CONFORMANCE ASSESSMENT

Conformance assessment of electrical energy meters investigated by risk analysis – a case study

Helge Karlsson, Justervesenet, Norway
Åge Andreas Falnes Olsen, Justervesenet, Norway
Leslie Pendrill, SP, Sweden

Abstract

This paper presents a case study of more effective decision rules for the conformance assessment of electrical energy meters in private households in Norway, and proposes how to use a specific risk analysis in order to set the time for the next meter test. The MID regulation today prescribes conformance assessment of electrical energy meters based on ISO standards for attribute sampling where decision rules are purely statistical decision rules and economic consequences are not explicitly taken into account. The risk analysis we introduce calculates the risks involved for erroneous decisions, either rejecting a conforming batch of meters (the producer risk) or accepting a non-conforming batch (the consumer risk). The consumer risk is sensitive to the period until the next test which becomes a quality characteristic of each batch. This time interval can be optimized by balancing the consumer risk against the producer risk. When the quality drops, the period until the next test will need to become shorter. But at a certain level of quality, the energy net supplier would rather replace the complete batch, than continue testing at such short intervals.

1 Introduction

In 2011, the Norwegian Ministry of Industry issued a press release stating that there would be no further replacement of electrical energy meters in private households in Norway until the introduction of smart meters for electrical energy. At the time, the transition to the modern instruments was scheduled to be completed by 2015, but this deadline was later extended to 2019. The practical implication of the press release was to suspend the testing regime of electrical energy meters, with the consequence that testing laboratories in Norway shut down their activities for these instruments.

The press release cited economic arguments for the decision, pointing out that the estimated replacement cost of a non-conforming instrument was around NOK 2 000. Since all instruments would be replaced within a few years anyway, this decision appeared to save money for the utility companies, and by extension the consumers of electrical power which would have faced a higher utility price when the companies recovered their costs. However, the Ministry of Industry appeared to neglect the possible costs associated with leaving non-conforming instruments on the market to measure consumption. In fact, by postponing the shift to smart meters, the accumulated cost from erroneous measurements potentially increased significantly. Indeed, if one considers the maximum permissible error (*MPE*) of the meter (see Figure 1), and multiplies this figure by a typical consumption level, the annual costs of measurement errors may be as large as NOK 600: extending the time in operation from 4 to 8 years would incur an additional cost of NOK 2400. A non-conforming instrument could represent an even higher cost.

While this simple cost estimate should act as a warning that the consumer costs can be significant, it does not accurately reflect realistic values. Firstly, the interplay between the failure modes of electrical utility meters, short-time fluctuations in the actual consumption, and short-term fluctuations in the pricing means it is difficult to quantify a "typical" consumption level and a "typical" utility price. Secondly, the test regime was based on grouping individual units in batches involving anything from a few tens of instruments to several thousands, sampling a subset of each batch for verification testing, and replacing the entire batch if a prescribed number of sampled units fail the test. Since it is exceedingly unlikely that every unit in a batch measures with the same large error, the batch average measurement error will almost always be smaller than the *MPE* even for rejected batches.

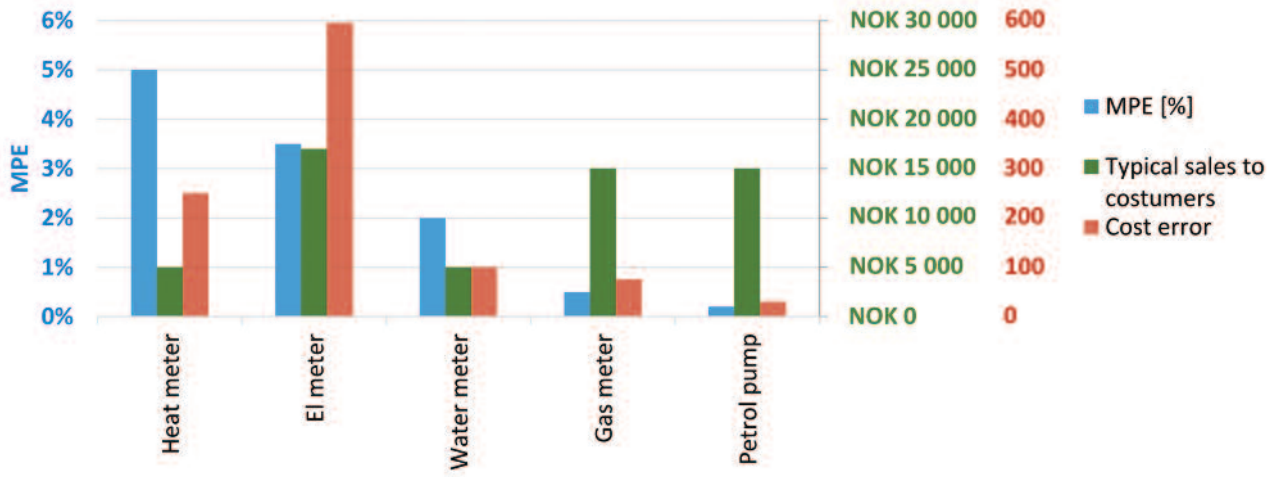

Figure 1: Blue bars: Maximum permissible error (*MPE*) defined in the MID for different types of meter. Green bars: Estimate of typical annual consumption of a private household in NOK, green figures. Red bars: "Maximum cost errors" for different types of utilities in NOK, red figures.

In this paper we present a much more detailed computation of the costs based on a large national database of test results. The database provides details about every sampled unit, such as its measurement errors at various loads, whether the unit failed the test, and whether the batch it belonged to failed the test. In addition we use detailed hourly data about the price of electrical power in Norway, typical hourly consumption profiles, and the actual distribution of household yearly consumption to compute a more realistic cost associated with erroneous measurements.

The consumer cost may then be directly compared with the producer cost (replacement of the entire batch). However, this direct comparison ignores the following issue: Since the verification was carried out using acceptance sampling, there is always a probability that the *actual* error rate (and therefore batch average measurement error) differs from the *observed* error rate. As a consequence the actual cost could have been higher or lower than observed. We propose here that a risk perspective is an attractive angle of attack on this issue. By multiplying the consumer cost with the probability that the actual error rate is greater than a tolerance, given the observed error rate, we obtain a *specific risk* value for the consumer, which may be compared with a corresponding producer risk.

The tolerance used in the probability calculation remains a free parameter. In fact, with acceptance sampling there are a number of parameters which need to be fixed, such as the sample size, acceptance and rejection thresholds, and the period between tests. Each of them affects the test performance, e.g. its cost, or the probabilities and consequences of wrong decisions. While typical acceptance sampling standards, often referred to in existing regulations, rely on probabilistic thresholds to fix these parameters, it is also possible to address the question within a risk analysis framework [1], [2], [3]. In our analysis we use the sample sizes given in the database, and acceptance and rejection thresholds from the Norwegian regulation and Welmec Guide 8.10 [4]. Interestingly, the two standards differ in the thresholds used, which also impacts the balance between specific risks.

The paper is structured as follows: The first part explains the test regime which electrical energy meters were subjected to in Norway, and shows how it compares with Welmec recommendations on an example dataset comprising almost 2 million instruments. The second part discusses how the cost of errors in metering electrical energy may be calculated, and shows how cost alone would affect accept/reject decisions of the dataset. The third part explains how to compute probabilities of non-conformity, and how the calculations may be combined to produce estimates of specific risks. The last part compares the different decision processes.

2 Statistical quality verification of electrical energy meters

The Norwegian regulation quite reasonably places the responsibility for quality assessment of electrical energy meters on the grid owners (producer side) rather than on the consumers. The typical consumer does not have the relevant knowledge or equipment to perform proper testing of their utility meter, and even if they did, they would

have no incentive to report errors in their favour. The regulation prescribes statistical acceptance sampling of uniform batches, with tolerances directed both towards the performance of individual devices, and towards entire batches in terms of the proportion of non-conforming items in the batch. Batches are retested at predetermined intervals, which range from 3 to 10 years.

