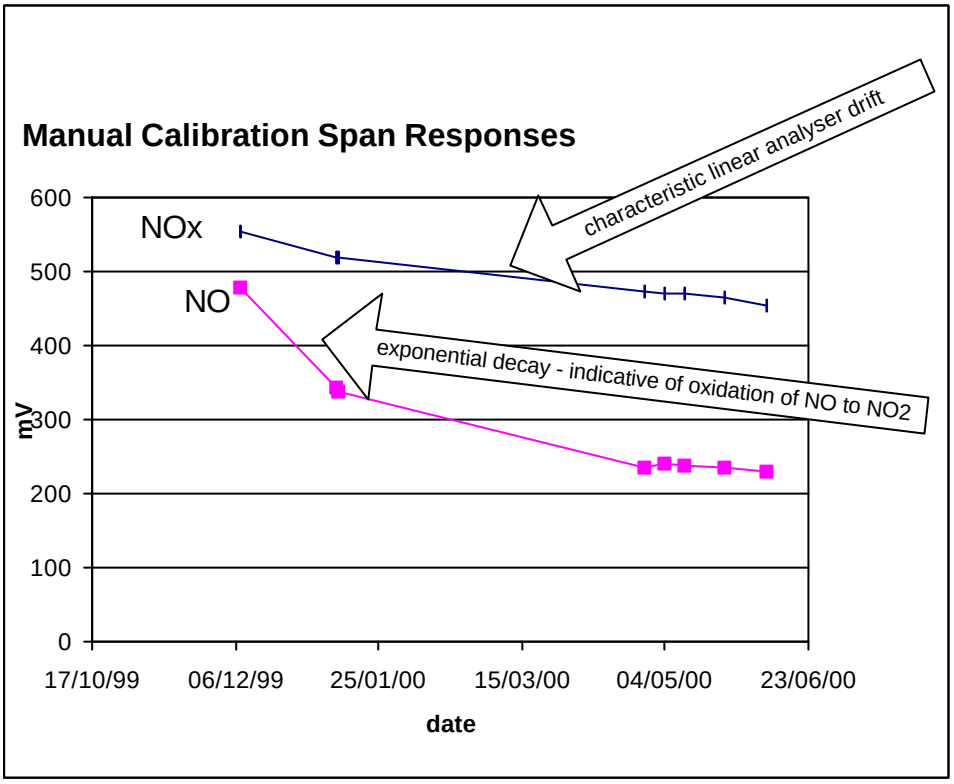


Figure 3: Calibration responses to span gas showing the effect of O2 in the bottle

The gas bottle was replaced under warranty. However the operator had lost two months of NO2 data.

Intuitively, one would expect that the calibration factors gained by calibrating the analyser against the new, good calibration cylinder could be used to recover the corrupt data. Yet can such data manipulation be justified?

2.2 DATA CORRECTION:

The calculation to rescale data is itself straight forward, following the straight line equation, y=mx+c. It is necessary to calculate the values for m and c to correct for each erroneous calibration and then apply these correction factors to the wrongly scaled data. The equation to calculate the correction factors is:

$$M = \frac{\text{expected span gas} - \text{expected zero gas}}{\text{measured span reading} - \text{measured zero reading}}$$

and

C = expected zero gas – (m x measured zero reading)

(the derivation of these equations is given in appendix A.)

The following worked example assumes that the analyser response has not drifted between the two dates. Temporal compensation will be introduced later.

(example incorrect calibration)
On 27th April, The NO span factor was calculated as 2839.97
The NO zero offset was calculated as –152.54
The NOx span factor was calculated as 1408.52
The NOx zero offset was calculated as –79.09

(reference calibration to new gas cylinder)
On 22nd June, The NO span factor was calculated as 1209.2
The NO zero offset was calculated as –62.59
The NOx span factor was calculated as 1201.4
The NOx zero offset was calculated as –63.36

So, these provisional correction factors are:

(1209.2+62.59)/(2839.97+152.54)=0.424991 = $M_{(NO)}$

-62.59 – (0. 42499 x –152.54) = 2.238136 = $C_{(NO)}$

(1201.4 + 63.36)/(1408.52+79.09) = 0.850196 = $M_{(NOx)}$

-63.36 – (-0.850196 x –79.09) = 3.881998 = $C_{(NOx)}$

Through a quirk in Signal's software the gain factors (M) need to be multiplied by 1000, giving:

NO gain = 424.99
NO zero = 2.238
NOx gain = 850.20
NOx zero = 3.882

However, before these gain factors can be applied it must be established to what degree the analyser performance is consistent between the periods using. the new and old calibration gas bottles.

This system implements a daily internal zero and span check (IZS) This employs a dedicated zero generator (scrubber) and a NO2 permeation tube as the span gas source. The implementation of this IZS is therefore orthogonal to the fortnightly manual calibrations which employ bottled gas as the span reference and a second zero air generator. The IZS responses are separately logged in mV, uncorrected by the scaling factors generated during manual calibrations. The IZS can therefore be used as an independent check of analyser performance.

