THE QUALITY AND VALUE OF ROAD WEATHER FORECASTS

John Edward Thornes

School of Geography and Environmental Sciences
The University of Birmingham
Birmingham, B15 2TT, United Kingdom
j.e.thornes@bham.ac.uk TEL.+44-121-414-5556

1. Abstract

In order to decide whether or not a weather service supplier is giving good value for money we need to monitor the quality of the forecasts and the use that is made of the forecasts to estimate their value. A number of verification statistics are examined to measure the quality of forecasts including miss rate, false alarm rate and the Peirce skill score. In order to assess the economic value of the forecasts a quality value index is suggested that takes into account the cost-loss ratio and forecast errors. It is suggested that a combination of these quality and value statistics could be used by road weather forecast customers to choose the best forecast provider and to set limits for performance related contracts.

2. Introduction

Users of road weather forecasts are operating within an increasingly commercial environment and have to attempt to prove that they are getting value for money from all of their expenditure. The introduction of compulsory competitive tendering (CCT) may have reduced costs but it does not guarantee forecast quality and can lead to new players entering the market with no track record. One method to ensure customer satisfaction is to use performance related contracts. For example, a customer may set a target for the percentage of correct forecasts over the period of a contract and if the forecasts are better than the target the forecast provider receives a bonus and if they are worse than the target the forecast provider is paid less money, according to an agreed scale. If the target was 86% accuracy and the value of the contract was £20,000, for example, it could be agreed that for every percentage above 86% the forecast provider receives an extra £1,000 and for every percentage point below 86% they receive £2,000 less. These figures are arbitrary however. What is the true value of better or worse forecasts? This paper presents a more sophisticated set of potential verification targets that can be used to judge the quality of forecasts and chosen to suit the customer's commercial interests.

The judgement of weather forecast quality and/or value has received considerable attention in the literature in recent years for example Mylne (1999), Stanski et al. (1989), Thornes (1995,1996), Thornes & Proctor (1999), Richardson (2000), Ryder (1996), Stephenson (2000) and Wilks (1995). This paper demonstrates some of this knowledge in order to verify road weather forecasts and shows how that information can be used by highway engineers and other users, to keep a 'sharp eye' on their weather forecast suppliers.

More than £2 million pounds per winter is being spent on road weather forecasts in the United Kingdom out of a total budget of approximately £140 million for winter road maintenance. The corresponding figures for Agency roads are approximately £0.33m on forecasts and £20m in total. It is difficult to independently assess the quality and value of these road weather forecasts and most highway authorities rely on a simple set of statistics provided by the weather service providers. The current guidance specification in the UK (Thornes, 1993) for road weather forecasts only calls for a 'percent correct' (PC) of 86% for frost forecasts on nights when the minimum road surface temperature is 5 deg.C or below. In this case, for simplicity, a frost is defined as when the road surface temperature falls to zero or below irrespective of surface moisture. This measure of PC = 86% is inadequate and new measures are suggested below.

Weather forecast providers are required to produce a 2x2 contingency table at the end of the winter for each forecast site. The minimum road surface temperature for each night is noted at each forecast site and compared with the forecast minimum road surface temperature. The results are entered into the contingency table just for the nights when the actual road surface temperature fell to 5 deg.C or less.

For example, during the winter of 1995/96 there were 77 such nights at a road weather site located at High Eggborough on the M62 motorway between Leeds and Hull. Both the Met Office and Oceanroutes were providing forecasts for that site for different customers. The results for the Met Office are given as an example in Table 1

Table 1 Contingency Table for the Analysis of Met. Office Road Surface Temperature Forecasts at High Eggborough for the Winter of 1995/96

		Observed		
		Frost	No Frost	Total
	Frost	a = 29	b = 6	a + b = 35
Forecast	No Frost	c = 4	d = 38	c + d = 42
	Total	a + c = 33	b + d = 44	n = a+b+c+d = 77

a : Frost Forecast and Frost Observed (29 nights): Forecast correct
c : No Frost Forecast but Frost Observed (4 nights): Type 1 error
b : Frost forecast but No Frost Observed (6 nights): Type 2 error
d : No Frost Forecast and No Frost Observed (38 nights): Forecast correct
Type 1 Error: Possibility of severe road accidents as roads may not be salted
Type 2 Error: Possibility of wasted salt as roads may be salted unnecessarily

'Percent Correct' is simply the percentage of correct forecasts:

$$\text{Percent Correct} = PC = ((a + d) \times 100)/n$$

For this case PC is 87%, which just exceeds the target of 86%.

