Hydrol. Earth Syst. Sci. Discuss., 9, 1335–1343, 2012
www.hydrol-earth-syst-sci-discuss.net/9/1335/2012/
doi:10.5194/hessd-9-1335-2012

Hydrology and Earth System Sciences Discussions

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

Received: 26 December 2011 – Accepted: 19 January 2012 – Published: 26 January 2012

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

Published by Copernicus Publications on behalf of the European Geosciences Union.

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 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 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

HESSD
9, 1335–1343, 2012

V. V. Vetrova and W. E. Bardsley

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 a data-sparse region of given size with a linear internal boundary in a scatter plot within a rectangular region as defined by the data. 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.

2 The test

The nature of the test can be illustrated with respect to the scatter plot in Fig. 1, which 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 zero-data requirement of the original 1999 test required a somewhat unrealistic location of the data-sparse boundary as being above this point (Bardsley et al., 1999, Fig. 3a). A better approach here is to deem "data sparse" in this case as permitting a single point within the region and placing a linear boundary just above the other data, indicated by the solid line in Fig. 1.

The significance test of the present paper can now be defined generally as finding the probability p that random swapping of data points will give rise to a data-sparse region (containing exactly m data points) which has an area greater than the original defined data-sparse area $\Delta(m)$ containing m data points. At each data reordering, the largest possible data-sparse zone containing m data points is found, and a check made as to whether that area is greater than the original $\Delta(m)$. The value of p is thus determined by a sufficiently large number of repeated random reorderings of the data set, where precision of p is determined from the binomial theorem in the usual way. Following standard practice, if p is less than 0.05 then the area of the original observed data-sparse zone is deemed sufficiently large so as to be unlikely to have arisen by chance.

HESSD
9, 1335–1343, 2012

V. V. Vetrova and W. E. Bardsley

As in Bardsley et al. (1999), Δ is expressed as a proportion of the rectangular area as defined by the outer limits of the data scatter (Fig. 1). An alternative approach would be to redefine the rectangular region for each random data reordering. This would represent a different form of test and may yield different results.

A general VBA macro which is unrestricted as to the size of m is described in the Excel spreadsheet in the Supplement 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.

The indicated sparse region above the solid line in Fig. 1 yields $p(1) = 0.001$ which can be compared to $p(0) = 0.02$ listed in Fig. 3a of Bardsley et al. (1999).

3 Application to validation scatter plots

It may happen that a hydrological model is calibrated to a data set with respect to some measure of best fit but fails in validation with respect to the same fit criterion. Bardsley and Purdie (2007) present an "invalidation test" as one means of detecting this situation. However, the model may still exhibit some degree of predictive ability in a conditional sense. 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 example 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 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,

HESSD
9, 1335–1343, 2012

V. V. Vetrova and W. E. Bardsley

high inflow forecasts may associate with high or low actual inflows. This lends itself to a data-sparse significance test which in fact indicates high significance of the sparse zone above the solid line with $p(0) = 0.0004$. There is concern, however, in that the small number of data points may suggest lack of robustness of this outcome in the event of a new data point appearing in the data-sparse zone. The algorithm of the present paper permits such investigations and inserting a synthetic data point in Fig. 2 yields $p(1) = 0.002$, which is still highly significant. This suggests that the autumn forecasting model has value for forecasting some future low inflows, while recognising there will be other low inflows which occur when high inflows are forecast.

Figure 3 shows the corresponding validation plot for spring inflow forecasts, suggesting that here too there may be a possible linear boundary with a positive gradient (solid line) to enable some degree of forecasting ability. However, the $p(1)$ value of 0.16 indicates that there is no confirmed predictive ability for spring inflows for this model.

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 checking validation scatter plots under situations of restricted forecasting ability.

HESSD

9, 1335–1343, 2012

Supplementary material related to this article is available online at: http://www.hydrol-earth-syst-sci-discuss.net/9/1335/2012/ hessd-9-1335-2012-supplement.zip.

Acknowledgements. We are grateful to Meridian Energy Ltd for providing the Pukaki/Tekapo lake inflow data. The Israel Meteorological Service and Alex Manes provided the Berurim rainfall data set.

References

Bardsley, W. E. and Purdie, J.: An invalidation test for predictive models, J. Hydrol., 338, 57–62, 2007. 1338

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. 1337, 1338, 1339

Delaigle, A. and Gijbels, I.: Data-driven boundary estimation in deconvolution problems, Comput. Stat. Data Anal., 50, 1965–1994, 2006. 1336

Florens, J.-P. and Simar, L.: Parametric approximations of nonparametric frontiers, J. Econometr., 124, 91–116, 2005. 1336

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. 1336

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. 1336

Kumbhakar, S. C., Byeong, U. P., Simar, L., and Efthymios, G. T.: Nonparametric stochastic frontiers: a local maximum likelihood approach, J. Econometr., 137, 1–27, 2007. 1336