The IZS responses are shown (see figures 4 and 5) to illustrate the analyser performance before and after changing the span gas bottle. Note that the zero response is remarkably stable, and that the analyser gain exhibits typical linear drift.

By plotting the IZS span responses for the 60 days before and the 60 days after the switching of gas bottles it can be established that the analyser response is substantially the same in both periods. This is shown as figure 6. As the two traces trivially correlate, no further justification is offered.

Although this establishes correspondence between the analyser performance in these two periods, it also proves that the new scaling factors calculated above cannot simply be applied, as they do not compensate for the long term analyser drift.

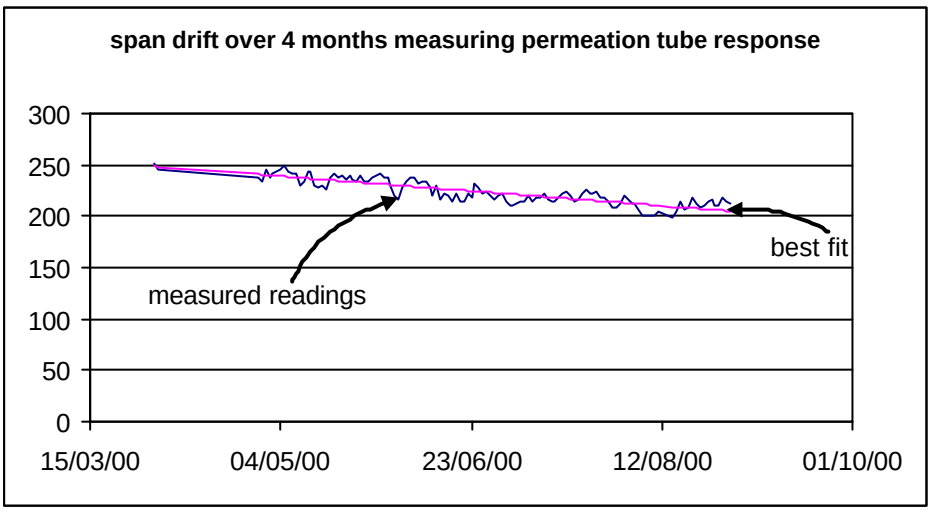


Figure 3: decrease in analyser gain as exhibited by permeation tube response

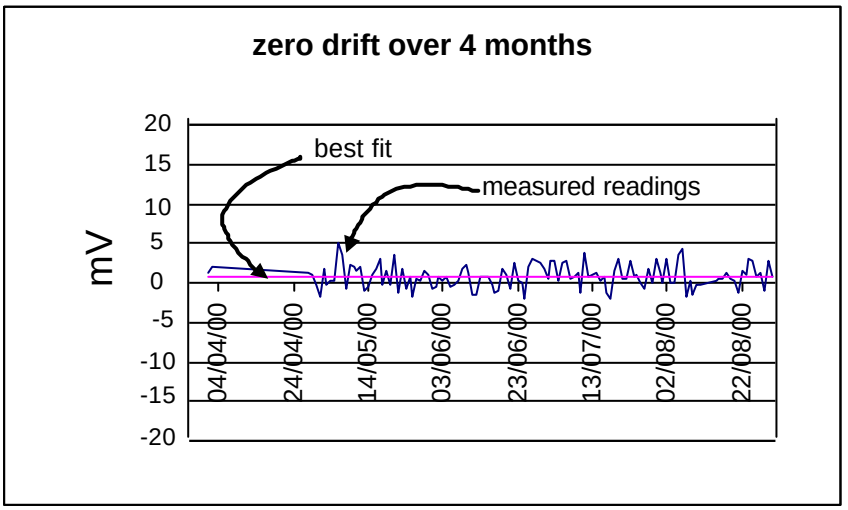


Figure 4: stability of zero response as exhibited by IZS zero

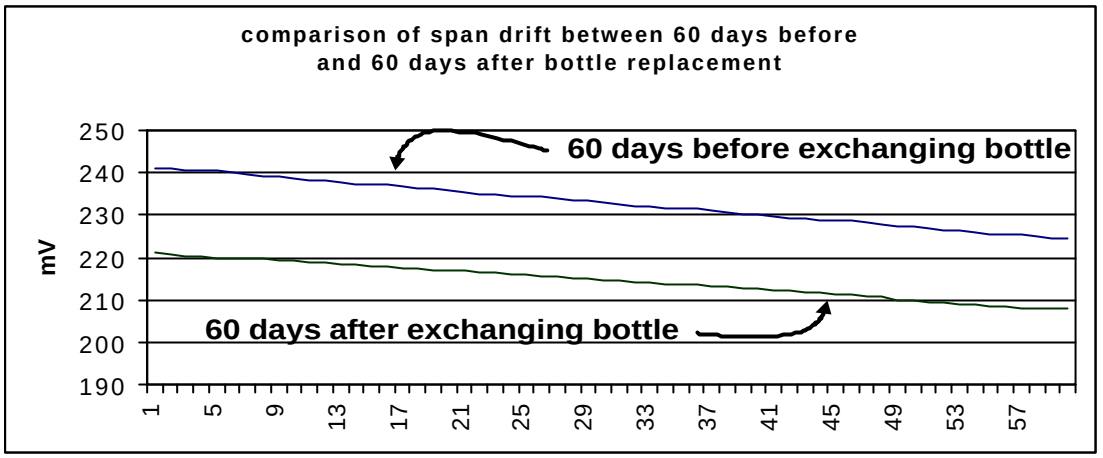


Figure 6: span drift before and after exchanging gas bottles.