There are two types of error in the forecast. A Type 1 error is defined as those nights when the road surface temperature was forecast to stay above zero when in fact it fell to zero or below. This is potentially dangerous for the road user as the maintenance engineer may decide not to salt the roads and if the road is wet, ice may form on the road surface and accidents take place. The number of nights with a Type 1 error are given in the contingency table as 'c'. Type 2 errors occur when the road surface temperatures are forecast to go to zero or below but in reality they do not. The maintenance engineer may then salt the roads unnecessarily. This does not effect road safety but is a waste of salt and money. The number of nights with a Type 2 error is given in the contingency table as 'b'.

The use of Percent Correct is an over simplistic check on forecast quality that does not take into account the proportion of Type 1 and Type 2 errors. Also a forecast accuracy of greater than 86% may not be of greater value than the loss suffered if the forecast accuracy is less than 86%. The costs and losses associated with Type 1 and Type 2 errors are discussed below.

The highway engineer is concerned about more than just road surface temperature. Forecasts of road wetness and snow are also of considerable importance. It should be possible to also monitor the quality of such forecasts in any verification scheme and the assessment of snow forecasts is discussed below.

3. What Makes A Good Weather Forecast? Quality And/Or Value?

There needs to be a clear link between quality and value, especially when considering road weather forecasts. It has been traditionally accepted in the industry that a slight bias (explained below) in the forecast of road surface temperature should be present. This is due to consequences of the Type 1 and Type 2 errors discussed above. A Type 1 error in the forecast which leads to the roads not being salted could lead to the local authority being sued if a motorist skids on an icy road. This could cost the local authority millions of pounds in compensation if the driver is badly injured and the local authorities insurance ‘excess’ is high (Mead, 1998). Also Type 1 errors are undesirable because as they contribute to accident figures they run counter to Government policy to reduce accidents significantly by 2010. A Type 2 error will only cost the local authority tens of thousands of pounds if the roads are salted unnecessarily. Hence there is a tendency to ‘err on the side of caution’ and ‘over forecast’ the number of frosts or the occurrence of snow. It is only a matter of time before the weather forecast provider is also sued as a result of an incorrect forecast (Millington, 1987). The definitions of quality and value are discussed below with examples.

4. Quality Of A Forecast

Stanski *et al.* (1989) review six attributes of a weather forecast that make up the total quality: reliability, accuracy, skill, resolution, sharpness and uncertainty. They also make the important point that:

No single verification measure provides complete information about the quality of a product.

A number of measures of forecast quality are therefore required, but in order to avoid confusion their use must be obvious, they must be easy to calculate and their statistical significance should be testable. Of the six attributes mentioned above the first three: reliability, accuracy and skill are the easiest ones to measure and will be considered here. Resolution is important in the forecasting of precipitation - being able to distinguish between, for example, snow, sleet, freezing rain, hail, drizzle and rain. Sharpness is a measure of the spread of the forecasts away from climatology e.g. a forecast method that can predict frosts in summer as well as winter shows high sharpness whereas a forecast method that can only predict frosts in winter has low sharpness. Uncertainty relates to the climate, for instance some areas of the United Kingdom have comparatively few road frosts (e.g. Cornwall) in comparison to others (e.g. Cumbria). This may effect the achievability of performance targets (Halsey, 1995) and if frost or snow are rare events then the ‘base rate’ effect (Mathews, 1997) comes into play as discussed below. Another important attribute of the forecast is the ‘precision’ with which the forecast can hit the right side of a threshold (e.g. zero degrees Celcius). There are many other thresholds that are important to forecast users and sometimes the customer only needs to know whether or not a threshold will be crossed e.g. high winds affecting traffic safety (Thornes, 1997).

