ELSEVIER

Journal of Hydrology 219 (1999) 1–6

Journal of Hydrology

A significance test for empty corners in scatter diagrams

W.E. Bardsley[a,*], M.A. Jorgensen[b], P. Alpert[c], T. Ben-Gai[d]

[a]*Department of Earth Sciences, Water Research Unit, University of Waikato, Private Bag 3105, Hamilton, New Zealand*
[b]*Department of Statistics, University of Waikato, Private Bag 3105, Hamilton, New Zealand*
[c]*Department of Geophysics and Planetary Sciences, Tel-Aviv University, Ramat-Aviv, 69978 Tel-Aviv, Israel*
[d]*Department of Geography, Tel-Aviv University, Ramat-Aviv, 69978 Tel-Aviv, Israel*

Received 20 February 1998; received in revised form 11 February 1999; accepted 9 March 1999

Abstract

Regression analysis is usually the statistical tool of choice in hydrological studies when there is a strong correlation between two variables. However, weak correlations can also be of interest if a region within the scatter plot is data-free. This could direct attention to seeking some underlying physical process that might create regions with low probability of generating data points. A necessary prior requirement here is to verify that the data-free area in the plot is sufficiently large to be a real effect and not a visual illusion. This check can be most simply carried out in a hypothesis-testing framework. A permutation approach to hypothesis testing is suggested for the particular case where a data-free region occupies one of the corners of a scatter plot, and a test statistic Δ is presented for testing the statistical significance of the size of this ''empty corner''. Application to some rainfall data from southern Israel shows that the new test can sometimes yield higher levels of statistical significance than linear regression when applied to the same data.

Keywords: Time series analysis; Regression analysis; Pattern recognition; Israel; Rain

1. Introduction

The scatter plot of rainfall data shown in Fig. 1 illustrates a fan-shaped dispersal of data points with an absence of data in the upper left corner. ''Empty corner'' patterns of this type occur quite frequently with scatter plots of hydrological and climatological variables. One approach to analysis would be to apply a regression model which allows variation with respect to the distribution of errors about the regression relation. See, for example, Koenker and Bassett (1978); Williams (1997). Another option is to seek analysis models which do not assume any underlying unique functional relation between the variables. This paper is concerned with introducing one particular kind of function-free analysis for application to scatter diagrams, whereby data patterns like Fig. 1 are interpreted as indicative of some physical process leading to low probability of data generation in a corner region.

The regression and empty-corner interpretations lead to focusing on different physical mechanisms generating hydrological observations. A regression analysis applied to Fig. 1 would lead to seeking explanations for the increasing trend in the October rainfall totals. In contrast, a statistically significant empty corner raises the more direct question as to why there were no relatively wet Octobers in the earlier part of the record for this region of southern Israel. As

* Corresponding author. Fax: +64-7-856-0115.
E-mail addresses: web@waikato.ac.nz (W.E. Bardsley), maj@waikato.ac.nz (M.A. Jorgensen), pinhas@cyclone.tau.ac.il (P. Alpert), yuval@iprg.energy.gov.il (T. Ben-Gai)

0022-1694/99/$ - see front matter
PII: S0022-1694(99)00043-8

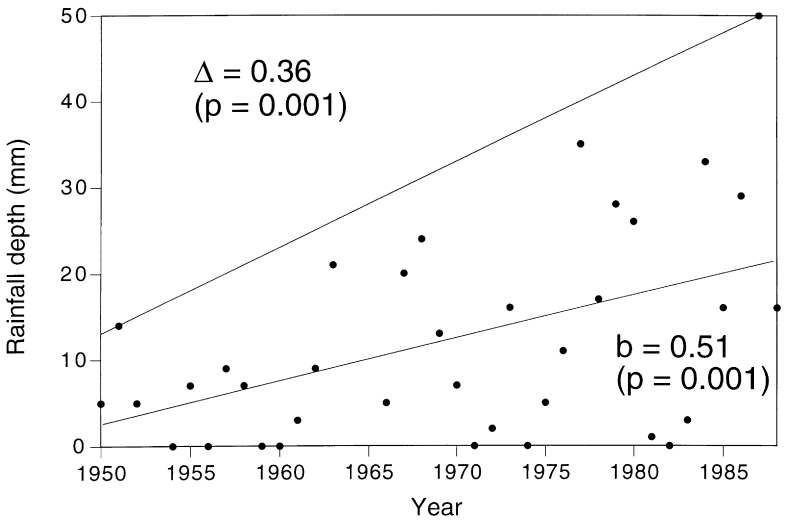

Fig. 1. Scatter plot of October rainfall totals as a function of time (1950–88) for Beit Govrin site, southern Israel. The upper line maximises a triangular empty area in the upper left corner. Δ is the test statistic for the empty corner and corresponds to the area proportion occupied by the empty corner. The regression coefficient *b* is the gradient of the least-squares line of best fit (lower line). The *p* values give the respective statistical significance of Δ and *b* as obtained from permutation testing.

indicated in Fig. 1, both the regression and empty corner models have similar strong statistical significance. However, if the evident empty corner is a non-random effect then the statistical significance of the regression gradient could be just an artefact of the absence of points in the empty corner-as opposed to a ''true'' underlying linear relation through the central portion of the data.