Furthermore, the compensation for the drift in gain cannot be calculated from the IZS responses. This is because the permeation device itself also depletes with time. This can be illustrated by normalising the IZS NOx span responses and the NOx span responses to bottled gas over several months. By plotting these on the same graph (figure 7) it will be observed that the drift measured against bottled gas shows less decrease than the permeation tube response. In this comparison the NOx response is used rather than the NO response because the IZS span source is purely NO2, and the bottled gas is also known to have a shifting ratio between NO and NO2 (which is the error to be corrected).

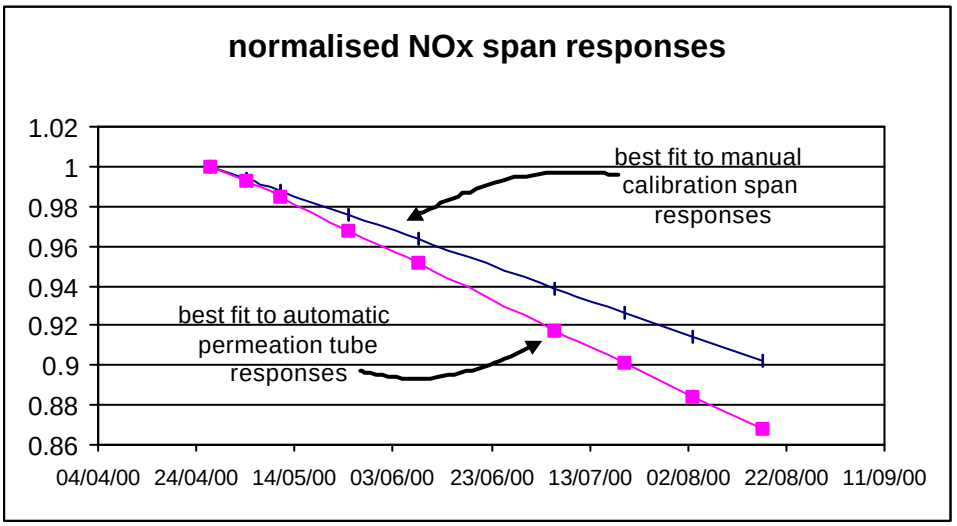


Figure 7: Comparison of the IZS and bottled gas NOx span responses

As the IZS span drift cannot be used to adjust the correction factors for temporal drift, it therefore needs to be established whether the compensation for temporal drift can be calculated from the manual calibration responses to the new, reference gas bottle.

It is known that the NO response to the corrupted bottle and the new bottle will be different, therefore only the NOx responses are considered. The NOx span responses for the 5 calibrations before the bottle was exchanged were normalised. The best fit

trend was calculated using the least squares method (via an Excel spread sheet). This best fit was then extrapolated to give the expected analyser drift in the first two months after changing the bottle.

The response of the next four calibrations, (which were against the new bottle) were then normalised to this extrapolated curve and these actual responses plotted against the prediction (see figure 8). This established that the manual calibrations before and after exchanging the gas bottle exhibit the same rate of drift.

This is an important result as it establishes that the analyser drift can be derived from the NOx response alone, which is unaffected by the changing ration between NO and NO2 in the bottle. This means that the calibrations carried out in the period when the bottle was faulty can still yield valid NOx measurements, even though the NO (and thus NO2) measurements are erroneous. It further establishes that the drift of the analyser with time can be predicted, and therefore compensation for temporal drift can be based upon the reference calibration to the new gas bottle..

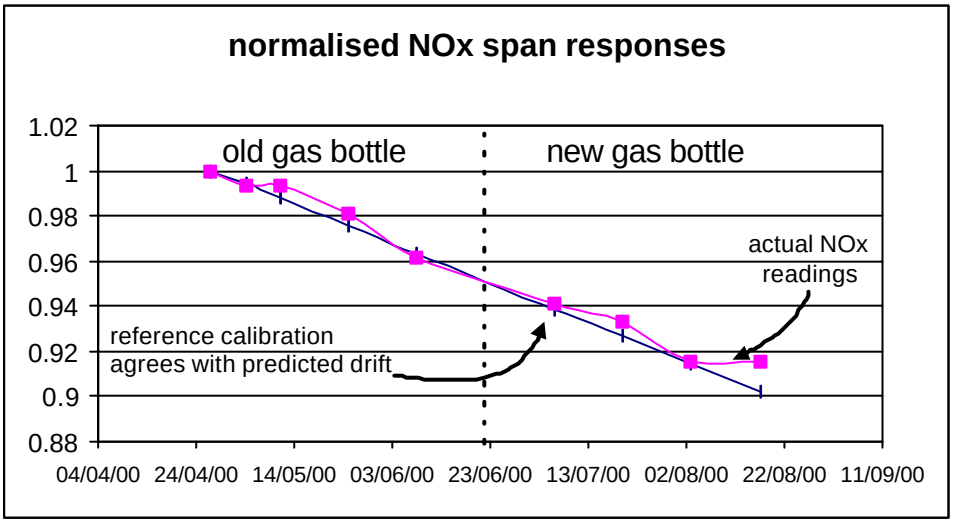


Figure 8: Illustrating that the rate of analyser drift is unchanged in the periods before and after changing the gas bottle, as evidenced by the NOx span response.

