Hydrol. Earth Syst. Sci., 16, 1255–1257, 2012
www.hydrol-earth-syst-sci.net/16/1255/2012/
doi:10.5194/hess-16-1255-2012

Hydrology and Earth System Sciences

Technical note: A significance test for data-sparse zones in scatter plots

V. V. Vetrova and W. E. Bardsley

Department of Earth & Ocean Sciences, University of Waikato, Hamilton, New Zealand

Correspondence to: W. E. Bardsley (web@waikato.ac.nz)

Received: 26 December 2011 – Published in Hydrol. Earth Syst. Sci. Discuss.: 26 January 2012
Revised: 27 March 2012 – Accepted: 29 March 2012 – Published: 26 April 2012

Abstract. Data-sparse zones in scatter plots of hydrological variables can be of interest in various contexts. For example, a well-defined data-sparse zone may indicate inhibition of one variable by another. It is of interest therefore to determine whether data-sparse regions in scatter plots are of sufficient extent to be beyond random chance. We consider the specific situation of data-sparse regions defined by a linear internal boundary within a scatter plot defined over a rectangular region. An Excel VBA macro is provided for carrying out a randomisation-based significance test of the data-sparse region, taking into account both the within-region number of data points and the extent of the region. Example applications are given with respect to a rainfall time series from Israel and also to validation scatter plots from a seasonal forecasting model for lake inflows in New Zealand.

1 Introduction

A visual examination of hydrological scatter plots is a useful first step toward considering possible relationships between variables, or for evaluation of the worth of hydrological forecasting models via validation plots of observed and predicted values. It is intuitive that we tend to focus on regions in scatter plots with greatest data density as this suggests highest degree of association and worth most effort in further refinements – see, for example, Green and Finlay (2008). However, a sufficiently extensive data-sparse zone in a scatter plot can be of interest also as this may suggest that for a specific magnitude range one variable might restrict the other.

For hydrological variables, the transition between data-sparse and data-dense fields in scatter plots will most likely be a poorly-defined boundary which can be thought of as a stochastic frontier, for which a range of estimation techniques are available (Hall and Simar, 2002; Florens and Simar, 2005; Delaigle and Gijbels, 2006; Kumbhakar et al., 2007). Our focus here is not on boundary estimation as such, but rather on providing a significance test against the null hypothesis that a data-sparse zone in a scatter plot has arisen by random chance. Specifically, the purpose of this short communication is to provide a practical significance test for the size of the area of an observed data-sparse region with a linear internal boundary in a scatter plot within the specific rectangular region which just encompasses all the data points. The test requires no assumptions concerning the data. For convenience, the data-sparse area is taken to mean its proportion of the rectangle area. Given that there are m data points within the data-sparse area $\Delta(m)$, the null hypothesis is that a data-sparse region at least as large and containing m data points could have arisen by random chance. Rejection of the null hypothesis does not imply any specific alternative with respect to correlation between the variables, but simply indicates that the data-sparse region is confirmed large enough so as to be unlikely to have arisen by chance. The approach adopted here represents a generalisation of an earlier test described by Bardsley et al. (1999) which was restricted in practical application because it required the data-sparse region to contain no data points at all ($m = 0$).

The nature of a data-sparse (as opposed to no-data) region is illustrated with respect to the scatter plot in Fig. 1. The pattern of data points suggests a possible linear rising trend in an upper boundary for October rainfalls at a site in Israel over the period 1951–1987, but with an unusually wet month in October 1986 as an outlier. The $m = 0$ requirement of the original 1999 test required a somewhat unrealistic location of a boundary as being above the outlier (Bardsley et al., 1999, Fig. 3a). A better approach is to deem "data sparse" in this particular case as permitting a single point within the

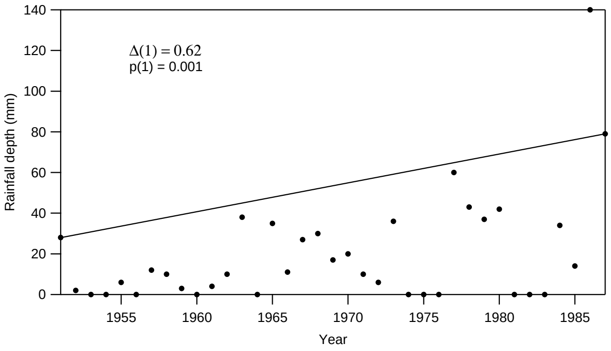

Fig. 1. Scatter plot of October Rainfall values (1951–1987) at Berurim in southern Israel. Line shows the linear internal boundary of the largest possible data-sparse region for $m = 1$.

