Dŵr Cymru
Welsh Water

delivering quality analysis
CM Water Economics Ltd

Non-Household Night Use Assessment:

Re-transformation Bias

The issue

DCWW's non-household night use (NHNU) allowance assessment involves the application of the UKWIR methodology (UKWIR, 1999) to samples of logged consumption data. However, evidence from DCWW studies has indicated that allowances generated using the functional specification outlined in the UKWIR methodology tend to underestimate the true volume of night use, most notably for larger users. In particular, the sums of prediction errors across sampled sites are always found to be negative (often significantly so). Furthermore, it has been possible to identify a positive relationship between DMA-level reported leakage and the volume of assessed NHNU.

The bias issue potentially arises whenever linear-transformed relationships such as the UKWIR logarithmic model are back-transformed to make predictions 'in the levels' (*i.e*. the logarithmic predictions are exponentiated). A logarithmic transformation typically reduces skewness, making the distribution more symmetric and closer to normality. This has led to the widespread use of logarithmic regression models in the water sector. One problem with this approach in the context of NHNU estimation is the difficulty of dealing with both outlier and zero observations (inactive night use), but significant problems also arise because simple exponentiation of the logarithmic predictions in the UKWIR model does not produce unbiased predictions on the original volumetric scale.

While this retransformation issue has been well-documented in both theoretical econometrics literature and applied studies in the healthcare sector (see, for example, Crow and Shimizu, 1988; Jones, 2010; Manning, 1988; Mullahy, 1998; Newman, 1993; Sprugel, 1983; and Stow *et al*, 2003) it often remains ignored in modelling applications within the water industry and also across many other sectors.

Source of bias

The UKWIR methodology employs a double-logarithmic specification to estimate the relationship between night use (NU) and average daily use (ADU) (Equation 1).

$$\text{NU} = \alpha \, . \, \text{ADU}^{\beta} \qquad \textbf{\textit{(1)}}$$

Newman (1992) provides a clear explanation for the source of the bias: since we have simply taken logs of both sides of the NU-ADU relationship, it might seem obvious that the retransformation back to levels would involve simply taking exponents. However for predictive purposes, Equation 1 is incomplete, since it omits the transform of the error term. As Newman (*ibid*) suggests, this oversight is understandable, as when making similar predictions with linear regression models involving untransformed variables, the error term does not appear to be incorporated. However, the mean of the errors in these types of model is zero, and thus can be ignored when making predictions. In contrast, in a regression employing log-transformed variables, the errors have a mean of zero in logarithmic units, but *not* in the original arithmetic (level) units. Because the mean will not be zero after retransformation, unless the error term ε in the estimated logarithmic relationship is retained, resulting NHNU predictions following retransformation will be downwardly biased by the quantity e^{ε}. In essence, the bias problem arises because the mean of log (NU) differs from the log of mean (NU).

Bias correction

To correct for the bias, it is necessary to estimate e^{ε} and include it in the retransformation. If the error term is normally distributed with variance σ_{ε}^2, then the bias correction is given simply by

$$\exp(0.5\,\hat{\sigma}_{\varepsilon}^2) \qquad \textbf{\textit{(2)}}$$

However, if the error term is not normally distributed or is heteroscedastic, then the adjustment given by Equation 2 will be biased. In such cases, so-called 'smearing' estimators for e^{ε} may instead be applied (see Baser, 2007; Duan, 1983; Manning, *op cit*). Alternatively, different modelling approaches such as nonlinear, generalised linear models and two-part models may offer a more flexible and robust solution (see, for example, Buntin and Zaslavsky, 2004; Jones, *op cit*; Manning & Mullahy, 2001; and Welsch and Zhou, 2006). An additional advantage of some alternative modelling approaches in the context of NHNU is their potential for dealing more effectively with outliers and inactive night use (where the lower end of the NU distribution is censored at zero).

Application

Until now, DCWW's NHNU allowance assessments have been based on the application of the UKWIR methodology to data derived from logging surveys undertaken between 2002 and 2006. The problem of re-transformation bias was identified by CMWE in 2011, and in 2015 DCWW embarked on a long-term logging programme of 1000 non-household meters. Logging periods spanned intervals of between 14 and 365 days during the period December 2015 - December 2016. In early 2017, Equation 1 was re-applied to logging data from appropriately structured samples of meters across 9 (of 15) user-type sectors, and appropriate bias adjustment procedures were applied.

The scale of the bias adjustment for Sector 1 (Agriculture) is graphically indicated in Figure 1.