From this curve we can calculate compensation factors to compensate for the drift against time.

27/04/00	1.065096
04/05/00	1.058586
11/05/00	1.052077
25/05/00	1.039058
08/06/00	1.026038
22/06/00	1

Revisiting the figures for the 27th April calculation used in the worked example above:

NO gain = 424.99 x 1.065 = 452.62

NO zero = 2.238 x 1.065 = 2.38

NOx gain = 850.20 x 1.065 = 905.46
NOx zero = 3.882 x 1.065 = 4.13

Using Signal's Ambidesk software it is trivial to rescale the data using these correction factors. Figure 9 shows the data before correction (along with about one weeks good data generated after the new calibration bottle was introduced). This graph was produced using Signal's Ambidesk reporter software:

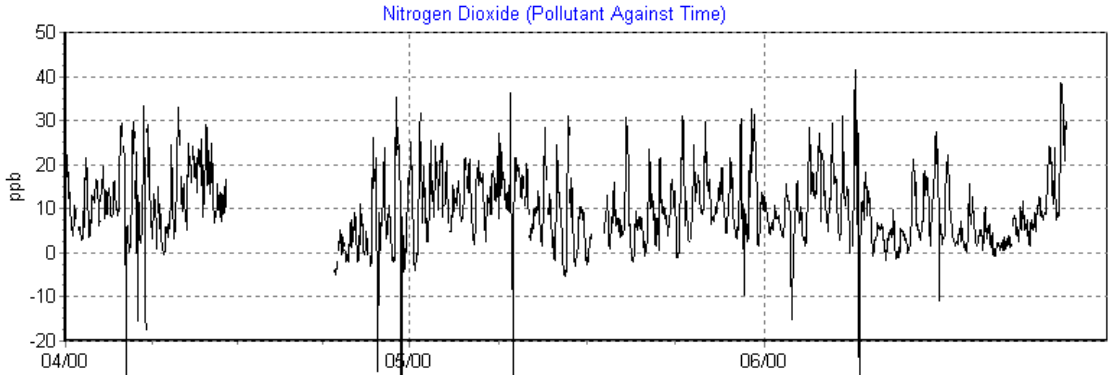


Figure 9: Data before correction

Figure 10 shows the effect of correcting the data. Note that there are now no negative values.

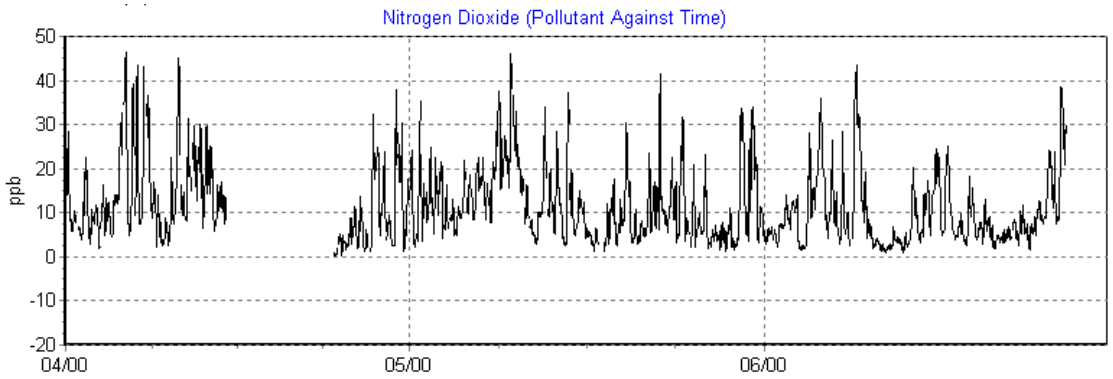


Figure 10: Data after correction

By inspection this data appears much more plausible, however, can we justify this correction more rigorously?

2.3 JUSTIFICATION OF THE DATA MANIPULATION

2.3.1 Intuitive justification

We know that the NO scaling factors are unreliable (and thus so are the calculated NO2 values), however, we have established the NOx scaling factors are correct, even when calculated against the faulty bottle.

We therefore have:

Raw data logs –	one minute averages in raw mV
Processed data logs –	15 minute averages already scaled to faulty bottle
Original calibration factors –	NOx (OK), NO (incorrect)
Adjustment factors –	these rescale the data already scaled to bad gas bottle, and compensate for drift against time.

To check the adjustment factors, a day was selected at random from the period where the data was unreliable and the following, independent calculations were carried out and the results compared:

Raw mV NOx data, was averaged to one hour, then scaled using original NOx calibration factors

Processed (blurred) NOx data, was averaged to one hour, then corrected using adjustment factors.