4.1 Reliability

The reliability of a forecast can be measured by calculating the bias. This will show if the forecasters are consistently over forecasting the number of frosts or snow. The bias tells us whether or not more forecasts of frost are being issued than frosts are observed. It is normal to find a positive bias in frost forecasts in order to hedge the chance of a Type 1 error. The bias is calculated as follows, using the notation of Table 1.

$$\text{Bias} = B = (a + b) / (a + c)$$

When B = 1 then the forecasts are said to be perfectly reliable. When the B > 1 then this indicates over forecasting and when the B < 1 this indicates under forecasting. The bias of the Met. Office forecast given in Table 1 is 1.06 i.e. slight over forecasting. A bias of 1 does not necessarily mean that the forecasts are accurate however.

4.2 Accuracy

Percent Correct (PC) has already been discussed above and is a measure of forecast accuracy. It relates to the terms 'a' and 'd' in the contingency table. There are several other measures of accuracy that attempt to look at the incorrect forecasts 'b' and 'c' and the two independent measures that are recommended here are Miss Rate (M) and False Alarm Rate (F), which are both calculated from the actual number of 'frosts' and 'no frosts' observed (i.e. the columns of the contingency table).

(a) The Miss Rate (M) is an important statistic as it calculates how many of the observed frosts were not forecast i.e. it relates directly to the number of Type 1 errors. We want this number to be as close to zero as possible. If 'c' is zero (i.e. no Type 1 errors) then the Miss Rate will be zero.

$$\text{Miss Rate} = M = c / (a + c)$$

The Miss Rate of the Met. Office forecast given in Table 1 is 0.12. The Hit Rate (H = (1 - M) = a / (a + c)) can be derived from the Miss Rate and both statistics are not therefore needed. The Hit Rate and the Miss Rate by themselves can be misleading, for instance if a frost was forecast every night then the Hit Rate would be 1 and the Miss Rate would be zero even though the forecast was of very poor quality.

(b) The False Alarm Rate (F) is also an important statistic as it considers the number of Type 2 errors, i.e. the number of nights that a frost was forecast but did not occur 'b'. These nights are when roads may be salted unnecessarily. If 'b' is zero then F is zero. The smaller the value of F the better. There is some confusion in the literature over the definition of the False Alarm Rate but for our purposes the False Alarm Rate is defined as:

$$\text{False Alarm Rate} = F = b / (b + d)$$

The False Alarm Rate for the Met. Office data given in Table 1 is 0.14. The Miss Rate and False Alarm Rate correspond to the two columns of data in the contingency table. It is better to examine the columns of the contingency table rather than the rows because it is the observations of frost or no frost that determine the quality of the forecasts.

4.3 Skill

There are many different skill scores that attempt to assess how much better the forecasts are than those which could be generated by climatology, persistence or chance. A forecast based on climatology would take, for example, the likelihood of frost based on the minimum road surface temperatures that have been observed on that day over the last 30 years. For road weather forecasts it is very unlikely that climatology will be of any use as the mean minimum road surface temperature in winter is above zero across most of the UK (Thornes, 1995). Hence Climatology would never predict a frost in most parts of the UK. Climatology can tell us that on average we can expect so many frosts in a winter but could not tell us when. Persistence is a very simple forecast method, for example, persistence would tell us that if there was a frost last night then there would be a frost tonight. This would score quite well for long periods of frost but would always be incorrect when the weather changes from frosty to non frosty nights and vice versa. Chance can be used to see if the distribution of scores in the contingency table is as expected from the frequencies of forecast and observed frosts (e.g. Heidke Skill Score discussed in Stanski et al.,1989) but the resulting statistic is difficult to interpret. It is proposed therefore to use a measure of skill that is easy to calculate and is appropriate for use in performance related contracts.

The Peirce Skill Score (PSS) was first published in 1884 and has since been rediscovered as Kuipers Performance Index and the True Skill Statistic (TSS) as discussed in Stephenson (2000). It is simply calculated from the Miss Rate (M) and the False Alarm Rate (F) as follows:

$$\text{Peirce Skill Score} = \text{PSS} = 1 - M - F$$

The closer the value of PSS to 1 the better. For the Met. Office data given in Table 1 PSS = 0.74. A weakness of PSS is that it treats M and F equally, irrespective of their likely differing consequences. Before targets can be considered for any of these measures the value of the forecasts must taken into account and these can then be incorporated into a performance related contract. Target setting is relatively arbitrary within a specified range, but providing targets are reasonable and consistent across all Agent areas they should be a fair measure of performance.

