What are the conditions for valid extrapolation of statistical predictions?

Demetris Koutsoyiannis

Department of Water Resources, Faculty of Civil Engineering, National Technical University of Athens, dk@itia.ntua.gr - http://www.itia.ntua.gr/dk

It is maintained that the most important conditions to obtain valid statistical predictions are (1) to be aware of the fundaments of probability, statistics and stochastics, (2) to formulate the problem as clearly as possible, (3) to know the statistical/stochastic properties of the variables involved, such as marginal and dependence properties, and (4) to use correct statistical results, i.e. those results that correspond to the nature of the problem and the variables involved. These conditions are not always met in scientific publications and practical applications.

Posted in http://landshape.org/enm/what-are-the-conditions-for-valid-extrapolation-of-statistical-predictions-answer-ii/, August 2006

1. A statistical prediction should be distinguished from a deterministic prediction. In a deterministic prediction some deterministic dynamics of the form $y = f(x_1, \ldots, x_k)$ are assumed, where y is the predicted value, the output of the deterministic model $f(\ )$, and $x_1, \ldots, x_k$ are inputs, i.e. explanatory variables. The model $f(\ )$ could be either a physically based one or a black box, data driven one. The latter case is very frequent, e.g. in local linear (chaotic) models and in connectionist (artificial neural network) models.

Now in a statistical prediction we assume some stochastic dynamics of the form $Y = f(X_1, \ldots, X_k, V)$. There are two fundamental differences from the deterministic case. The first, apparent in the notation (the upper-case convention), is that the variables are no more algebraic variables but random variables. Random variables are not numbers, as are algebraic variables, but functions of the sample space. This is very important. The second difference is that an additional random variable V has been inserted in the dynamics. This sometimes is regarded as a prediction error that could be additive to a deterministic part, i.e. $f(X_1, \ldots, X_k, V) = f_d(X_1, \ldots, X_k) + V$. However, I prefer to think of it as a random variable manifesting the intrinsic randomness in nature.

2. To avoid confusion it is always advisable to formulate the stochastic model in such a way that all $X_1, \ldots, X_k$ are observable or, better, observed, so that we can directly apply it to obtain predictions that are conditioned on $X_1 = x_1, \ldots, X_k = x_k$, where $x_1, \ldots, x_k$ denote observations of $X_1, \ldots, X_k$. Predictions can be of point or interval type. The point prediction is $y = E[f(X_1, \ldots,$

X_k, $V|X_1=x_1$, …, $X_k=x_k$) = $E[f(x_1, …, x_k, V)]$. Here $E[\ ]$ denotes expectation and in the last part of equation it was assumed that V is independent of X_1, …, X_k. Interval predictions are intervals (y_b, y_a) satisfying $P\{y_b < Y < y_a || X_1=x_1, …, X_k=x_k\} = \alpha$, where $P\{\ \}$ denotes probability and α is a confidence coefficient. In simple cases these are calculated analytically; in other cases analytical solutions are not feasible and the method of choice is Monte Carlo simulation.

3. In addition to the inherent uncertainty that is described by the variable V, we have also uncertainty in parameters of the model $f(\ )$ because these parameters are usually estimated from a sample rather than by theoretical reasoning. This obviously influences our predictions, point and interval, and should be taken into account for a consistent description of uncertainty. Its quantification could be done using the notion of confidence limits of estimation, a notion very different from the prediction limits discussed in point 2.

4. In natural systems, all variables X_1, …, X_k are dependent to each other. This is usually missed. For instance the classical statistical law that relates the width of confidence intervals to the square root of the sample size is no longer valid if there is dependence, particularly long-range dependence. Sadly, numerous (if not most) published results on related issues have been based on this and other classical statistical laws that are valid merely when X_1, …, X_k are independent. The error in statistical predictions from such misuses could be huge.

5. To summarize, I think that the most important conditions to obtain valid statistical predictions are (1) to be aware of the fundaments of probability, statistics and stochastics, (2) to formulate the problem as clearly as possible, (3) to know the statistical/stochastic properties of the variables involved, such as marginal and dependence properties (particularly, the behaviour of the distribution tails is very important for extrapolations), and (4) to use correct statistical results (formulae, estimators etc.), i.e. those results that correspond to the nature of the problem and the variables involved.