EMS Annual Meeting Abstracts
Vol. 12, EMS2015-485, 2015
15th EMS / 12th ECAM

Spurious skill? Varying event frequencies, non-collapsibility and Simpson's Paradox in forecast verification

Roger Harbord
Met Office, Exeter, United Kingdom (roger.harbord@metoffice.gov.uk)

The climatological frequency of high-impact weather events varies markedly between locations and seasons. Measures of forecast skill can be misleading if samples of locations or dates with markedly varying climatological event frequencies are combined, as has previously been pointed out (Mason 1989 *Aust. Meteor. Mag.* 37:75-81; Juras 2000 *Weather Forecast.* 15:365-366; Hamill & Juras 2006 *Q. J. Roy. Meteor. Soc.* 132:2905-2923). Within the statistics literature this phenomenon has been recognised since at least 1903 (Yule, *Biometrika* 2:121-134) and is now generally known as non-collapsibility (e.g. Greenland, Robins & Pearl 1999 *Stat. Sci.* 14:29-46). It can affect all skill measures and all forecast types (binary, categorical, continuous, probabilistic), though the precise conditions for its occurrence vary between measures.

Somewhat more recently, statisticians realised that this can affect not just the magnitude but also the sign of the measure, a phenomenon known for binary events as Simpson's Paradox (Blyth 1972 *J. Amer. Statist. Assoc.* 67:364-366), but this phenomenon does not appear to have been recognised as yet in the forecast verification literature.

We present an illustrative example of Simpson's paradox in forecast verification and outline the conditions necessary for its occurrence. We then explore its frequency of occurrence, and of substantial non-collapsibility falling short of sign reversal, in some recent Met Office forecast verification data. We conclude by discussing some possible approaches that may avoid or minimise it, including recalibration of thresholds and meta-analytic methods.