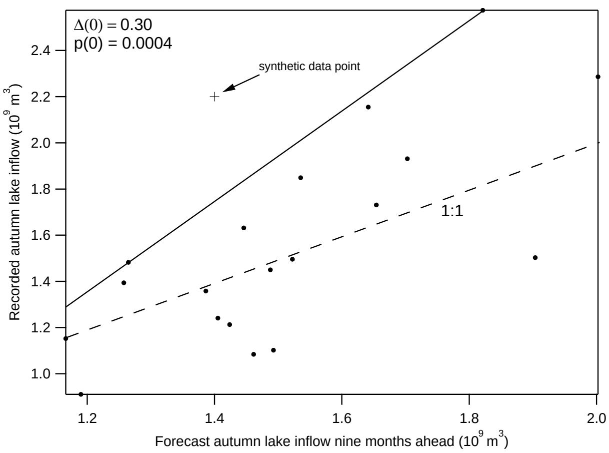

Fig. 2. Validation plot for a model forecasting combined autumn river inflow volumes into Lakes Tekapo and Pukaki (New Zealand).

data-sparse region ($m = 1$) which now gives a better linear boundary location just above the other data points (Fig. 1).

2 The test

Following Bardsley et al. (1999), the significance test of the present paper is based on a standard randomisation approach. That is, the x-coordinates of the data points are randomly reassigned, giving rise to a different pattern of points in the scatter plot in the rectangular region. For example, if the x-axis represented yearly values, then this would amount to a random reordering of years. After a given random reordering of x-coordinates, a check is made in the algorithm whether the largest upper-left region (with linear internal boundary) containing m data points is larger than the upper-left data-sparse area in the original scatter plot.

This random reordering of the x-coordinates is repeated many times, and the proportion of times p that the original data-sparse area is exceeded is calculated. This p value is the probability of obtaining a data-sparse area at least as large as observed in the original scatter plot, given that the null hypothesis is true. Therefore, if p is sufficiently small, say less than 0.05, then the size of the original data-sparse region $\Delta(m)$ is deemed statistically significant. The number of random reorderings needed for the required precision of p is determined from the binomial theorem in the usual way (see Excel spreadsheet in the Supplement).

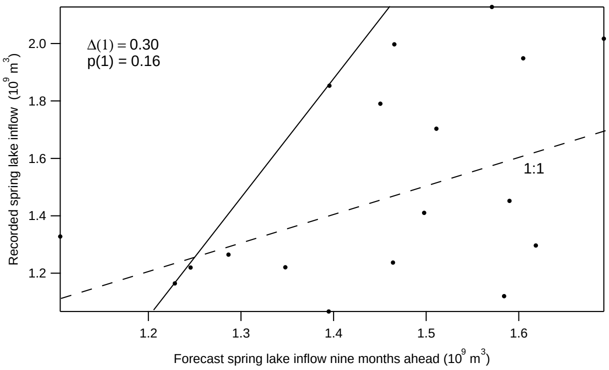

Fig. 3. Validation plot for a model forecasting combined spring river inflow volumes into Lakes Tekapo and Pukaki (New Zealand).

A general VBA macro which is unrestricted as to the size of m is described in the Excel spreadsheet supplementary to this paper. The macro appears efficient in trial runs but inevitably will become slower for large numbers of points in the scatter plots coupled with large m. When the macro is applied to the indicated data-sparse region above the line in Fig. 1 ($m = 1$), the resulting p value is obtained as $p(1) = 0.001$, which is a higher level of statistical significance then the value of $p(0) = 0.02$ listed in Fig. 3a of Bardsley et al. (1999) for the case of $m = 0$. Of course, there is no general guarantee that higher levels of significance will be obtained for the test proposed here, as this is dependent on the data pattern of the scatter plot.

3 Application to validation scatter plots

Scatter plots most commonly serve as a graphical indication of some degree of association between two variables. In addition, scatter plots are often used in hydrology to give a graphical indication of how well some model fits a set of validation data. The ideal here is to have points scattered close to the 1:1 line and Bardsley and Purdie (2007) present an "invalidation test" as one means of testing departure from this situation. However, a validation scatter plot may indicate failure in the sense of poor 1:1 fitting but nonetheless still possess some degree of predictive ability as evident from the pattern of points. For example, the location of a data-sparse region in a validation scatter plot may suggest that low predicted values tend to be associated with low observed values, but increasingly large predicted values result in high or low magnitudes being as likely.

