Testing for Elevated Risk near Point Sources, with Special Reference to Childhood Leukaemia near Nuclear Installations

John F. Bithell, Oxford

Summary

The excess of cases of childhood leukaemia and non-Hodgkin lymphoma near the Sellafield re-processing plant in Cumbria is well-documented, though a causal link with the plant is still not generally accepted. It has prompted numerous other enquiries into whether there might be similar excesses near other installations, although the operations at Sellafield are effectively unique in the UK. In 1990 the UK Electricity Industry sponsored a systematic study of incidence around all nuclear installations in Britain, using registration data from the Childhood Cancer Research Group based in Oxford. This was effectively the most comprehensive analysis of its kind to date.

The results of this investigation are used to illustrate the issues involved in selecting a suitable test procedure. Using a classical frequentist framework, it is shown how to construct the best test against a given alternative hypothesis - i.e. against a given Relative Risk Function. This approach defines a class of tests not previously used in the literature and here called „Linear Risk Score Tests". The selection of the best of these is of course hampered by our ignorance of the true alternative and the pertinent question therefore concerns the choice of a test that will be reasonably good against a range of alternatives. Simple LRS tests are compared with the order-restricted MLR test due to Stone, which has recently become popular. The analytical intractability of the latter and the discreteness of the data make it necessary to use simulation, but this method lends itself well to the comparison of tests using the Expected Significance Level as a criterion.

Further considerations include the choice of conditional or unconditional tests, a question of considerable practical importance. The issues are illustrated using the nuclear installation data.

1 Introduction

This paper is concerned with the question of the choice of test for detecting raised disease risk near a putative point source of hazard S, say. The question arises in particular with claims that the risk of childhood leukaemia is higher near nuclear installations and a major study of this question is used to exemplify the issues.

As usual in data analysis, there is no shortage of proposals for such tests, and although one might hope that the choice should not be critical, it often will be with the small data sets that occur in practice. Conclusions from such small data sets must necessarily be tentative and there is a particular danger in imputing more importance to a highly significant result than it really warrants. Conversely, it is clear that some tests, particularly the Standardised Incidence Ratio test, are less likely to demonstrate a real effect as significant owing to a lack of power. It is important to select a test of reasonable power in order to maximise the credibility of negative results. Although it is known how to construct the test which is most powerful against a given alternative hypothesis, we do not in practice know which alternative is most appropriate, so the more practical question is that of which tests are reasonably powerful against a range of alternatives.

We assume throughout this paper that the source of putative risk is specified *a priori*; without this assumption, the methods we consider are not valid.

2 Nature of the data

We suppose that data are available in the commonest form encountered in geographical epidemiology, namely counts X_i of cases of disease in small subregions A_i, $i = 1,2,...,k$, within a study region R, which will typically be a circle around S. We also assume that we are able to calculate expectations e_i for the counts under the null hypothesis H_0 that the risk is as determined by some externally available, perhaps national, rates. Under a typical alternative hypothesis H_1 these expectations will be replaced by $\lambda_j e_i$, where the λ_i may be thought of as subregion-specific relative risks. Independence of occurrence of the cases of disease leads to the assumption that the X_i are independently Poisson distributed with means $\lambda_i e_i$ [2].

It follows that the unconditional likelihood of the data can be written as

$$L = \left\{\frac{e^{-\theta}\,\theta^{n}}{n!}\right\} \times \left\{n!\prod_{i=1}^{k}\frac{p_i^{x_i}}{x_i!}\right\} \qquad 1$$

where $\theta = \Sigma\, e_i\lambda_i$; $p_i = \lambda_i e_i \,/\, \Sigma\, \lambda_j e_j$ and $n = \Sigma\, x_i$, the sum of the observed values of the X_i. This factorisation of the likelihood demonstrates how the information in the data can be partitioned into that due to an overall excess incidence described by the Poisson distribution of $N = \Sigma\, X_i$ and that due to spatial non-uniformity described by the multinomial distribution determined by the $\{p_i\}$.