The correspondence between the resulting two data sets was remarkable, exhibiting an average error of only 0.103% of analyser range. This gives confidence that the algorithm for correcting between the two bottles, and then compensating for drift is robust.

2.3.2 Quantification of improvement using statistical methods:

Verification of the NO2 data requires the selection of an appropriate reference from the available diagnostic measurements. This system logs the raw mV outputs of the analyser separately from the data which has been corrected to the calibration. As the error here is known to be due to calibration, the raw mV data will be unaffected. Furthermore, as the NO2 data is calculated from the difference between the NO and NOx responses, the effect of instrument drift is minimised.

The raw mV NO2 values, prior to any calibration are therefore a suitable reference. Several days were selected at random to establish the validity of this comparison, one example is presented here. The “blurred” data, the corrected data and the reference data are shown as figure 12. In this example the improvement is apparent by inspection, this is for illustrative purposes only, and the technique can be used for less trivial examples!

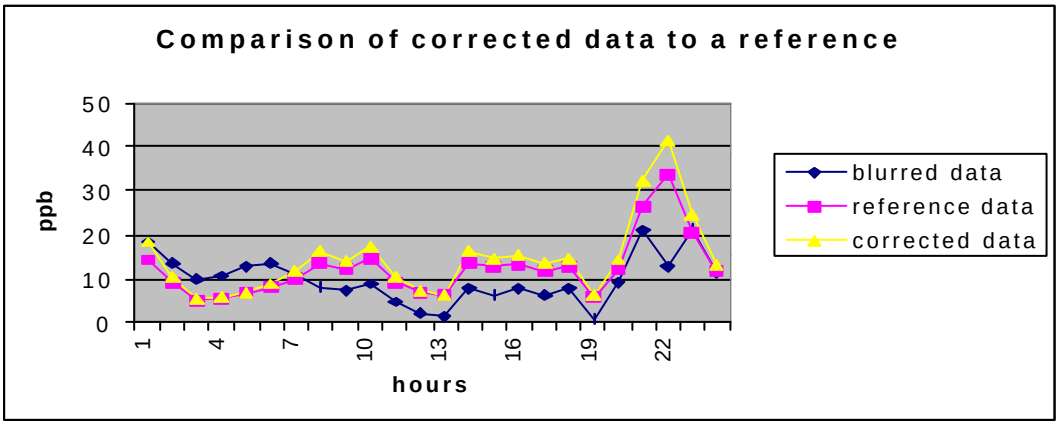


Figure 11: Comparison of blurred, corrected and reference data

Note that the selected data sets, using only one day's data averaged to into one hour averages, are now quite manageable. It should be noted that the built in statistical functions in spreadsheet packages, such as Excel, are usually limited to only 30 data points. If that number of points is exceeded, then errors can be introduced. "*In an examination of blue chip companies using large spreadsheets (more than 150 rows) Coopers and Lybrand found that more than 90% of models contained at least one calculation error. In 21 of 23 models reviewed results were inaccurate by more than 5%.*"**[7]**

The improvement achieved by correction of the blurred data set may be established by simple statistical analysis, and reference to the percentage points of the **t** distribution. The **t** distribution is an adjusted version of the normal distribution which compensates for small data sets. The **t** distribution is usually used to establish whether there the differences between two data sets are statistically significant. However, here we can use it as a metric of how much we have improved the data set, because the degree of difference between the improved data set and the reference data set should be significantly less than the difference between the blurred and reference data sets.

The calculations are simple, particularly when using a spread sheet and small data sets. First calculate the factor **S** from the two standard deviations.

$$S = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where,

σ_1 = standard deviation of data set 1

n_1 = number of samples in data set 1

Next calculate, the factor **F** from the arithmetic means and the factor **S**.

$$F = \frac{\bar{x}_1 - \bar{x}_2}{S}$$

where,

$\bar{x}_1$ = arithmetic mean of data set 1

Were the data set sufficiently large to approximate to the normal distribution, (say more than 150 data points) then the resultant **F** factors could now be looked up in the statistical tables. However, as we are dealing with a small data set, the degree of freedom (**DoF**) must be calculated: this is simple:

$$DoF = n_1 + n_2 - 2$$

This gives a **DoF** of 46 = 24 + 24 –2

Comparing the blurred and the reference data gives:

$F_{br} = 2.25$

Comparing the corrected and reference data gives:

$F_{cr} = 0.93$

We can therefore quantify the improvement in the data set by reference to the table of percentage points of the **t** distribution: **[8]** These tables express the probability that two data sets are NOT related.

α=	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
DoF = 40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460

The reference and blurred data are statistically different with an possibility of error (α) of only 2.5%. After correction, the possibility of error has increased to over 10%. Inverting the logic we may say that there is an over fourfold increase in confidence that the data sets are related.

For this particular data set the improvement could be noted by inspection. This can also be confirmed by calculating the correlation coefficients.

The correlation between the blurred and reference data is 0.534
The correlation between the corrected and reference data is 0.999

Thus a weak correlation has been changed to a very strong correlation. This calculation is also trivial to perform using a spreadsheet.