The immediate problem for any possible empty corner analysis centres on what constitutes a ''statistically significant empty corner''. That is, a significance test must be applied by which a given empty corner can be demonstrated to have a high probability of being a non-random effect and not just a visual illusion arising from a wide scatter of data points. The purpose of this paper is to present such a test (henceforth referred to as the Delta test) for application to empty corners in scatter diagrams, where the data-free corner region is defined to be triangular in form. The test can be readily generalised and is based on a permutation approach which avoids any distributional assumptions. Examples from southern Israel rainfall time series illustrate that the Delta test sometimes yields greater statistical significance than linear regression applied to the same data. Interested readers are referred to Ben-Gai et al. (1993) for climatological background, site locations, and possible causal mechanisms for the change in the Israel October rainfalls.

2. The Delta test

Suppose there are N data points on an x–y scatter plot, and define the ranked X and Y data values respectively as $X_1 \leq X_2 \ldots X_N, Y_1 \leq Y_2 \cdots Y_N$. Define the rectangular data space of area A by the four points (X_1,Y_1), (X_1,Y_N), (X_N, Y_N), (X_N,Y_1). The test statistic Δ is defined as the normalised area $\Delta = U/A$, where U is the maximum triangular corner area that can be achieved in a specified corner while still excluding all data points.

The measured area proportion Δ can be tested for statistical significance via permutation by calculating different values of Δ for a sufficiently large number of the $N!$ possible permutation orderings of the X values. The calculation must be with respect to the same empty corner as was used to obtain the original test statistic Δ from the data. A statistically significant result is obtained in the event of a sufficiently small value of p, where p is the proportion of permutation-derived Δ values which equal or exceed the original

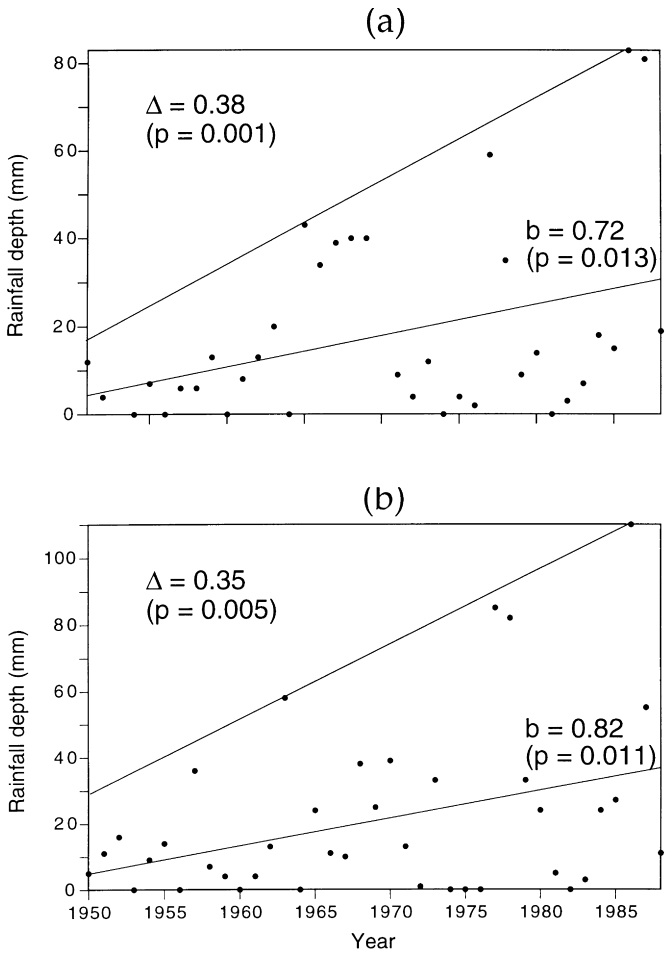

Fig. 2. Scatter plots of October rainfall totals as a function of time for southern Israel, showing higher statistical significance for the empty corners than for linear regressions: (a) Gvar'am site; (b) Nir Galim site.

recorded Δ value. That is, significance in the Delta test is obtained when the test statistic is located sufficiently far out in the right tail of the permutation distribution of Δ.

The maximum empty triangular corner does not always define the location of its internal boundary, for a given set of data points. This does not affect the Delta test, which is with respect to the triangle's area and not the position of its boundaries.