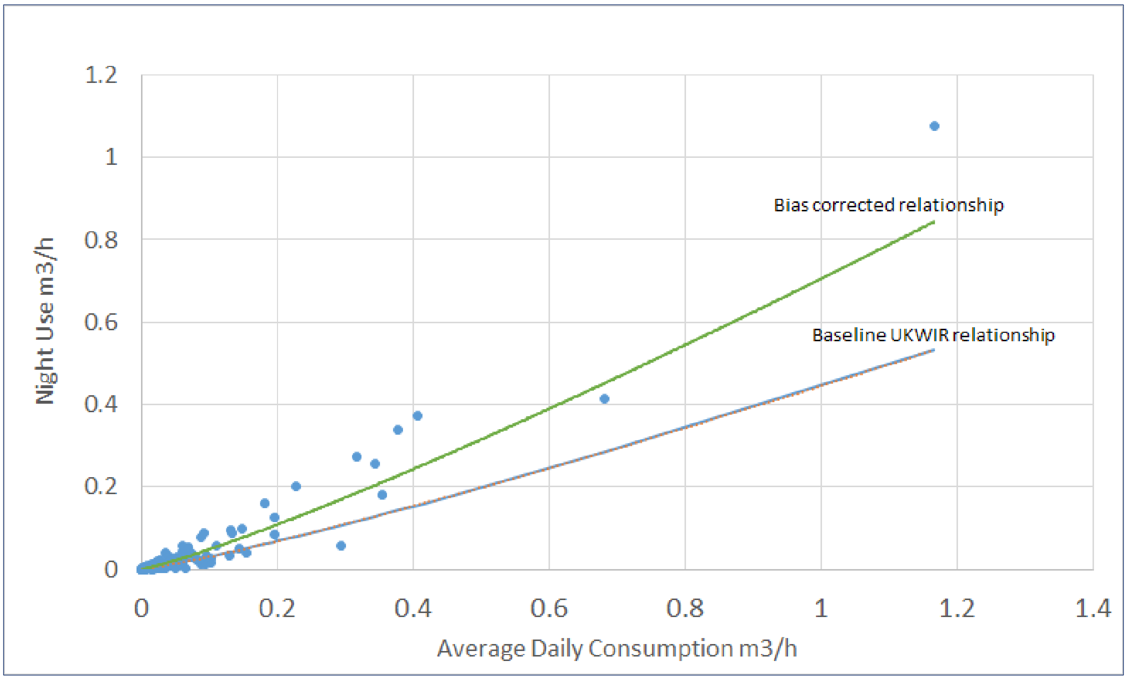

Figure 1: Bias adjustment for Sector 1 (Agriculture)

A pragmatic but conservative approach to bias correction was adopted in this study and the extent of the indicated bias is consistent with both the relatively high error variances generally evidenced in the logarithmic NHNU relationship, and also with evidence from applications elsewhere in applied research, which suggests biases in excess of 25% are not uncommon (see, for example, Newman, *op cit;* and Duan, *op cit*).

Conclusion

Application of the existing UKWIR methodology leads to biased NHNU allowance estimates. The choice of approach for bias correction is not always straightforward and depends in large part on the statistical properties of the error term, which indicates the need for appropriate statistical testing procedures. Moreover, alternative modelling approaches may have preferable statistical properties which could potentially give rise to further accuracy improvements in NHNU allowance assessment.

While it may be conservative, relative to DCWW's previous allowance assessment, the adjustment identified in this study almost certainly provides a more accurate representation of NHNU across the company as a whole. However, additional and more detailed modelling work is required to further improve NHNU allowance accuracy, in relation not only to the issue of retransformation bias, but also the more robust treatment of outliers and inactive night use, appropriate user-type stratification, seasonal variations, and consistency with UKWIR's recently proposed leakage reporting guidelines ('convergence').

Caroline Manning
28/04/17

References

Buntin M. B. and A.M Zaslavsky (2004). Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures, *Journal of Health Economics*, Vol 23(3); 525–542

Crow, E.L and K. Shimizu (1988) Eds. Lognormal Distributions: Theory and Applications. Volume 88 of Statistics: A Series of Textbooks and Monographs, *CRC Press*, Marcel Dekker Inc – New York

Duan, N. (1983) Smearing estimate: a nonparametric retransformation method. *Journal of American Statistical Association*, Vol 78; 605-610

Jones, A.M (2010) Models for Health Care, University of York, HEDG Working Paper 10/01.

Manning W.G. (1998). The logged dependent variable, heteroskedasticity, and the retransformation problem. *Journal of Health Economics,* Vol 17; 283-295

Manning W.G. and H. Mullahy (2001) Estimating log models: to transform or not to transform? *Journal of Health Economics,* Vol 20; 461-94

Mullahy, J. (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics,* Vol 17; 247-81.

Newman, M.C. (1992). Regression analysis of log-transformed data: statistical bias and its correction. *Environmental Toxicology and Chemistry*, Vol. 12, pp 1129-1133

Sprugel D.G. (1983). Correcting for Bias in Log-Transformed Allometric Equations. *Ecology*, Vol 64(1); 209-210.

Stow, C.A., K.H Reckhow, and S.S. Qian (2006) A Bayesian approach to retransformation bias in transformed regression. *Ecology*, Vol 87(6); 1472-1477

UKWIR (1999), Estimating Legitimate Non- Household Night Use Allowances. Report 99/WM/08/26

Welsch A.H. and X. H. Zhou (2006). Estimating the retransformed mean in a heteroskedastic two-part model. *Journal of Statistical Planning and Inference,* Vol 136; 560-881