3 CASE STUDY TWO – THE VALUE OF DIAGNOSTIC LOGS

This case study concerns continuous emissions measurements from a power station in southern Europe. The site operator had been less than diligent in maintaining the equipment, and problems which would have been easy to rectify had been allowed to obtain for several months. The effects of the instrumentation errors are fortunately recorded as diagnostic logs in the PC logger, which facilitates data recovery.

3.1 Diagnosing the problem:

The symptoms presenting themselves to the operator are clearly observed in figure 12. Basically, the NO and NOx readings follow the ambient air temperature.

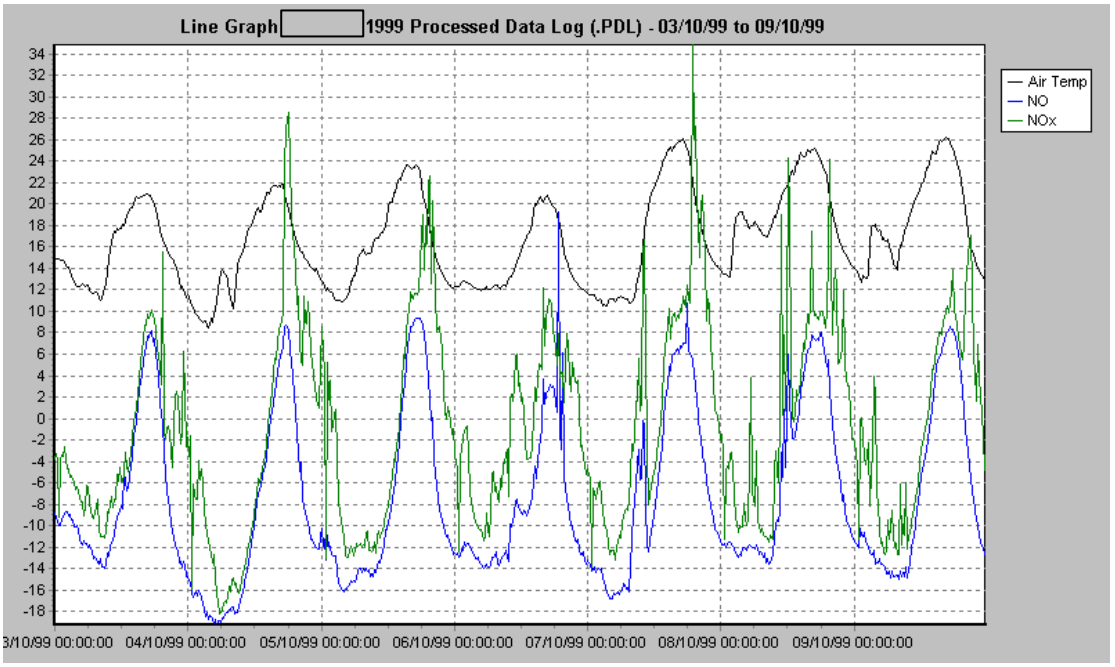


Figure 12: showing correlation between NO/NOx and air temperature

The reason for this was found to be that the diurnal temperature fluctuations were causing a corresponding fluctuation in the temperature of the photo-multiplier tube. See figure 13.

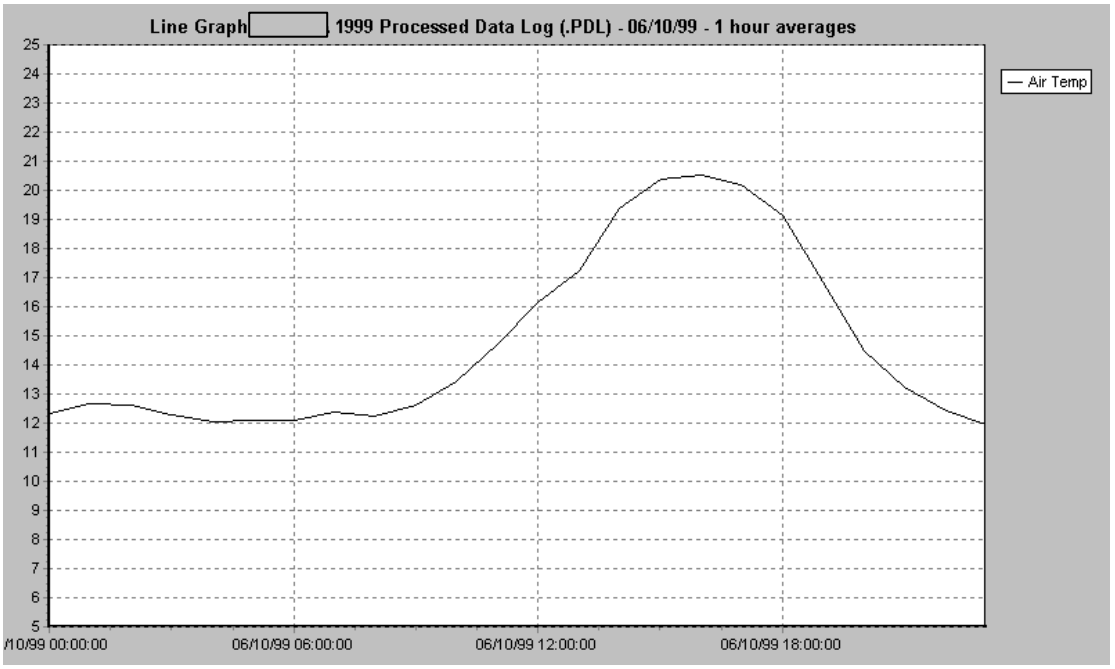


Figure 13(a) diurnal ambient temperature distribution (day selected at random)