Sampling plans are chosen so that the Operational Characteristic (OC) curve for the test correctly passes beneath two critical points, the acceptance quality level (AQL) and the rejection quality level, also called the limiting quality level (LQL). AQL is a level of quality corresponding to a probability of acceptance of 95 %, with a non-conformity of less than 1 %: P_a (p < 1 %) = 95 %. LQL is a level of quality corresponding to a probability of acceptance of 5 %, with a non-conformity of less than 7 %: P_a (p < 7 %) = 5 %.

The MID Directive [6] is the basis for regulation of electrical energy meters in Norway. Normative documents EN 50470-1/-2/-3 [7] and OIML R 46-1/-2 [8] describe essential metrological requirements and tests. Taking into account measurement uncertainty [9] in each meter test result, each unit tested is declared to be inside or outside specifications (conforming or nonconforming) according to principles laid down in JCGM 106:2012 [10]. Accept or reject decisions, however, are made on the basis of statistical verification of homogeneous batches of meters rather than treating them as individual units, and statistical sampling uncertainties associated with limited sampling also needs to be accounted for. Requirements for statistical verification are described in Welmec 8.10 [4] and ISO 2859-1/-2 [11] using either a single sampling plan or a double sampling plan. When the number of non-conformances, M, is very small or very large, a double sampling plan is more efficient than a single sampling plan, because a conclusion is often clear already after the first part, effectively reducing the sample size, n. When the number of nonconforming units fall between critical values M_{Ac} and M_{Re}, the second part of the sampling plan is carried out. We will use index 1 or 2 for M, n and p to indicate the first or second part of a double sampling plan.

Table 1: Acceptance and rejection thresholds in the Norwegian regulation for verification of electrical energy meters. Indexes refer to the first and second part of the sampling plan.

Batch size	n_1	n_2	M_{Ac1}	M_{Re1}	M_{Ac2}	M_{Re2}
65 - 1200	32	64	0	2	1	2
1201 - 3200	50	100	1	4	4	5
3201 - 10000	80	160	2	5	6	7
10001 - 35000	125	250	5	9	12	13

Welmec Guide 8.10 [4] provides guidance to manufacturers of measuring instruments and to Notified Bodies responsible for conformity assessment of their products. The sampling plans chosen in the Norwegian regulation differ from examples given in the Guide as "best practice": The examples in Welmec Guide 8.10 are stricter than the sampling plans in the Norwegian regulation.

In order to illustrate the difference between the Welmec recommendation and the Norwegian regulation, we have analysed data in the national test result database. It contains values for more than 60 000 sampled electrical energy meters, representing roughly 2.2 million meters in Norway. We have limited our attention to batches of size $N \geq 200$, which comprise test results for 24 600 sampled mechanical type of electrical energy meters (divided into 502 batches, representing 910 933 devices), and 17 619 electronic type of electrical energy meters (sampled from 254 batches, representing 996 264 devices). The overall error rate ($\sum M_i / \sum n_i$) is 3.57 % for mechanical devices and 0.51 % for electronic devices.

For each batch we have extracted the number of tested units, the number of units in the batch, and the number of units which failed the test (i.e. exceeded the *MPE*). The numbers provide a best estimate of the failure rate in each batch. This estimate is then used to extrapolate the expected number of failing devices in a sample of a different size than that actually used: according to the Norwegian regulation on the one hand and the Welmec Guide on the other. The extrapolated number of failing devices is then compared with the prescribed threshold number for a rejection. Table 2 summarizes the results.

Table 2 illustrates how sensitive the test conclusions are with respect to the sampling scheme parameters. The Welmec recommendation results in many more rejections. The difference is particularly striking for electronic devices, where the number of rejected batches is significantly higher than the corresponding numbers using the sampling parameters from the Norwegian regulation.

Table 2: The number of failing batches, and the number of units they represent, for sampling parameters using Welmec Guide 8.10 and the Norwegian regulation. The average accumulated cost is the cost of the measurement errors for a batch if left to measure for 8 years, weighted by the size of the batches. See the cost section for more.

Observation	Welmec Guide	Norwegian regulation
Mechanical devices		
# batches accepted	360	383
# batches rejected	142	119
Failure rate, batches	28.3 %	23.7 %
# units in failing batches	48 483	36 791
# units per failing batch (average), *N*	341	309
Electronic devices		
# batches accepted	229	248
# batches rejected	25	6
Failure rate, batches	9.8 %	2.4 %
# units in failing batches	78 189	19 246
# units per failing batch (average), *N*	3 128	3 208

3 Cost of measurement errors

The cost errors of two different decisions

The cost of rejecting a batch and replacing the meters is the purchase of a new device, typically NOK 500, plus the cost of installation, typically NOK 1 500 (in 2011) multiplied by the number of items in the batch which are to be replaced. If this decision was an erroneous decision, this cost of replacement is the producer cost error, Δc_r.

If the decision is to accept the batch and continue to measure with these measurement errors, we calculate the accumulated consumer cost errors due to actual measurement errors for all units in the batch. The annual cost error for a utility meter is dependent on the measurement errors, as well as typical electrical power profiles of a household (frequency of use), and the actual price profile through the year, 8760 hours in one year (spot price and grid fee). The annual cost for instrument number i in the batch is calculated from equation 1:

$$c_i = c_i(W_i, P_i, e_i) = \sum_{q=1}^{8760} P_{iq}\left[W_{iq}\left(1 + \frac{e_{iq}}{100}\right)\right] \quad (1)$$

where:

W_q = the actual electric power during hour number q
e_q = measurement error in % during hour number q
P_q = price of electrical energy during hour number q

The annual cost error for instrument number i is given by Δc_i:

$$\Delta c_i = c_i(W_i, P_i, e_i) - c_i(W_i, P_i, 0) \quad (2)$$

We have calculated the annual cost error for a random meter in a particular batch with a Monte Carlo simulation for a large number of randomly selected sets of data. These sets of data are:

1) actual measurement errors for a randomly selected measuring device contained in the sample;
2) actual frequency of use, one hour resolution, randomly drawn from a set of typical power profiles (this curve is scaled to match a randomly drawn annual energy consumption from the distribution of annual consumptions of electric energy); and
3) price of utility during one year, one hour resolution.

Repeating calculations of equation (2) for 2×10^6 randomly selected triplets of data sets (measurement errors, scaled power profiles and price profiles) results in a distribution of the annual cost errors for a random meter in the batch. From Hafslund Energy, we have received 38 qualified power profiles for the period August 2011 – July 2012, and also the distribution of annual electrical energy consumption from more than 500 000 private households in eastern Norway, including Oslo. The Oslo spot price of electrical energy for the same period is taken from http://www.nordpoolspot.com.

This calculation treats negative and positive measurement errors within the same utility meter fairly; the costs cancel out depending on the size of measurement error, frequency of use and price. A positively signed error is to the benefit of the producer, a negatively signed error is to the benefit of the consumer. Both signs are equally important for *MPE* in the regulation, so this is also the case for cost errors. Independently of sign, the average annual cost error due to measurement errors for a randomly selected meter in the batch is calculated by equation (3):

$$\Delta c_m = \frac{1}{n}\sum_{i=1}^{n} |\Delta c_i| \tag{3}$$

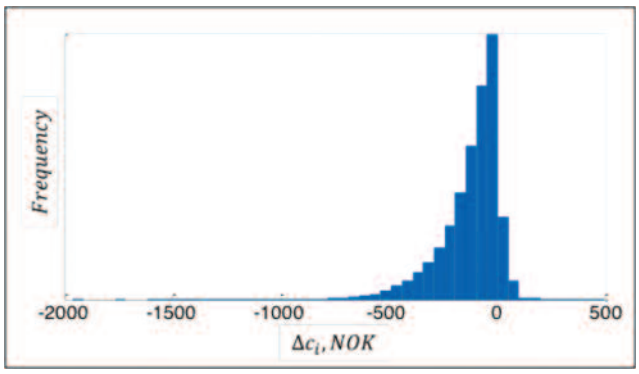

Figure 2: Results of a Monte Carlo simulation of cost errors, Δc_i, due to measurement errors in electric energy meters. A negative cost error is to the consumers benefit, positive cost error is to the producers benefit.