HESSD

9, 1335–1343, 2012

Significance test for data-sparse zones

V. V. Vetrova and W. E. Bardsley

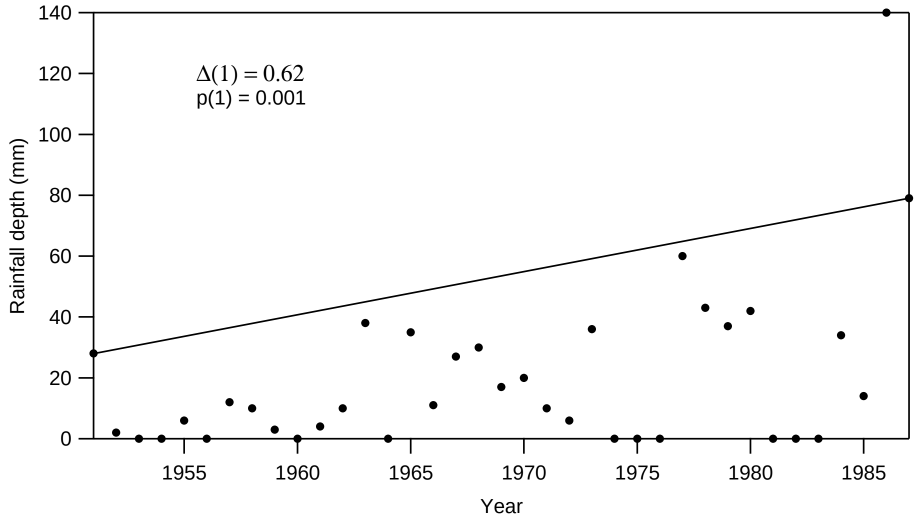

Fig. 1. Scatter plot of October Rainfall values (1951–1987) at Berurim in southern Israel. Solid line shows the linear internal boundary of the largest possible data-sparse region for $m = 1$.

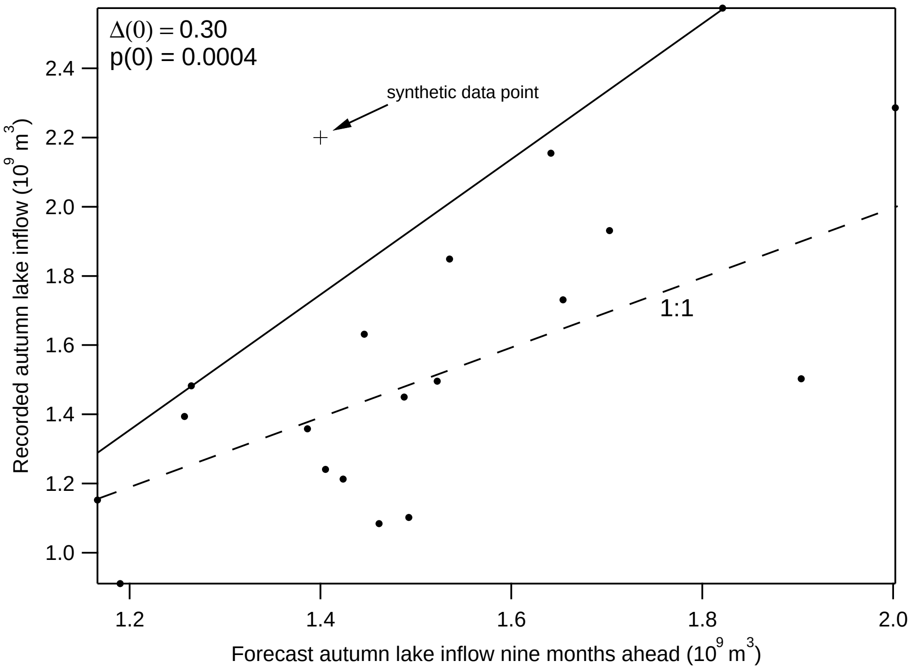

Fig. 2. Validation plot for a model forecasting combined autumn river inflow volumes into Lakes Tekapo and Pukaki (New Zealand).

HESSD

9, 1335–1343, 2012

V. V. Vetrova and W. E. Bardsley

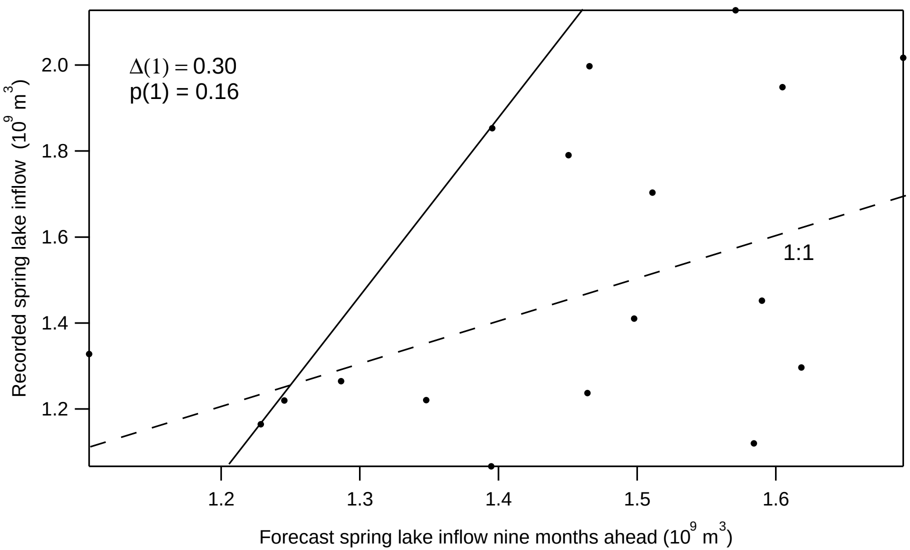

Fig. 3. Validation plot for a model forecasting combined spring river inflow volumes into Lakes Tekapo and Pukaki (New Zealand).

HESSD
9, 1335–1343, 2012

Significance test for data-sparse zones

V. V. Vetrova and W. E. Bardsley