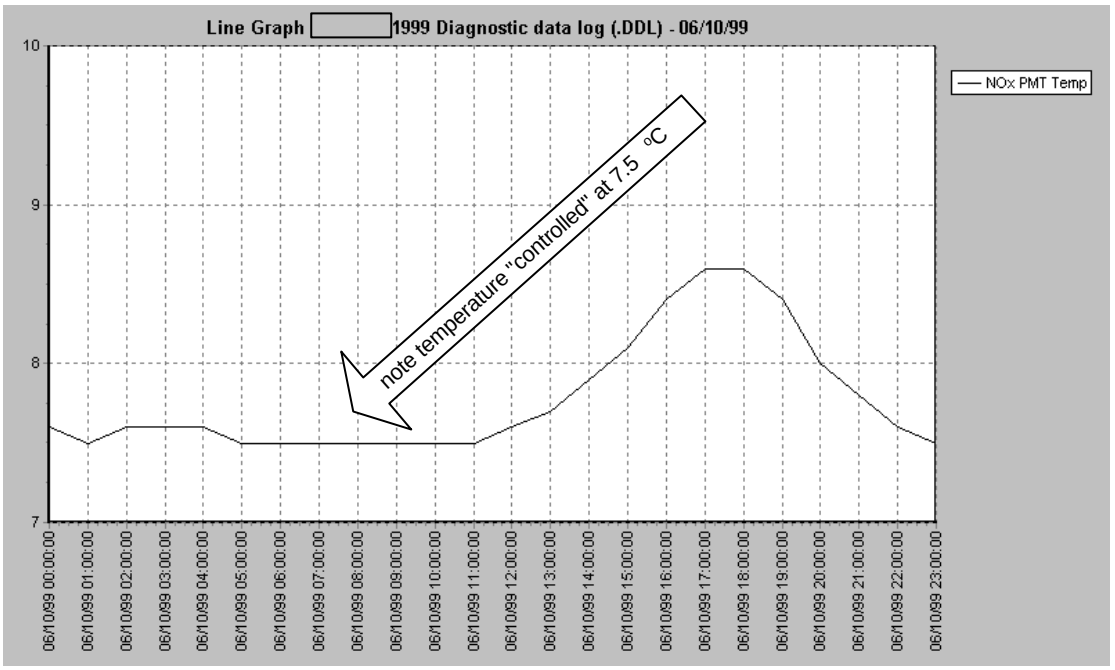


Figure 13(b) diurnal variation of PMT temperature (the same day)

The photo-multiplier tube (PMT) amplifies the light emitted by chemiluminescence, as NO reacts with O3. It is a known feature of the PMT that the zero offset (dark current) linearly follows temperature, however the gain of the PMT is not temperature dependant. The effect of temperature variation therefore effects NO and NOx readings equally, by adjusting the zero baseline.

The variation in temperature should have been well with the capabilities of the instrument's cooler, which was a malfunction.

5.2 Recovering the data

As the cause of the error is well understood, it is also known that the problem will effect the NO and NOx readings equally, therefore the differential NO2 reading is unaffected. For reporting to ambient air standards, where only NO2 is usually referred to, then no further action needs to be taken.

In this case, the NOx readings were also of interest. The route to recovering this data was a bit more involved. However, the operator was prepared to undertake considerable effort to recover the data, in order to satisfy the regulatory authority, which was otherwise threatening to close the plant!

The PMT temperature log is available, and can be used to correct the NO and NOx data, provided the effect of the diurnal temperature variations can be quantified. Observation of this log reveals that when the temperature control is working correctly it controls the PMT temperature at 7.5° C, this can therefore be regarded as a base line for the correction.

This system has two calibration systems. Each day an automatic internal zero and span (IZS) check is performed. The operator also performs periodic manual calibrations. The IZS is performed in the middle of the night, whereas the manual

calibrations are performed in mid afternoon. The temperature variation between the two sets of calibrations is therefore quite pronounced.

The zero stability of the instrument is known to be good, therefore by plotting the variation in NO zero response against the PMT temperature variation, the effect can be quantified. This amounts to 16mV per °C. (See figure 14). A further calibration factor is applied to convert this mV deflection to engineering units.