Figure 2 displays a typical result of a Monte Carlo simulation of annual cost errors for a random meter in a particular batch of meters. The distribution has a maximum close to zero, slightly on the negative side. It is also skewed to the negative side.

Taking only a cost perspective into consideration for the decision, a simple criterion for rejection of a batch is that the 8 year accumulated average excessive cost due to erroneous measurements is larger than the average cost of replacement, NOK 2 000 per item. Figure 3 and Table 3 summarize the results for mechanical and electronic electrical energy meters based on cost evaluation only.

We have so far considered the rejection ratio for essentially two separate methods: the traditional acceptance sampling method, which can be seen as a purely probabilistic approach, and a comparison between the cost of replacement on the producer side and the accumulated cost of measurement errors on the consumer side.

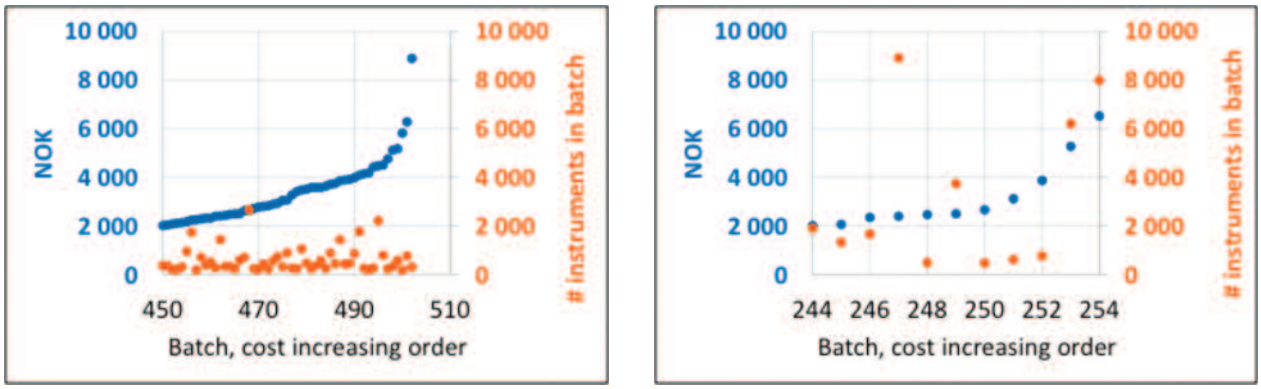

Figure 3: 8 year accumulated average cost for mechanical type (left) and electronic type (right) of electrical energy meter. 53 batches of mechanical meters and 11 batches of electronic meters have average cost error > NOK 2 000,-.

Table 3: Rejection of batches based on costs only.

'Cost only'	Mech.	Electr.
Number of batches with Cost, 8 years accumulated < NOK 2 000	449 batches	243 batches
Number of batches with Cost, 8 years accumulated > NOK 2 000	53 batches	11 batches
Fail rate, batches	10.6 %	4.3 %
Number of instruments in rejected batches	32 491	34 101
Average size of rejected batches, *N*	613	3 100

4 Specific risk analysis

The computed cost is an average whose value is based on the assumption that the observed measurement errors are representative of the actual performance of all units in the batch. We can relax this assumption by feeding the cost of measurement errors into a calculation of a specific consumer risk given the observations. A similar calculation of the specific producer risk allows a new rejection criterion based on a risk balance.

The specific risk of an erroneous decision from the producer point of view is the probability that a conforming batch is rejected multiplied by the error cost of replacing the entire batch with new electrical energy meters. The specific risk faced by the consumer is the probability that a non-conforming batch is accepted multiplied by the error cost of a continuation of erroneous measurements.

Computing the probabilities of conformance and non-conformance is non-trivial and explained below. We assume first that the probabilities are determined by the batch sampling, which appears to neglect the measurement uncertainty in the laboratory testing. However, because the measurement capability in the laboratories is high, at least greater than 10, we can safely ignore this contribution. The observed number of non-conforming units in a batch, M, and the sample size from the batch, n, provides an observed error rate $\hat{p} = M/n$. But the true error rate p could differ; in fact, given the observations the probability distribution of p is the normalized binomial distribution

$$f(p; n, M) = const \cdot p^M (1-p)^{n-M} = \frac{p^M (1-p)^{n-M}}{\int_0^1 u^M (1-u)^{n-M}\, du} \tag{4}$$

This expression is the beta-distribution with form factors α and β given by:

$$\alpha = M + 1 \tag{5}$$

$$\beta = n - M + 1 \tag{6}$$

A few examples of the shape of $f(p; n, M)$ are shown in Figure 5, in which two important features are highlighted. Firstly, the peak of the distribution is at $\hat{p}$ regardless of the sample size; secondly, for small $\hat{p}$ or small n the distribution is asymmetric with a long tail for high true rates p. The latter feature is particularly important in our case because small values of both $\hat{p}$ and n is a common scenario in assessing electrical energy meters.

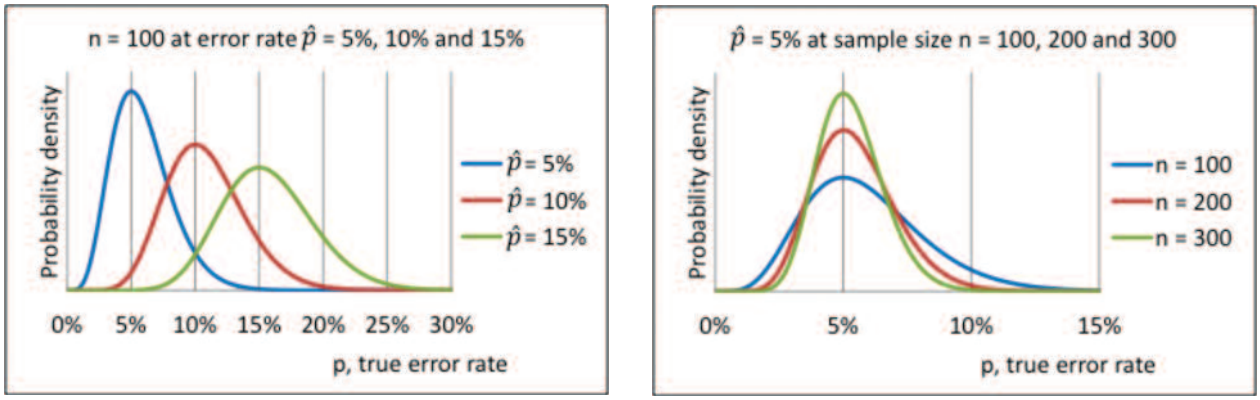

Figure 4: Left: Probability density distributions for the true error rate, p, at different observations of the error rate, $\hat{p}$ and sample sizes, n.