3. Examples

The Delta test was applied to some rainfall time series from southern Israel (Figs. 1–3). The frames of the figures have been drawn to coincide with the rectangular data space defined in the previous section (Section 2). Diagrams constructed in this way will always have at least four data points located on the frames.

For each permutation, the ordinary least squares regression gradient was calculated in addition to the Δ value. This allows fair comparison between the regression and empty-corner models. For the regressions, p is defined as the proportion of the permutation regression gradients which equal or exceed the least-squares gradient b, obtained from regression analysis of the original data. This represents a one-sided

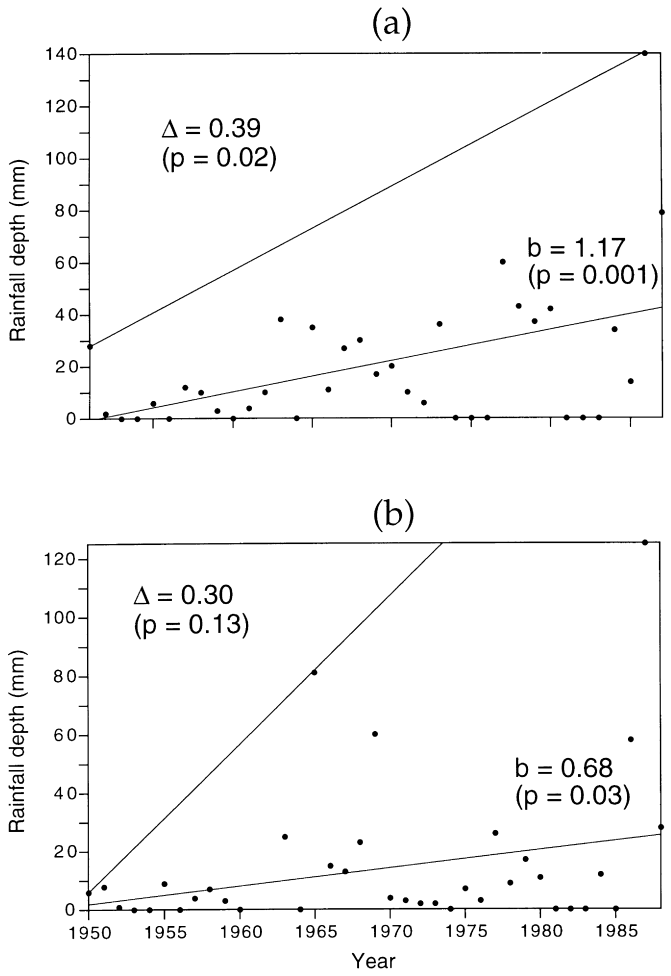

Fig. 3. Scatter plots of October rainfall totals as a function of time for southern Israel, showing higher statistical significance for linear regressions than for the empty corners: (a) Berurim site; (b) Sa'ad site.

significance test of the regression gradients. A sufficiently large number of permutations were generated from each data set to allow the listed *p* values to be calculated to the decimal place accuracy indicated in the figures.

An Excel spreadsheet macro ''Deltatest'' was written to carry out the Delta test procedure, and to generate the regression gradients. The macro can be applied to any *x*–*y* data and those interested in applying the Delta test to their own data can download the Excel workbook (see ''Electronic Supplements'' on this journal's homepage: www.elsevier.com/locate/jhydrol or www.elsevier.nl/locate/jhydrol). The Israel rainfall data sets used in this publication are also included, together with a description of the macro. The macro is reasonably efficient and for low *p* values the calculation time on current PC's should be not much more than 0.5 h for the precision needed for significance testing.

The data sets produced the not unexpected result that the Delta test has its highest significance when the empty corner is bounded by an approximately linear margin along the data points. This emphasises the fact that the empty corner is defined by a small subset of the data points. In contrast, a least-squares regression line is influenced by all data points and fares best from

a significance viewpoint when data points are not too far removed from the line. Fig. 1 represents something of a compromise between these two situations, with both regression and empty-corner models yielding similar significant p values. The plots in Fig. 2 have roughly linear upper data boundaries and the Δ values have higher levels of significance than the regression gradients. The plots in Fig. 3 have poorly-defined upper data boundaries and the regression gradients produce higher levels of significance than the empty-corner Δ values.

The regression approach is evidently more useful for detecting significance with the data shown in Fig. 3(b), where the Δ value is not statistically significant. This illustrates the sensitivity of the Delta test to even single data point departures from an approximately linear data field boundary. This has particular implications for when there are large numbers of data points involved, making it likely that one or two data points will appear in the low probability corner region. The Δ statistic would have little power in this situation but a generalisation of the Delta test is discussed in the next section (Section 4) whereby the ''empty corner'' is permitted to obtain some data points.