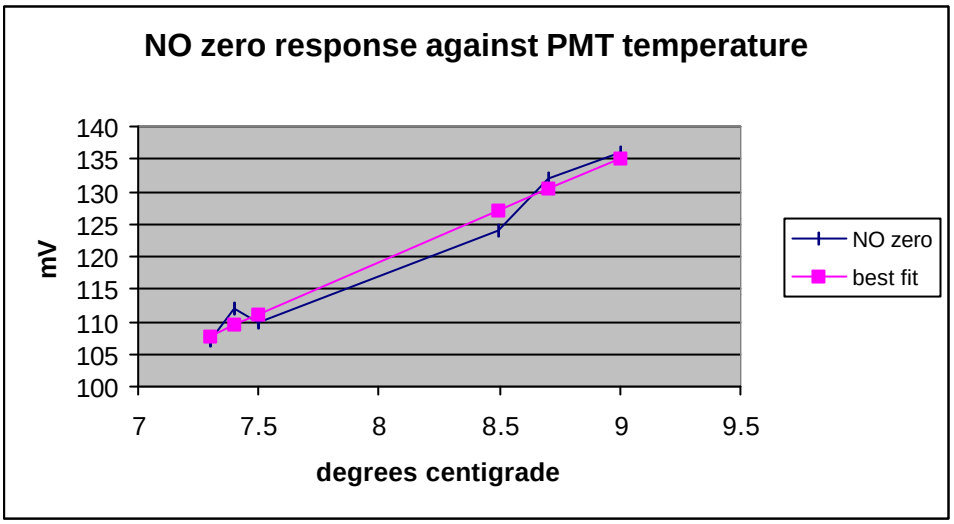


Figure 14**: **NO zero response plotted against PMT temperature

The PMT temperature log was therefore normalised and then scaled into engineering units. The resultant correction was then subtracted from both the NO and NOx logged data. Thus producing a corrected NOx data set for reporting to the regulatory authority.

4 CONCLUSIONS

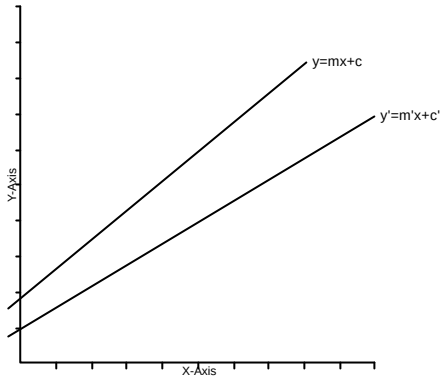
Both of the case studies here are real life incidents, where the data was corrupted, but still bore a quantifiable relationship to the correct measurement. Borrowing the terminology of SEVA technology, this data may be regarded as “blurred”, rather than “blind” or bad.

In both cases the justification of the data correction was accepted by the regulatory authorities.

Recovery of the data, and the justification of that recovery requires:

1. Trained staff, acquainted with simple statistical techniques
2. System and analyser health diagnostics to be logged – easy with a PC logger
3. Appropriate software tools
4. Support from the instrument manufacturers in interpreting diagnostic logs.

APPENDIX A: Derivation of correction factors



To correct one curve:

y=mx+c

to another curve:

y'=m'x+c'

then make y' a function of y: $\frac{y-c}{m}=\frac{y'-c'}{m'}$

so, $y'=\frac{m'}{m}y+c'-\frac{m'}{m}c$

but: $\frac{m'}{m}=\frac{y'-c'}{y-c}$

so, the span correction factor is: $\frac{y'-c'}{y-c}$

and the zero offset correction is: $c'-\frac{y'-c'}{y-c}.c$

where:

y' = span expected
c' = zero expected
y= span actual
c= zero actual

ABOUT THE AUTHOR:
Andrew Newman is a chartered engineer and has a first class honours degree in Digital Systems Engineering from the University of the West of England. He is currently studying part time for an MSc in software engineering at Oxford University, with the aid of a "Year of Engineering Success" Scholarship from the Institution of Electrical Engineers (IEE) Andrew designed and supervised the development of Signal's Ambidesk software which is used in both AQM and CEM applications in the UK and overseas. Ambidesk has recently been adopted to manage the new AQM network in Saudi Arabia by the King Abdulaziz City for Science and Technology.
He has many years experience working as a development engineer for gas analyser manufacturers, firstly with Systech Instruments, and later for Rotork Analysis, who were acquired by the Signal group in 1997

REFERENCES:

1. "*HMIP Technical Guidance Note (Monitoring) M2*", HMSO, 1993

2. See for example, *"*, Bower, J.S., "*Quality Assurance and Control for Air Monitoring*.; JRC ISPRA, Italy, October 1996.

3. Dettmer, R.; "*Self Validation in Process Control"*, IEE Review, July 2000. pp 29-32.

4. See p789 *"Process Control, Instrument Engineers' Handbook, 3rd Edition"* Béla Lipták (Ed.). Butterworth Heinemann, 1995..

5. Dettmer, p31

6. Newman, A. "*Centralised QA/QC for CEM applications using AQM Software"*. Proceedings of the International Conference on Continuous Emissions Monitoring, CEM99.

7. http//www.uk.coopers.com/financialservices/junglebriefing/spmodel.html

8. Murdoch J. and Barnes J. "*Statistical Tables for Science, Engineering, Management and Business Studies – 3rd Edition*". Macmillan, 1986.