Finding the probability of conformance $P_C(\hat{p})$ and non-conformance $P_{NC}(\hat{p})$ is now straightforwardly achieved by integrating $f(p; n, M)$ between appropriate limits:

$$P_C(\hat{p}) = \int_0^{p_{Accept}} f(p;\hat{p})dp \tag{7}$$

$$P_{NC}(\hat{p}) = \int_{p_{Reject}}^{1} f(p;\hat{p})dp \tag{8}$$

The integration limits are in principle computed from the standardized values of rejection and acceptance thresholds as discussed in section 2. Figure 6 illustrates the integrals for a hypothetical case where the initial sample size is $n_1 = 50$ and the number of non-conforming units is $M_1 = 2$. The probabilities $P_C(\hat{p})$ and $P_{NC}(\hat{p})$ are shown as green and red areas, respectively, with the limits of integration taken from Table 4. The second part of the sampling plan is invoked in this case, with another two devices failing: the resulting integrals for $n_2 = 100$ and $M_2 = 4$ are also shown.

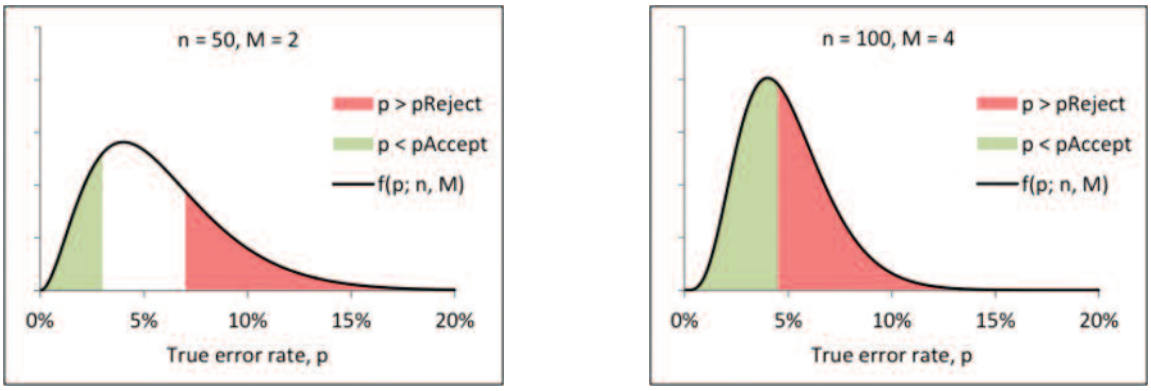

Figure 5: In this example $n_1 = 50$ and $M_1 = 2$, $\hat{p}_1 = M_1/n_1 = 4.0$ % in the first part of the sampling plan, and $n_2 = 100$ and $M_2 = 4$, $\hat{p}_2 = M_2/n_2 = 4.0$ % in the second part of the sampling plan. The probability of conformance, P_C, is indicated by the green area, and the probability for non-conformance, P_{NC}, is indicated by the red area.

The random sampling leads to a coarse resolution for the observed error rate of $1/n$. To adjust for the fact that the actual error rate can take on values also between adjacent values of observed error rates, we will use lower and upper watershed specification limits introduced by D.J. Wheeler [12]. The watershed specifications for p_{Ac} and p_{Re} are calculated from M_{AC} and M_{Re}:

$$p_{Ac} = (M_{AC} + 0.5)/n$$
$$p_{Re} = (M_{Re} + 0.5)/n$$

Transforming acceptance and rejection criteria for M in Table 1 into watershed specifications for p, rounded to two decimals, gives the following table for our sampling plans, see Table 2. There is no alternative to either accept or reject the batch after the second part of the sampling plan is finished, and this is seen in the table by the fact that $p_{Ac} = p_{Re}$ for the second part of the sampling plans.

We may now compute the probabilities $P_C(\hat{p})$ and $P_{NC}(\hat{p})$ for different values of observed M. Figure 6 shows an example for different values of the observed error rate, $\hat{p} = M/n$ for the same sample sizes as before ($n_1 = 50$ and $n_2 = 100$). As the observed error rate increases, the probability of conformance decreases while the probability of non-conformance increases. At some point the two curves intersect, providing a new, purely probabilistic batch rejection condition: $P_{NC}(\hat{p}) > P_C(\hat{p})$.

At the observed error rate $\hat{p}_1 = 2/50 = 4.0$ % and $\hat{p}_2 = 4/100 = 4.0$ %, the probability of non-conformance is higher than the probability of conformance. We would reject this batch at such observations of M. In contrast, the Norwegian regulation rejects the batch at $M_2 = 5$, ($\hat{p}_2 = 5$ %). In fact, we may perform a similar calculation for all batch sizes referred to in Table 4 and compare which method is more strict; the results are summarized in Table 5.

Table 4: Sampling plans in the Norwegian regulation with watershed specifications.

Batch size	n_1	n_2	M_{Ac1}	M_{Re1}	M_{Ac2}	M_{Re2}	p_{Ac1}	p_{Re1}	p_{Ac2}	p_{Re2}
65 - 1200	32	64	0	2	1	2	1.56 %	4.69 %	2.43 %	2.43 %
1201 - 3200	50	100	1	4	4	5	3.00 %	7.00 %	4.50 %	4.50 %
3201 - 10000	80	160	2	5	6	7	3.13 %	5.63 %	4.06 %	4.06 %
10001 - 35000	125	250	5	9	12	13	4.40 %	6.80 %	5.00 %	5.00 %

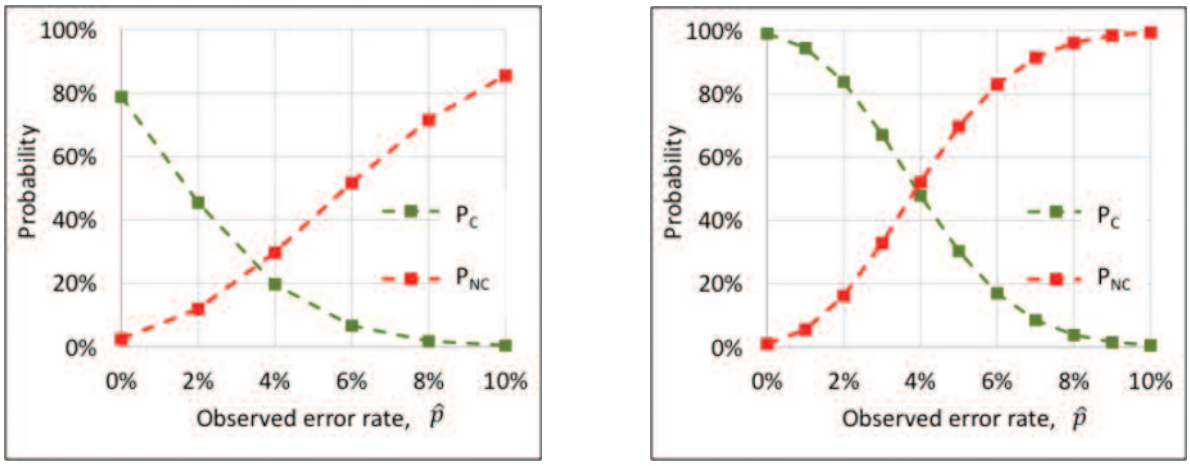

Figure 6: P_C (green curve) and P_{NC} (red curve) for different observations of the error rate, $\hat{p} = M/n$. Left: First part of the sampling plan, $n_1 = 50$. Right: Second part of the same sampling plan, $n_2 = 100$.

The Norwegian regulation allows more non-conforming units than a pure comparison of probabilities for conformance / non-conformance for all sampling plans, except for the second part of the smallest batch, where the Norwegian sampling plan is just slightly stricter than the probability perspective calculated from the observed error rate.

Table 6 summarizes the results for mechanical and electronic electrical energy meters based on probabilities only.

Table 5: Red colour: Norwegian regulation is stricter than the probability perspective.
Blue colour: Norwegian regulation is less strict than the probability perspective.
Green colour: Norwegian regulation is similar to the probability perspective.