5. The Value Of A Forecast

Unlike skill, the value of a forecast depends on user requirements. Thompson & Brier (1955) proposed the simple cost/loss ratio for judging value. It can be applied to situations in which (a) the effect of adverse weather on an operation and the cost of taking action to avoid weather damage, is known in monetary terms; (b) the decision maker's dissatisfaction with a loss is a linear function of the monetary value of the loss; and (c) the probability of occurrence of adverse weather is known precisely. For winter road maintenance it should be possible to make reasonable estimates of (a) and (b) whereas assumption (c) is known after the event.

It is normal to denote the cost of taking action as C, in this case to salt the roads, and to denote the loss incurred as L, if the roads are not salted and accidents and delays occur, taking into account the savings made by not salting. On a given night if 'p' is the expected probability of adverse weather i.e. frost or snow then:

if $p > C / L$ it will pay to salt the roads
if $p < C / L$ it will not pay to take action
if $p = C / L$ it doesn't matter either way

Obviously it is assumed that $0 < C / L < 1$ i.e. that $C < L$

Relating this to the contingency table of Table 1 we find that:

Type 1 Errors = c Therefore costs incurred = c x L = cL
Type 2 Errors = b Therefore costs incurred = b x C = bC

If we determine that for a particular local authority the cost C of salting the roads for one night is £20,000 and that the Loss L incurred by not salting is likely to be £160,000, (Thornes (1999) has found a benefit/cost ratio of 8:1 for winter maintenance of roads) then C / L = 0.125 and for the frequencies of error given in Table 1 then:

cL = 4 * £160,000 = £640,000
bC = 6 * £20,000 = £120,000

So that the total money lost to the local authority due to errors in the forecast is estimated to be (cL + bC) = £760,000. The cost for the nights when the roads were salted correctly = a * £20,000 = 29 * £20,000 = £580,000.

The total cost therefore for the winter = aC + cL + bC = £1.34 million

It is usual to compare this to the costs that would have been incurred if no forecasts were issued and the roads were salted on all 77 marginal nights = 77 * £20,000 = £1.54 million

The forecasts therefore saved the local authority £1.54 - £1.34 = £0.2 million.

If the forecasts had been perfect then the roads would have been salted on just the nights that there was a frost ie (a + c) = 33 nights. This would have cost 33 * C = 33 * £20,000 = £660,000.

Perfect forecasts would have saved the local authority £1.54 - £0.66 = £0.88 million.

To put these figures into perspective note if the roads were never salted the total Loss would be:

(a + c) * L = 33 * £160,000 = £5,280,000

To summarise therefore the expense E of the various options is given by:
With perfect weather forecasts they would have spent E(P)=£0.66 million
With the quoted accuracy of Table 1 they would have spent E(A)=£1.34 million
If the roads were salted every marginal night it would cost E(S)=£1.54 million
If the roads were never salted it would cost E(N)=£5.28 million

These figures are only illustrative but they show the value of accurate forecasts and that Type 1 errors are to be avoided. There is still much to be gained by increasing the accuracy of the forecasts. One way to reduce Type 1 errors is to issue more forecasts of frost but that will increase the chance of a Type 2 error. This is acceptable up to a limit because the cost of a Type 2 error is so much less than that of a Type 1 error. In order to compare the quality and value of forecast providers we need an index that takes into account the number of Type 1 and Type 2 errors as well as the size of the Cost/Loss ratio. The relative value 'V' of a forecast system, as defined by Richardson (2000), compares the mean expense 'ME' of a forecast with that of a forecast based on climatology such that:

$$V = (ME(climate) – ME(forecast)) / (ME(climate) – ME(perfect))$$

V will have a value of 1 if the forecast system is perfect and will have a value of zero if the forecast is no better than climatology. A forecast based on climatology would, for example, predict a frost on a certain night if on average over the last 30 years there had been a frost on that night. Unfortunately the users of weather forecasts rarely have access to climate data and often their forecast sites are nowhere near a climate station.