3 Tests considered

The tests we consider are as follows:

3.1 The Standardised Incidence Ratio (SIR) test

This has been used traditionally in the investigation of risk around a point source S, and indeed it is natural to ask about the incidence within R. However, it is important to appreciate that this test is far from powerful against any realistic alternative; it is in fact most powerful against an alternative in which there is a step function of risk - elevated and constant within R but normal elsewhere. It is of course critically dependent on the size of R considered.

3.2 LRS Tests

Linear Risk Score (LRS) tests are based on a statistic

$$T = \Sigma x_i \log\lambda_i\,, \qquad 2$$

rejection being for values of $T \geq t_0$, say. This is equivalent to summing, for each case, a score defined as the log of the Relative Risk appropriate to his or her small area A_i. Each test is characterised by the set of λ_i and it is known to be most powerful against

alternatives in which the Relative Risk in each A_i is proportional to λ_i.

The difficulty with this of course is that the λ_i are not known in practice, so the practically important question becomes that of how to choose a set of scores that perform reasonably well against a range of alternative hypotheses. It has been found that the canonical test with scores equal to *1/distance* or *1/(distance rank)* meet this criterion. Other surrogates of distance can of course be used instead and to some extent these can be determined by scientific considerations. For example, environmentally mediated effects might argue for scores based on *distance* or *distance*^{2}, while an occupationally mediated effect might argue for the use of ranks.

Ranks are also less dependent on population distribution, but they start to lose power with small values of e_i, say appreciably less than unity. More sophisticated surrogates for distance can easily be substituted, for example to take account of wind direction or smoke plumes.

3.3 Stone's tests

Stone [9,10] proposed a class of tests based on the maximum likelihood ratio (MLR) test in which the λ_i are estimated so that they maximise the likelihood ratio, but subject to the restriction that they are non-increasing with distance from S, i.e.

$$H_1: \quad \lambda_1 \geq \lambda_2 \geq \lambda_3 , \ldots , \lambda_k \geq 1 . \qquad \mathbf{3}$$

Stone conjectured that such a test would be reasonably powerful against a wide class of alternatives in which the λ_i obey this condition. It is clear from the above analysis that this test cannot be most powerful against any possible alternative, though it may well be nearly 100% powerful in many cases.

The analysis associated with this estimation process is somewhat intractable [6], [7] and Stone therefore proposed the much simpler "Pmax" test, based on $\lambda^\wedge_1$.

This statistic has the natural attraction that it is the maximum value of the estimated Relative Risk as we move further out from S:

$$\text{Pmax} = \max_r \left[\sum_1^r X_i \right] = \lambda^\wedge_1 \qquad \mathbf{4}$$

It is dominated by early excesses and is therefore likely to be more powerful for alternatives in which the risk is concentrated at the origin than where it persists throughout R.

3.4 Conditional and unconditional tests

Reverting to the factorization of the likelihood exhibited in Equation (1), we can see that the first factor derives from the distribution of the total N of the cases and therefore contains the information employed by the SIR test. The second factor derives from the conditional distribution of the cases within R and therefore contains the information relating to the spatial distribution of the cases without reference to the overall incidence. To examine this distribution we would use the *conditional* version of the LRS test (or of either of Stone's tests); however, an overall excess would not contribute to this and, if we believed that the calculated expectations were really appropriate to R, it would be more sensible to use the *unconditional* version, based on the (Poisson) distribution for the data as a whole and using the whole of the information in the likelihood. If we use the conditional test, we may reject H_0 because of a *deficit* of cases in the outer part of R rather than an excess near S.