Batch size	n_1	n_2	M_{Ac1}	M_{Re1}	M_{Ac2}	M_{Re2}
65 – 1 200	32	64	0	2	1	2
1 201 – 3 200	50	100	1	4	4	5
3 201 – 10 000	80	160	2	5	6	7
30 001 – 35 000	125	250	5	9	12	13

Table 6: Rejection of batches based on probabilities only.

'Pure probabilities'	Mech.	Electr.
Number of batches with $P_{NC} < P_C$ (accepted batches)	169 batches	131 batches
Number of batches with $P_{NC} > P_C$ (rejected batches)	333 batches	123 batches
Fail rate, batches	66.3 %	48.4 %
Number of instruments in rejected batches	352 506	206 256
Average size of rejected batches, N	1 059	1 677

Cost risk curves

The probability curves, as illustrated in Figure 6, can be used to compute corresponding risk curves. The probability of a conforming batch, given the observations, is multiplied with the cost of device replacement to give a producer side risk. Similarly the probability of a non-conforming batch is multiplied with the cost of measurement errors to give a consumer side risk. Figure 7 shows the risk curves for the same example as shown in Figure 6 assuming an annual consumer side cost of NOK 133 with 8 years continued operation (8 × NOK 133 = NOK 1 064). The observed error rate is a function of M, $\hat{p} = M/n$, and to emphasize that our cost risk curves are functions of the *observed* error rate, we display cost risk curves directly as functions of M, see Figure 7.

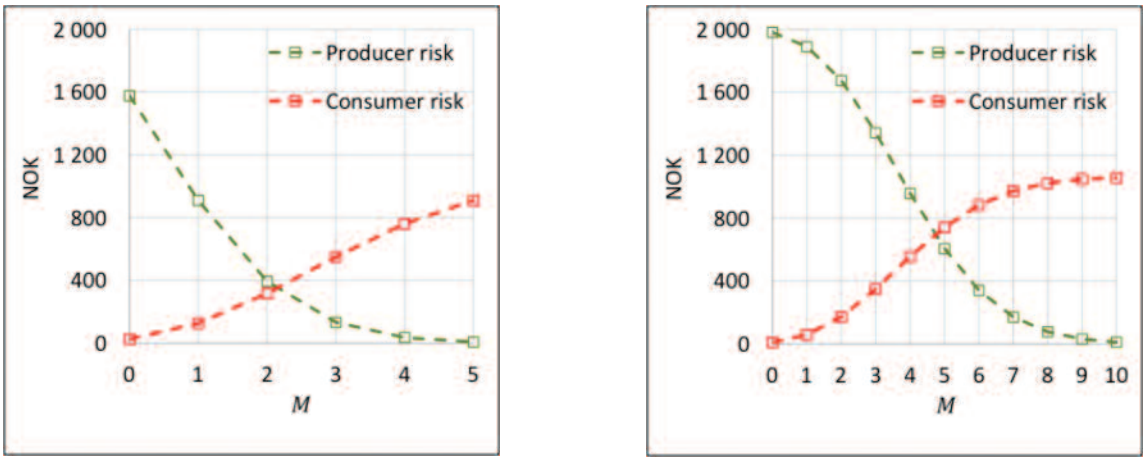

Figure 7: Average risk per instrument for producer (green curves) and consumer (red curves) for the first (n_1 = 50) and second (n_2 = 100) part of the sampling plan using watershed specifications. We have assumed average cost NOK 2 000 for replacing a meter, and average annual invoice error of NOK 133 and 8 years until next test.

We see that for our example of n_1 = 50 and n_2 = 100, the specific cost risk is higher for the producer than the consumer at M_1 = 2 ($\hat{p}_1$ = 4 %) and M_2 = 4 ($\hat{p}_2$ = 4 %). We avoid the higher risk if we decide to accept the batch at such observations of M: hence, adding in the cost modifies the test decision we would have taken from a purely probabilistic point of view.

Increasing the replacement cost shifts the producer risk curve upwards, reinforcing the decision to accept the batch. On the other hand, if the cost of measurement errors were higher (due to larger measurement errors, a higher utility price, or a longer time until the next test) the consumer risk might rise above the producer risk at M_2 = 4.
The probabilities P_C and P_{NC} are computed for each batch in the database using information on sample size n, number of rejections M, and which part of the sampling plan was used, 1 or 2. The consumer side cost is calculated for each batch individually based on the actual measurement errors registered for each sample, see Annex "The dataset, computation of costs". Table 7 summarizes the results for our data, which should be compared to the figures in Table 6. It is obvious that including the costs modifies the decisions significantly, and drastically decreases the number of rejected batches.

Table 7: Rejection of batches based on cost risk.

'Specific cost risk'	Mech.	Electr.
Number of batches with Cons cost risk < Prod cost risk (accepted)	326 batches	218 batches
Number of batches with Cons cost risk > Prod cost risk (rejected)	176 batches	36 batches
Fail rate, batches	35.1 %	14.2 %
Number of instruments in rejected batches	232 209	106 004
Average size of rejected batches, N	1 319	2 945

5 Discussion

The previous sections describe different ways to analyse the outcome of sampling. The standard recommendation in Welmec and the existing Norwegian regulation prescribe a certain statistical sample size and establish rejection rules by requiring tolerance thresholds for non-conforming units given certain %-confidence thresholds (expressed as the AQL and LQL levels described in section 2). The emphasis is placed on reducing the probability of an erroneous decision in favour of either side below certain limits. From the statistical sampling, one can use the prescribed sample sizes and the observed errors in sampled instruments to compute a cost associated with the measurement errors, as described in section 3. The emphasis is shifted to a purely economic analysis, completely disregarding the possibility that the sampled items do not accurately represent the batch. Finally, section 4 first describes a simple criterion based on the probability that the true error rate in the batch, given the observations, is smaller than or greater than a predefined value (in our case we use prescribed numbers from the Norwegian regulations). We then merge this probability perspective with the economic perspective by constructing a risk parameter as the product of the probability of an erroneous decision with the cost of the decision.

Figures 8 and 9 summarise the results in the preceding sections. The number of rejected batches (Figure 8) is strongly dependent on the analysis method, with a similar tendency for the number of units the batches represent (Figure 9). Regarding the two traditional sampling plans (left pane in both figures) the Welmec recommendations reject more batches (and units) than the Norwegian regulations. The difference arises from the fact that the Welmec recommendations require a very small error rate among units in operation (a 95 % or higher probability of accepting the batch if the actual error rate is below 1 %). The Norwegian regulations, on the other hand, emphasize a high probability of rejecting the batch above a certain error rate (5 % or lower probability of accepting the batch if the actual error rate is above 7 %).

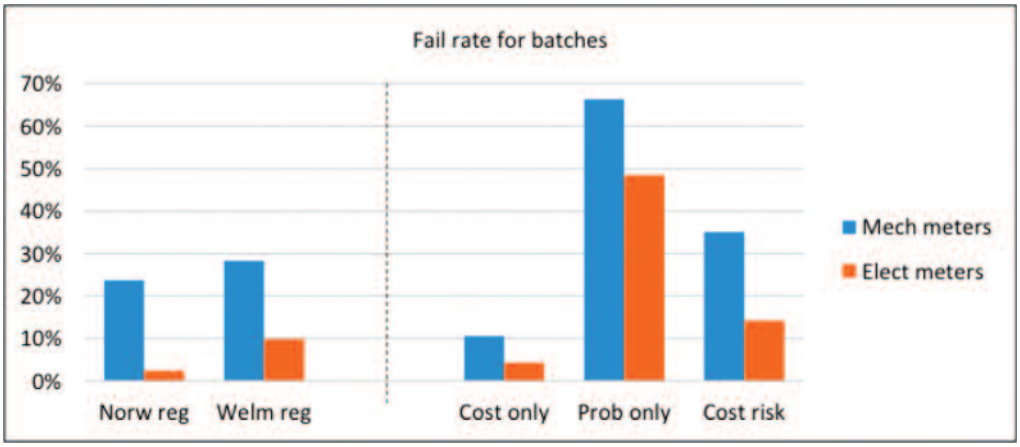

Figure 8: Proportion of batches failing the requirement according to different methods of analysis: Norwegian sampling plans, Welmec sampling plans, cost only, probability only and specific cost risk.