The Value Index (V) is therefore defined as:

$$V = (ME(without\ forecast) – ME(forecast)) / (ME(without\ forecast) – ME(perfect))$$

Where ME(without forecast) can relate to climatology, persistence or chance, or whatever is used to compare with the forecast. For example one could compare the expense of salting all marginal nights or salting all nights or not salting at all, whichever is the cheapest method that does not use a forecast. In the example used above it is cheaper to salt all marginal nights (E(S) = £1.54m) than not to salt at all (E(N) = £5.28m). Therefore we can state that:

$$V = (E(S) – E(A)) / (E(S) – E(P))$$

Using the figures from above V = (1.54 – 1.34) / (1.54 – 0.66) = 0.23
It can be shown that the V can be simply calculated as follows

$$V = ((c + d) - (c/p)) / (n - W)$$

Where p = C / L, n = number of nights RST <= 5 deg.C, W = winter index (a + c)

Let us summarise the proposed quality and value statistics in Table 2

Table2. Forecast Quality and Value Statistics for High Eggborough for 77 Nights During the Winter of 1995/96

1) PC = (a + d) * 100 / n	= 87%
2) Bias = (a + b) / (a + c)	= 1.06
3) Miss Rate = c / (a + c)	= 0.12
4) False Alarm Rate = b / (b + d)	= 0.14
5) Peirce Skill Score (1 - M) - F	= 0.74
6) Total Cost (aC+bC+cL)	= £1.34m
7) Potential Cost With No Forecast (nC)	= £1.54m
8) Potential Cost Perfect Forecasts (a + c)C	= £0.66m
10) Value Index V	= 0.23

The Value Index normally varies between zero and 1. If the V is negative it means that the forecasts are so poor that it would be more cost effective to salt the roads every marginal night. Care should be taken to define the size of d, in other words to ensure that only marginal decisions are included in the contingency table. For example in Tables 1 and 2 above only 77 out of 151 winter nights (1st November - 31st March) were considered when the minimum road surface temperature was observed to fall to 5 deg.C or below. Otherwise d would be very large and make the calculations less meaningful. . If we had considered all 151 nights from 1st November to 31st March then (n – W) = 151 – 33 = 118 and V = 0.71. Also because the denominator of the Value Index is (n – W) if the index is to be used to compare forecast performance in different geographical locations then we need to ensure that (n – W) does not vary significantly from Area to Area and from winter to winter. Fortunately in a colder area or a colder winter both n and W will increase compensating each other to a large extent. The size of V will also depend upon p = C/L. Table 3 shows the effect of varying p between 0.1 and 1.0 for the Met. Office values given in Table 2

Table 3 Values of V for a Variety of C/L Values

C/L	V
0.1	0.05
0.125	0.23
0.2	0.5
0.4	0.73
0.6	0.8
0.8	0.84
1.0	0.86

For the setting of performance targets it is necessary to agree on a realistic C/L ratio first and then set reasonable targets based on the likely number of Type 1 and Type 2 errors.

6. Snow And The Base Rate Effect

Highway engineers receive a text prediction of whether snow is likely in the next 24 hours. The number of days with snow falling in the UK is many less than the number of days when the road surface temperature falls below zero. During the winter of 1995/96 at High Eggborough for the 77 nights when the road surface temperature fell to 5 deg.C or below there were 33 frosts and 16 days with snow. In most winters in this part of the UK there are less than 10 days with snow. The C/L ratio and the SCR are not normally considered for snow in the UK but should be in climates with more snow. Some weather events occur much more frequently than others and this can affect the quality of the forecasts. This is called the Base Rate Effect. Mathews (1997) uses this effect to show that when 'rain' is forecast it is less likely to be accurate than when the forecast is 'no rain'. We can use the snow forecasts produced by two forecast providers for the same High Eggborough site during the winter of 1995/96 to illustrate this effect as shown in Table 4.