3.5 Execution of the tests

As indicated above, Stone's MLR test is difficult to execute analytically, though some numerical approximations are available [6]. By contrast, the LRS tests have the advantage that they have good normal approximations, at least when the expectations are not too small. In the case of the leukaemia data discussed below, the expectations are of the order of one case per ward on average and for this reason we carried out all the tests using simulations, i.e. by comparing the observed test statistics with values simulated under the null hypothesis. It is convenient to commpute all the test statisticssimultaneously using the same simulated data sets. Typically we used 1000 simulations, though this was increased where the estimated P-value was small or near to a critical value.

We also used one-tailed tests throughout. This seems reasonable since it is inconceivable that a greater distance from S implies a greater risk, at least under the distance model we have used.

4 Comparing the LRS and Stone's tests

The relative sensitivities of the tests described above could be assessed by estimating their power against given alternatives. However, following Stone [9], we prefer to use the Expected Significance Level (ESL) as an alternative to power. This criterion was introduced by Dempster and Schatzoff [5] and is defined as the expected value of the P-value of the test in repeated sampling under H_1. It has certain advantages over the power [4], in particular being easier to estimate by a simulation experiment, which is necessary when studying Stone's tests because of the intractability of the distributions involved. Although for many applications we recommend the use of unconditional tests, simulation experiments are best carried out for the conditional versions for simplicity; it seems likely that the relative sensitivities of different tests should be similar in the conditional and unconditional cases.

As an example of the comparison of ESL's, we show, in Figures 1-3, the results of simulation experiments with three different alternative hypotheses, using reciprocal, negative exponential and Gaussian functions to determine the Relative Risks at distance d:

$$f_1(d) = 1 + 10 / (1 + \beta d)$$

$$f_2(d) = 1 + 10\exp(-\gamma d) \qquad \mathbf{5}$$

$$f_3(d) = 1 + 2\exp(-(\delta d)^2)\,,$$

It is supposed that there are 25 subregions with equal expectations $e_i = 4$ and located at distances from S given by $\sqrt{\mathit{rank} - 0.5}$; this spacing corresponds approximately to a uniformly distributed population.

In each case, the decay parameter measured along the horizontal axis of the figure is the parameter multiplying the distance in the corresponding equation of (5), i.e β for f_1, γ for f_2 and δ for f_3. The greater the value of this parameter in each case, the faster this „Relative Risk Function" (RRF) decays towards unity and the more the excess risk is concentrated near S. In each case, the ESL has a minimum since we move from a situation in which the RRF is very attenuated and risk is elevated throughout R (which is not easily detected with the conditional test) to one where raised risk is very concentrated near S and affects only a few of the subregions. In between these extremes we have the smallest ESL's, where there is maximum differentiation between the risks in R, but allowing for the different weights given by each test to the nearer subregions.

The points + and • plotted in each case represent estimated logarithms (to base 10) of the ESL's of the Stone's MLR and Pmax tests respectively from 1000 simulations. The smooth curves represent log ESL's calculated using normal approximations for three LRS tests: those using *1/distance* and *1/rank* as scores and that with the score giving the most powerful test at the corresponding alternative, according to the value of the decay parameter determined by the position along the horizontal axis; as would be expected the latter curve provides a lower bound to the ESL's in each case. It should be noted that the LRS *1/rank* test can be expected to behave approximately like that with score *1/distance*^{2} in this example since we have assumed a population distribution that is approximately uniform spatially.

Looking first at Figure 1, we can see that, for the reciprocal RRF, the MLR test of Stoneis somewhat superior to Pmax for moderate values of β , presumably because the RRF decays rather slowly; for larger values of β there is little to choose between them. The two canonical LRS tests both perform as well as or better than Stone's tests across the range of parameter values; the *1/distance* test is slightly better than *1/rank*. This latter difference seems to be sustained across the whole range of values of the decay parameter.

The results in Figure 2 for the negative exponential RRF also show a superiority of the MLR over Pmax for small values of the decay parameter γ - i.e. where the risk extends over most of R - and in this range it is again comparable with the LRS tests. As the risk is more concentrated near S, however, Pmax assumes a clear advantage over the MLR test and its ESL converges to that of the two LRS tests, for which the initial advantage of *1/distance* also disappears.