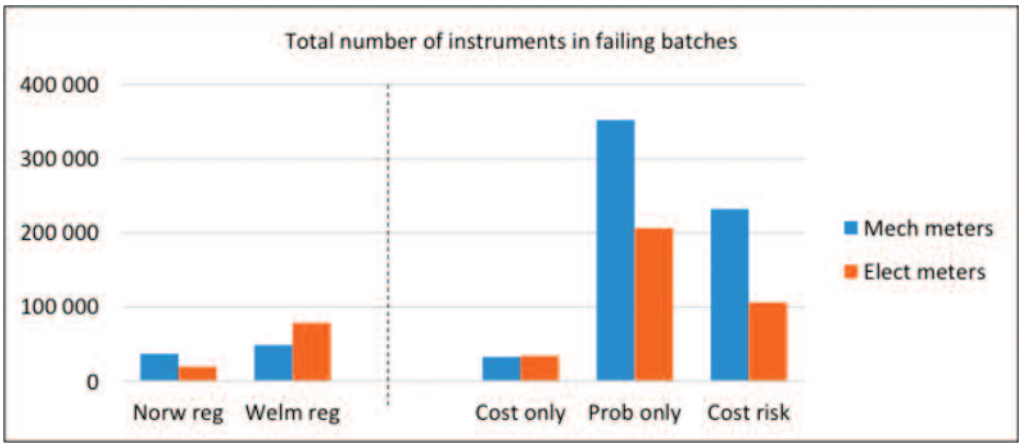

Figure 9: The total number of instruments in those batches which failed according to different methods of analysis: Norwegian sampling plans, Welmec sampling plans, cost only, probability only and specific cost risk.

It is conceivable that the cost calculation could have modified not only the number of failing batches, but also which batches were rejected. Figure 10 graphically represents the sets of rejected batches according to each method, where the overlap between methods is indicated by the common area of the geometric shapes, which represent each method. With a few exceptions a stricter method will just add more batches to the rejection list rather than select a completely different set. For example, using the risk balance method will reject more batches than the Welmec recommendation, but all the batches rejected by the latter are also rejected by the former. We can thus adequately characterize the methods according to the rejection ratio, which we will call *strictness*.

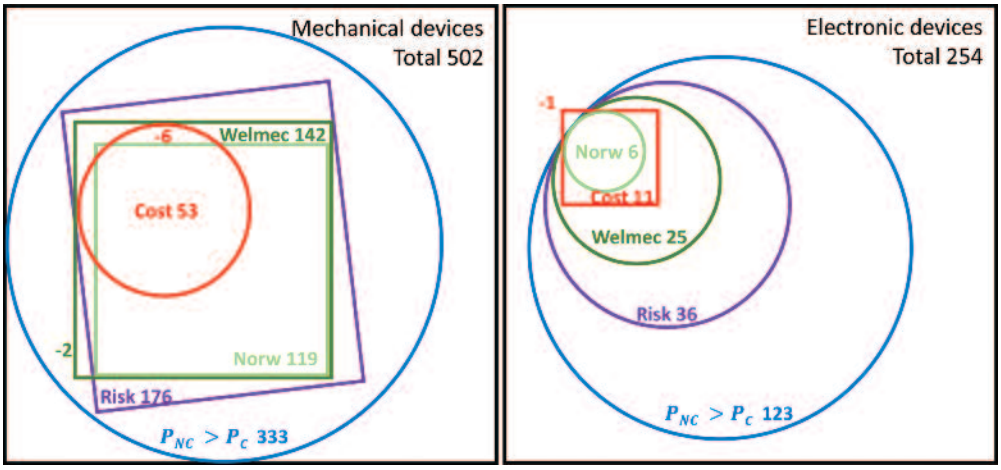

Figure 10: The number of rejected batches using different methods of analysis, mechanical devices on the left, and electronic devices on the right. Areas representing the number of rejected batches are drawn to proportional sizes. The overlap between different shapes indicates the set overlap between methods, so a smaller shape completely contained within a larger contains only batches common to both sets. Negative numbers indicate the number of failing batches *not* contained in the stricter method.
Light green: Norwegian regulation;
Green: Welmec sampling plans;
Red: Cost of 8 years continued use > NOK 2 000,-;
Blue: Probability, $P_{NC} > P_C$;
Purple: Risk analysis, consumer cost risk > producer cost risk;
Black: Total number of batches investigated.

The strictness clearly varies greatly between the rejection criteria, and yet the observations so far do not offer any guidance as to which method is most appropriate. In fact, there are a number of other parameters which could have been varied, such as the period between tests, the number of sampled units, the acceptance/rejection thresholds, and the error tolerance (*MPE*). Fixing these values requires a wider perspective, and the choices will place different emphasis on the grid owner or consumer protection. Traditional thinking tries to establish combinations of sample sizes and rejection limits from the probabilities of true error rate and desired confidence limits by phrasing the problem in terms of AQL and LQL levels (section 2). Pendrill ([1], [2], [3]) adopts a different approach by optimizing sample sizes with respect to a test cost and sampling uncertainty, thus bringing in yet another aspect of the verification assessment.

We would argue that once the measurement devices are put into operation, and the required sampling sizes for conformance assessment have been determined, the cost of replacement of units must impact the accept/reject decision. However, as shown by Figure 10 the two ways to include the cost (specific risk and pure cost) have very different strictness. In fact, it seems that the Norwegian regulation better matches a pure cost balance, while the Welmec recommendation better matches the method of risk balance. If one takes into account the probability that the true error rate differs from the observed error rate (essentially what the risk analysis does) we arrive at a practical rejection behaviour, which matches the Welmec recommendations best. However, with the sample sizes used, the authors of the Welmec Guide were forced to use values for M_{Ac} which imply a much smaller error rate than the AQL level of 1 %. This is appropriate in conformity assessment for type approval, but once the devices are in operation the additional cost represented by batch replacement requires a better producer side protection. The Norwegian regulations compensates this by relaxing the M_{Ac}; however, at the cost of consumer protection. A different strategy would be to ignore test costs and take larger samples in each test: while this would increase the test cost it would also decrease the probability of unwarranted batch replacement due to sampling uncertainty.

The calculation of consumer cost is tedious and demanding, and needs access to data other than the unit test measurements. Compiling a set of fixed rejection thresholds, as traditionally done, is a much simpler method to implement in a test laboratory. However, the biggest hurdle to basing the decision on a cost calculation is the sensitivity to the time to the next test. The consumer risk can be reduced simply by testing more frequently, and in an extreme case a grid owner could avoid unit replacements simply by deciding to perform another test within a short enough period of time. The risk analysis offers no guidance about how to handle this conundrum.