Table 4 Snow Forecasts for High Eggborough for 77 Nights During the Winter of 1995/96

	Provider A	Observed		**Provider B**	Observed	
		Snow	No Snow		Snow	No Snow
Forecast	Snow	9	7	Snow	15	15
	No Snow	7	54	No Snow	1	46

	Provider A	Provider B
1) PC	= 81%	= 79%
2) Bias	= 1.0	= 1.88
3) Miss Rate	= 0.44	= 0.06
4) False Alarm Rate	= 0.12	= 0.25
5) Peirce Skill Score	= 0.45	= 0.69
6) Value Index (C/L=0.125)	= 0.082	= 0.64

Thus although the Provider A percent correct looks better at 81% than Provider B at 79% and the Provider A Bias is 1.0 compared to the Provider B Bias of 1.88 the rest of the statistics tell a very different story. Provider A has a Miss Rate of 0.44 compared to Provider B's Miss rate of only 0.06. Thus one has to be very careful in interpreting the Snow forecasts . These contrasting results for Provider A are a consequence of the small Base Rate of only 16 days when snow fell out of 77 days. The results for Provider B show the benefits of a large positive bias i.e. 'over forecasting' snow, which reduces the chance of a Type 1 error. The Peirce Skill Score and the Value Index are much higher for Provider B and clearly show that Provider B provided better snow forecasts than Provider A. There is still much room for improvement of both Providers.

7. Conclusion

With the use of a simple contingency table a number of very useful statistics can be calculated by the customer. These results can be written into performance related contracts, or at the very least be demanded from the weather forecast service providers at the end of the season. Also these statistics can be used to choose the best forecast provider and if a new provider comes into the market, then a performance related contract would safeguard against possible poor performance.

The new Value Index should make the setting of value targets more understandable but it should be noted that it is very dependent upon the cost-loss ratio. It is important therefore that both the customer and the forecast provider agree on this value before a contract is entered into. The Peirce Skill Score should only be used for setting performance targets if C/L is close to one, in other words if the Miss Rates and False Alarm Rates are of similar economic consequence to the user.

8. References

Halsey, N. G. J. (1995). Setting verification targets for minimum road temperature forecasts, *Meteorol. Appl.*, 2: 193-197.

Matthews, R. (1997). How Right Can You Be?, *New Scientist*, No 2072, 28-31.

Mead, J. (1998). There's a killer on the loose. *Proceedings of the Cold Comfort 98 Conference*, September 1998, Northampton, (Organised by the *Surveyor Magazine*)

Millington, (1987). Weather Forecasting and 'The Limitless Seas', *The Law Quarterly Review*, 103: 234-245.

Mylne, K. R. (1999). The use of forecast value calculations for optimising decision making using probability forecasts. *17th Conference on Weather Analysis and Forecasting,* 13-17 September 1999, Denver, Colorado, 235-239.

Richardson, D. (2000). Skill and economic value of the ECMWF ensemble prediction system. Quart. J. Royal Met. Soc., 126, 649-668

Ryder, P. (1996). *Analytical Quality-Value Relationships in Meteorological Forecasting and Warning Services.* Unpublished Met. Office Internal Report.

Stanski, H. R. & Wilson, L. J. & Burrows, W. R. (1989). *Survey of common verification methods in meteorology*. WMO/TD-No. 358, World Meteorological Organisation, Geneva, Switzerland, 114 pp.

Stephenson, D. B. (2000). Use of the 'odds ratio' for diagnosing forecast skill. *Weather and Forecasting*, 15: 221-232.

Thompson, J.C. & Brier, G.W. (1955). The Economic Utility of Weather Forecasts, *Mont. Wea. Rev.*, 83: 249-254

Thornes, J. E. (1993). Review of the Efficiency and Effectiveness of the National Ice Prediction Network in England, Unpublished Final Research Report to Highways Agency

Thornes, J. E. (1995). A comparative real-time trial between the Met. Office and Oceanroutes to predict road surface temperatures, *Meteorol. Appl.*, 2: 113-119.

Thornes, J. E. (1996). The Quality and Accuracy of a Sample of Public and Commercial Weather Forecasts in the UK, *Meteorol. Appl.*, 3: 63-74.

Thornes, J. E. (1997). Transport. In *Applied Climatology*, ed by A. Perry & R. Thomson, Routledge, ch 12,

Thornes, J. E. (1999). UK road salting – an international benefit/cost review, *Journal of the Institute of Highways and Transportation*, July/August: 22-26.

Thornes, J. E. & Proctor, E. A. J. (1999) Persisting with Persistence: The Verification of Radio 4 Weather Forecasts, *Weather*, 54: 311-321.

Wilks, D.S. (1995). *Statistical Methods in the Atmospheric Sciences*, Academic Press, 465p