Because the Delta test is most powerful in association with approximately linear data boundaries, it may sometimes be useful to apply a transformation to the Y values if this achieves a better linearity of the data boundary. This is analogous to transforming data to improve linearity for the application of linear regression.

Permutation testing as described here becomes more awkward if the time series data has a high degree of serial dependence. This would make the permutation exercise very difficult because each permutation would have to preserve the correlation structure. The serial correlations for the data sets considered here were all low, so simple permutation was utilised. The highest r^2 for serial correlation was 0.22 for the Gvar'am site (Fig. 2(a)). All other sites had r^2 values not exceeding 0.14.

4. Discussion

The usual cautions need to be stated about retrospective statistical tests because the statistical hypotheses to be tested should be formulated prior to viewing the data. However, many hydrological and climatological data sets were not gathered as part of an experiment with a specific hypothesis in mind, and preliminary data inspection is often used to decide, for example, between the use one- or two sided parametric tests. Retrospective testing could also arise with use of the Delta test whereby the investigator might first view the scatter plot in order to determine which empty corner to test. To avoid such bias, a modified test statistic Δ^* could be defined as $\mathrm{Max}(\Delta a, \Delta b, \Delta c, \Delta d)$, where $\Delta a, \Delta b, \Delta c, \Delta d$ represent the largest empty corners achievable in each of the respective four corners, for a given data permutation. With respect to the examples considered here, a considerable number of rainfall stations in southern Israel display a sparsity of data points in the ''upper left'' corner over the time period 1950–1990 (Ben-Gai et al., 1993). It is this regional effect which in this case provides the justification for consistently selecting the upper left corner, rather than the form of any one data set.

The Delta test involves partitioning a data space into two sectors with and without data, and the internal boundary of the empty corner can be interpreted as a crude estimate of the true data field boundary over the range of the X variables. The physical nature of this boundary for the Israel data has been left undefined in this paper as it is not critical to the application of the test. It may be that the boundary is a true bound (zero probability of data points in the empty region) but in most hydrological or climatological situations it will be more meaningful to interpret the data field boundary as a zone of rapid transition to a region of relatively low probability of generating data points.

It is emphasised that Δ is a normalised area measure and its statistical significance is therefore with respect to the relative size of a specific area as opposed any particular boundary. For example, a statistically significant value of Δ cannot be used to infer statistical significance for the gradient of the internal boundary of the empty corner. If estimating the ''true'' underlying data boundary is of most interest, then stochastic frontier methods could be utilised-see, for example, Nepal et al. (1996) and references therein.

The Delta test is amenable to generalisation. For example, a generalised test statistic Δ_i could be defined as the maximum corner area achievable,

subject to exactly *i* data points being located within the (almost) empty corner concerned. This could be of use with large numbers of data points, as mentioned in the previous section. The test statistic Δ would then be classified as Δ_0. Another generalisation would be to substitute the triangular empty region with more flexible geometric forms. For example, a test statistic could be defined as the empty corner area which lies beyond a convex polygon covering all data points. The ''Deltatest'' algorithm in effect calculates this polygon on the way to obtaining Δ, so only minor code modification would be required for this generalisation. Whatever the nature of any subsequent generalisation of Δ, the permutation approach offers one practical way of testing statistical significance because the theoretical sampling distributions of the generalised test statistics would be difficult to derive. The difficulties of deriving sampling distributions are also avoided by bootstrap resampling methods, and this approach might be investigated as an alternative to permutation testing.

5. Conclusion

The Delta test gives a measure of the statistical significance of triangular empty corners in scatter diagrams and thus tests different aspects of the data than does regression. The test is less robust than regression against the effect of outliers in the corner region, but the Delta test can produce higher significance levels for fan-shaped data scatters when an approximately linear upper (or lower) boundary to a data field is evident on the data plot. Such data are not uncommon in hydrology and climatology so the Delta test and any subsequent generalisations should find some useful application.

Acknowledgements

We gratefully acknowledge the advice provided by Professor B.F.J. Manly, Department of Mathematics and Statistics, University of Otago. We thank the Israel Meteorological Service and Dr Alex Manes for providing the rainfall data.

References

Ben-Gai, T., Bitan, A., Manes, A., Alpert, T., 1993. Long-term change in October rainfall patterns in southern Israel. Theoretical and Applied Climatology 46, 209–217.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.

Nepal, S.K., Somers, G.L., Caudill, S.B., 1996. A stochastic frontier model for fitting tree crown shape in loblolly pine (*Pinus taeda L.*). Journal of Agricultural, Biological and Environmental Statistics 1, 336–353.

Williams, M.S., 1997. A regression technique accounting for heteroscedatic and asymmetric errors. Journal of Agricultural, Biological and Environmental Statistics 2, 108–129.