Rather than using a fixed period to the next test to provide the acceptance thresholds we could adopt the acceptance thresholds from elsewhere (e.g. Welmec) and compute the period until the next test instead. Such a test

regime not only circumvents the test frequency conundrum, but would also retain the attractive ease of use in tabulated acceptance thresholds while also taking a cost perspective into account, which is updated according to the observations after a test. The time to the next test, τ, may be defined as the ratio between the producer risk and the *annual* consumer risk, (see section 4, with the annual cost of measurement errors in place of the accumulated cost of measurement errors):

$$\tau = \frac{P_{NC}}{P_C} \cdot \frac{\Delta c_r}{\Delta c_m} \tag{9}$$

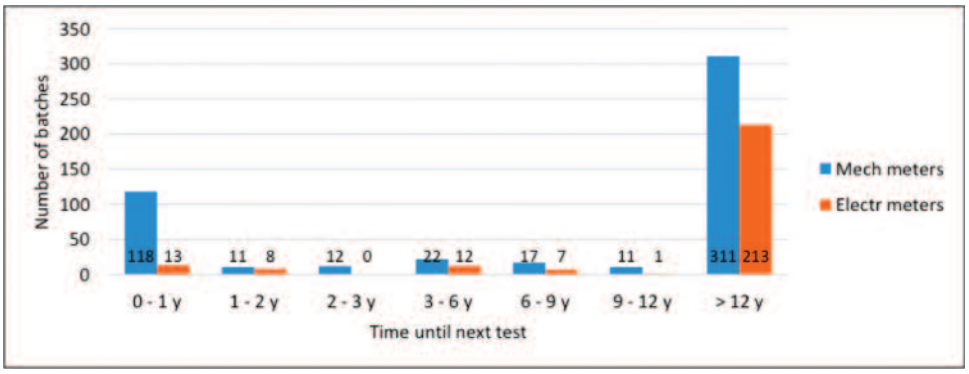

Figure 11: Time until the next test for each batch calculated from equation (9); acceptance criteria are from the Norwegian sampling plans.

We have computed τ for our data, and the results are shown in Figure 11. Most batches have $\tau > 12$ years. A smaller cluster of batches have a very short τ of less than a year, while a small, but significant number fall in between. While a very short period until the next test coincides with rejection, there can be cases where its value is unreasonably long. There is reason to expect units to have a finite lifetime, with a rapidly increasing error rate at the end of it. To avoid the situation where a batch undergoes accelerated failure rates long before its scheduled retest there should be an upper limit of τ fixed by other means, e.g. a knowledge of typical lifetimes of electricity meters.

The period until the next test can act as a quality indicator for each batch, which takes into account not only the measurement errors, but also their economic consequences. If the quality of a batch falls this might be seen as a decrease in τ, and could warn the net owner of an impending failure of the devices. A replacement of the batch could be planned a few years ahead based on actual measurements, which we hope could be a valuable asset to the utility companies. The quality feedback might also encourage utility companies to invest in better measurement devices in the first place, because they can reduce the workload associated with testing. In particular, since the majority of the cost associated with replacement stems from the installation work, it might even be possible to improve the device performance noticeably with a modest additional investment. Furthermore, as smart meters are introduced the grid owner might be able to exploit the measurements to purge poor individual units and thus extend the period until the next test, for instance by detecting unrealistic patterns in the consumption (e.g. a constant consumption for weeks, or suspiciously low values, or similar features).

The probability calculations we have carried out use the acceptance criteria from Table 4, which amounts to a sample size dependent tolerance level for the actual error rate. For the period until the next test computation one could perhaps favourably choose a fixed maximum tolerable error rate (typically 5 %). This would affect batches of small to moderate size by shifting the balance point between the consumer side risk and the producer side risk towards a higher value of observed error rate (section 4, in particular Figure 7), and hence increase τ. The large batches would be unaffected since the maximum tolerated error rate approaches 5 % with larger sample sizes.

Without regard to which method of analysis we choose (a pure cost analysis, a pure calculation of probabilities, or a combination of these two) we would argue that there are other important reasons for enforcing a test regime on measuring instruments. Firstly, there is a moral obligation to ensure a fair distribution of costs among consumers: measurement errors will shift costs between individuals in a completely arbitrary fashion. Secondly, correct measurements are the basis for levying taxes. Finally, correct measurement of consumption may induce consumers to act to conserve energy in the most cost effective way. Figure 1 actually suggests that electrical energy meters are already subject to rather wide tolerance levels, and any suspension of test regimes should be considered with much more diligent analysis than the press release indicated.

6 Conclusion

We propose to make accept/reject decisions for each individual batch after legal metrological testing is completed according to simple statistical rules described by the sampling plans in Welmec Guide 8.10 or similar plans modified in order to also provide better protection against false rejection (producer side risk). The Norwegian sampling plans provide such protection.

We further propose to use a risk analysis approach for those batches accepted by the rules of the regulation in order to calculate a period until the next test, which balances the consumer side risk with the producer side risk. A flat 5 % rate of non-conformances can be used for all sampling plans for this purpose. For each acceptance test which is carried out, the conditions for the risk analysis are updated with respect to actual measurement errors, varying price profiles through the year, typical consumption profiles, frequency of use and the cost of replacement of electrical energy meters.

The period until the next test acts as a quality characteristic of the batch, where a shorter period until the next test indicates degrading quality. By monitoring this quality characteristic it is possible to predict when each batch is expected to be replaced. At a short period until the next test, the net owner on their own initiative might choose to replace the batch, even though the batch was accepted according to the regulation.

7 Acknowledgements

The work on this paper is supported by the European Metrology Research Programme (2012–2015) in project NEW04 with a grant from the European Commission via EURAMET and which resulted in a Guide . We would like to express our gratitude to the stakeholder, Norsk Energi, who supported the initiative for this work, and especially to Håkon Jahr at Hafslund Fakturaservice who delivered a set of hour by hour profiles on electric power consumption as well as the distribution of annual energy consumption in eastern Norway. Henning Kolbjørnsen (JV) assisted the project by establishing contacts with Hafslund, and gave the project a good start. Kristian Ellingsberg (JV) has commented on differences in calculation of the power / energy measurements from 1-phase and 3-phase types of electrical energy meters. Hossein Piltan (JV) has contributed with the analysis of results stored in the database and many good discussions throughout the project. Nils Magnar Thomassen (JV), Eli Mogstad Ranger (JV) and Bjørn Fjeldstad (JV) have shown an interest in the results of the project and contributed with good discussions concerning the regulation. Stefan Svensson (SP) has commented on our calculations of excessive costs due to erroneous measurements, and Francesca Pennecchi (INRIM) has proof read and commented on parts of this paper in its various stages.

8 References

[1] Pendrill, Leslie. Optimized Measurement Uncertainty and Decision-Making in Conformity Assessment. Measure, Vol 2, 2007, Vol. 2007, 2.

[2] Operating ‘cost’ characteristics in sampling by variable and attribute. Accreditation and Quality Assurance, Volume 13, Issue 11, 2008.

[3] Using measurement uncertainty in decision-making &conformity assessment. Metrologia 51, 2014.

[4] Welmec Guide 8.10 Measuring Instruments Directive (2004/22/EC): Guide for generating sampling plans for statistical verification according to Annex F and F1 of MID 2004/22/EC. 1, 2011.

[5] Montgomery, Douglas C. Introduction to Statistical Quality Control. s.l.: John Wiley & Sons, Inc, 2009. ISBN 978-0-470-16992-6.

[6] ‘MID’: Directive 2004/22/EC of the European Parliament and of the Council of 31 March 2004 on measuring instruments. EU. 2004.

[7] EN 50470:2006 Electricity metering equipment (a.c.) Part 1, Part 2, Part 3. CEN. s.l.: CEN, 2006.

[8] OIML R 46-1 & -2:2012 Active electrical energy meters, Part 1, Part 2. OIML. s.l.: OIML, 2012.

[9] OIML G 1-100:2008 'GUM' Evaluation of measurement data - Guide to the expression of uncertainty in measurement, 2008.

[10] OIML G 1-106:2012 Evaluation of measurement data - The role of measurement uncertainty in conformity assessment, 2012.

[11] ISO 2859-1 /-2 Sampling procedures for inspection by attributes, Part 1, Part 2. ISO. 1999.

[12] Donald J. Wheeler, David S. Chambers. Understanding Statistical Process Control. Knoxville : SPC Press, 2010. ISBN 978-0-945320-69-2.