An illustration of this situation is given in Fig. 2, which shows a validation data set with respect to a seasonal lake inflow forecasting model seeking to anticipate total autumn inflow from the standpoint of autumn in the previous year. The lakes concerned (Tekapo and Pukaki) are adjacent New Zealand hydro storage lakes and it is convenient to consider seasonal forecasts of the combined inflow volumes of both lakes. The forecasting model itself will be described in a subsequent publication but for the purposes of the present paper the point of interest is that the validation scatter plot can be interpreted as the forecasts giving a low probability to high inflows when low lake inflows are forecast. However, at the same time high inflow forecasts may associate with high or low actual inflows. This lends itself to a data-sparse significance test ($m = 0$) which in fact indicates high significance of the sparse zone above the solid line with $p(0) = 0.0004$.

Although an $m = 0$ test may appear sufficient here, there could be concern over robustness of the conclusion because of the small number of data points involved. The $m > 0$ test gives an empirical means of robustness checking because artificial data points can be inserted into the data-sparse zone and a check made to see if statistical significance is maintained. For example, inserting the single synthetic data point indicated in Fig. 2 yields $p(1) = 0.002$, which is still highly significant. The forecasting model in this case should probably be robust therefore against a future real data point appearing in the sparse zone. Further synthetic data points could be inserted if required. The autumn forecasting model here is restrictive in that low forecast flows will tend to be below the solid line. However, forecast high flows in reality could be anywhere within the magnitude range. The forecasting value is with respect to a high probability that a forecast flow will not be in the data-sparse zone, as opposed to being near or far from the 1:1 line.

This view of forecasting value is also illustrated in Fig. 3, showing in this case the validation results of a model for forecasting spring inflows into the two lakes, where the model is forecasting from the previous spring. The predictive model clearly fails in the sense of any 1:1 matching, but the hope might be that the indicated solid line approximates an upper bound to actual inflows when forecasts are in the range 1.20–1.45×10^9 m^3. As with the autumn model, the validity of this upper bound is tested for significance via the macro with respect to the relative size of the upper left ($m = 1$) data-sparse empty corner. It happens in fact that the macro-derived $p(1)$ value of 0.16 indicates the sparse zone size is no larger than expected from chance. The predictive model therefore fails not only in the 1:1 sense, but also in the sense of establishing the existence of an upper boundary which might permit an estimated upper bound to some forecast inflows.

4 Discussion and conclusion

There is an element of subjectivity introduced for the test considered here with $m > 0$, in that sometimes it will not be evident which value of m best defines a data-sparse region. Some trial and error process will most likely be required in such instances. With respect to further development, the test approach considered here should be amenable to generalisation such as allowing for curved inner boundaries and incorporating multiple dimensions. However, the randomisation algorithms may become complex and slow.

As noted in Bardsley et al. (1999), there will be data situations where linear regression is the most appropriate analysis technique. In other situations where data-dense and data-sparse fields are separated by an approximate linear boundary, the test given here should find practical applications for both associations between variables and also for checking validation scatter plots under situations of restricted forecasting ability.

Supplementary material related to this article is available online at: http://www.hydrol-earth-syst-sci.net/16/1255/2012/hess-16-1255-2012-supplement.zip.

Acknowledgements. We are grateful to Meridian Energy Ltd for providing the Pukaki/Tekapo lake inflow data. The Israel Meteorological Service and the late Alex Manes provided the Berurim rainfall data set. We appreciate the helpful comments of referees on the original manuscript.

Edited by: B. Schaefli

References

Bardsley, W. E. and Purdie, J.: An invalidation test for predictive models, J. Hydrol., 338, 57–62, 2007.

Bardsley, W. E., Jorgensen, M. A., Alpert, P., and Ben-Gai, T.: A significance test for empty corners in scatter diagrams, J. Hydrol., 219, 1–6, 1999.

Delaigle, A. and Gijbels, I.: Data-driven boundary estimation in deconvolution problems, Comput. Stat. Data An., 50, 1965–1994, 2006.

Florens, J.-P. and Simar, L.: Parametric approximations of nonparametric frontiers, J. Econometrics, 124, 91–116, 2005.

Green, M. B. and Finlay, J. C.: Detecting characteristic hydrological and biogeochemical signals through nonparametric scatter plot analysis of normalized data, Water Resour. Res., 44, W08455, doi:10.1029/2007WR006509, 2008.

Hall, P. and Simar, L.: Estimating a changepoint, boundary, or frontier in the presence of observation error, J. Am. Stat. Assoc., 97, 523–534, 2002.

Kumbhakar, S. C., Park, B. U., Simar, L., and Tsionas, E. G.: Nonparametric stochastic frontiers: a local maximum likelihood approach, J. Econometrics, 137, 1–27, 2007.