[13] L.R. Pendrill, H. Karlsson, N. Fischer, S. Demeyer, A. Allard. A guide to decision-making and conformity assessment. Deliverable 3 3 1, EMRP project (2012-5) NEW04 Novel mathematical and statistical approaches to uncertainty evaluation. s.l. : Euramet, 2015.

Annex - The dataset, computation of costs

Figure A1 presents the measurement error curve for one particular electrical energy meter. We see that all the errors are negative, and that they lie within the specifications. When both positive and negative measurement errors are present they will to some extent cancel each other out when calculating the invoice error, depending on the frequency of use and the price of the utility. We have simplified the calculations of the calibration curve indicated in Figure A1, giving only rough estimates of the cost errors due to actual measurement errors.

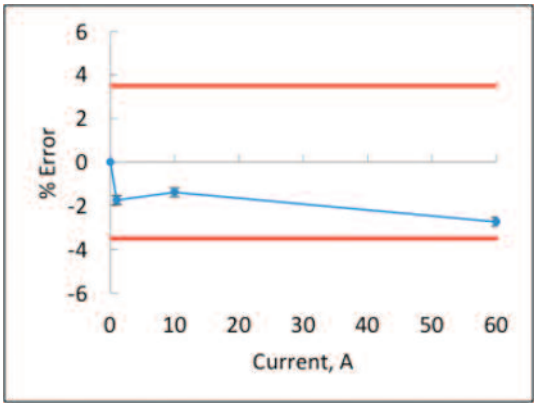

Figure A1: Calibration curve for a particular conforming meter in batch number 2783. The measurement errors are indicated with blue dots with measurement uncertainty bars. A stepwise linear interpolation is used to model errors between calibration points.

Figure A2 presents a typical power profile (frequency of use) for a private household. We have used only 38 such profiles from one vendor only. Some profiles received were excluded from this set of data because of strange behaviour, which was difficult to explain. For example, a variation period of 24 hours was expected, but we found a periodicity of ca. 32 hours. Also, some profiles showed constant power for several weeks.

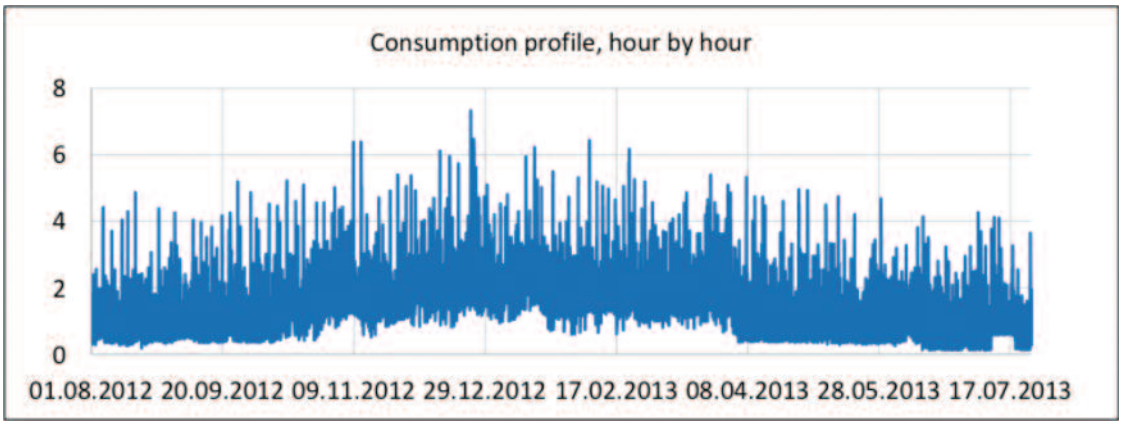

Figure A2: One example of a typical power profile for a private household in eastern Norway.

Figure A3 presents the distribution of the annual consumption of electrical energy for private households. The distribution of the annual consumption is highly skewed, with some excessively high consumptions. This curve is based on 500 000 electrical energy meters, all from one vendor, limited to eastern Norway, including Oslo. The annual energy consumption measured by our 38 typical power profiles in Figure A2 is marked with a blue "x" in Figure A3.

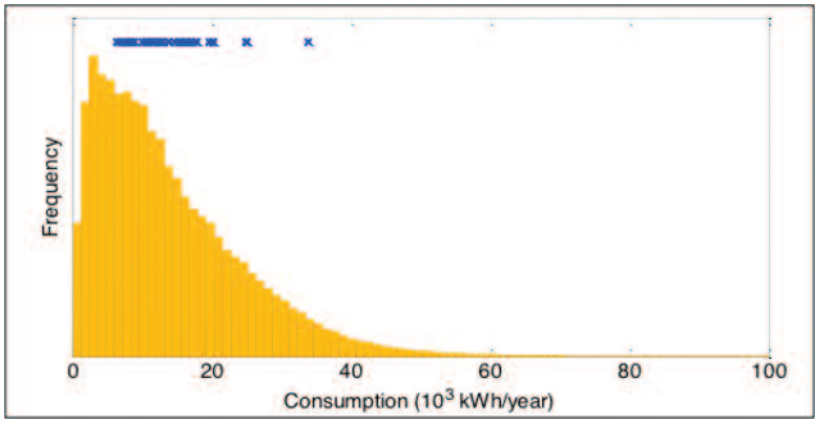

Figure A3: Annual electric energy consumption in private households in east part of Norway. Blue "x" indicate the annual electric energy consumption for our 38 available typical power profiles.

The scaling of each power profile by a randomly selected annual energy consumption is a rough way to simulate representative variations in actual power profiles. Our simulation might overstate the variation in power in some cases, or in other cases reduce variations. When variation in power exceeds I_{Max} these calculations have been deleted, effectively reducing the number of draws in the Monte Carlo simulation.

Figure A4 presents data for the price of electric power in the Oslo area. We have combined data sets for the same geographical area, and we have not simulated any variation between the different parts of Norway. Heating of private households in Norway is to great extent based on electricity, and there might be large variations in different parts of the country, especially during winter. Prices are highly correlated between different parts of Norway, but total energy consumption might have larger variations.

Figure A4: "Spot" price profile for electric power in the Oslo area. A grid fee of 390 NOK/MWh is added to this "spot" price.

The spikes on the spot price profile are of short duration, and have a small influence on the total annual cost. Many different price tariffs which avoid large spikes in the price profile are available. However, price tariffs which are different from the spot price tend to give higher annual prices on the annual energy consumption, as well as higher cost errors.

There are two shortcomings in our cost estimates: We have assumed a perfect power supply, measurement errors due to a phase difference between current and voltage are neglected, as well as measurement errors due to a voltage different from the standard value 230 V. Secondly, we have scaled a low number (38) of real example consumption profiles in a linear way to different yearly consumptions in order to simulate all the variations in yearly consumptions. This kind of scaling may not reflect fluctuations in real consumer profiles in a good way.

The probability density function for $\hat{p}$ was found using a binomial model for the sampling. For large batches, $n/N < 10$ %, this is a good approximation. Otherwise, a hypergeometric model for the sampling would be more accurate. We have limited our analysis to batch sizes $N \geq 200$, and also because of curtailment of the test, the requirement $n/N < 10$ % is not met for only 87 batches (total 756 batches), at a maximum value for $n/N = 30$ %.

We have presumed zero measurement errors for new electrical energy meters, which replace rejected electrical energy meters. This is a valid assumption when old mechanical meters with large measurement errors are replaced with electronic meters with low measurement errors. For electronic meters the distribution of measurement errors is narrower, but there may be a few meters with large measurement errors. When the electrical energy meter has very large measurement errors, it is possible to detect this by other means, and replacement of individual meters could be done without statistical sampling. ■

The Authors:

Helge Karlsson (JV)

Åge Andreas Falnes Olsen (JV)

Leslie Pendrill (SP)