For the Gaussian RRF f_3 we reduced the maximum Relative Risk in Equation (5) to produce ESL's of similar magnitudes to those for f_1 and f_2. The results, shown in Figure 3, are broadly similar to those for the negative exponential RRF, though the superiority of the MLR test for attenuated risks is more pronounced and sustained for greater values of δ.

5 Application to childhood leukaemia

The theoretical work on the properties of the statistical tests discussed above was motivated by the need to select a good test for application to a data set on childhood leukaemia and non-Hodgkin lymphoma (L & NHL) in relation to nuclear installations. This data set was obtained from the National Registry of Childhood Tumours maintained in Oxford by the Childhood Cancer Research Group; it consisted of 11,283 cases of leukaemia and non-Hodgkin lymphoma registered under the age of 15 in England, Wales or Scotland between 1966 and 1987. The cases were allocated to one of 9836 electoral wards (or equivalent areal units), for which expected numbers of cases were calculated allowing for various socioeconomic and demographic factors [3]. A fuller account of these calculations will be found in Bithell et al. [1].

23 sites of nuclear installations in England and Wales were identified, ten of which are generating stations operated by Nuclear Electric (NE) plc, the remainder being various research and production facilities. The wards whose population centroids were within 25 km of each site were identified and the distances of these centroids from the site were calculated. The definitive test used was the LRS *1/rank* test, though the results of Stone's MLR test were also examined and reported where significant.

The results of these analyses are described in full in Bithell et al. [1]. In summary, none of the 23 circular regions examined showed a significantly elevated SIR. This was true even for that around Sellafield, (24/18.5) where the notorious excess is concentrated in the nearest ward of Seascale (6/0.51). The LRS test, however, shows a very highly significant result for Sellafield ($P = 2 \times 10^{-5}$) and a more modest value of $P = 0.031$ for the Atomic Weapons Research Establishment at Burghfield. The latter result seems to be likely to be due to chance as it is believed that radioactive discharges from the plant have been insignificant. The NE generating station at Hinkley Point gave a P-value of 0.10 using the LRS test; for both this site and Burghfield Stone's MLR test returned a larger P-value. An analysis of the effect of using a conditional rather than an unconditional test is given by Bithell [2].

6 Discussion

Experience in analysing the leukaemia data confirms what the ESL studies indicate, that the LRS *1/rank* test is quite a good all-purpose test of risk in relation to a point source, though it is worth repeating the caveat that this may be less true if the expectations are very small; in particular Sharp et al. [8] report unfavourable results of power calculations in relation to areal units in Scotland which are smaller than those in the England and Wales study. These authors favoured Stone's MLR test as their definitive test, though the LRS *1/distance* test should also work better than *1/rank* for very small areas.

The ESL simulation methodology can be used to explore other issues, such as the effect of the aggregation of the sub-regions into larger areal units and variability in the expectations. There is still plenty of scope for investigations of such questions, but we feel that a useful start has been made and that compelling arguments have been established for using an analysis that is more sensitive to spatial distribution than the SIR test.

Acknowledgements

This research was supported in part by a grant from Nuclear Electric plc. The work of all who helped in the acquisition and preparation of the childhood leukaemia data is gratefully acknowledged, especially that of the staff of the Childhood Cancer Research Group. The CCRG is supported by the Department of Health and the Scottish Home and Health Department.

References

1. Bithell J.F., Dutton S.J., Draper G.J., and Neary N.M.: Distribution of childhood leukaemias and non-Hodgkin's lymphomas near nuclear installations in England and Wales. British Medical Journal, 309 (1994) 501-505.
2. Bithell, J.F.: The choice of test for detecting raised disease risk near a point source. Statistics in Medicine, 14 (1995) 21, 2309-2322.
3. Bithell J.F., Dutton S.J., Neary N.M. and Vincent T.J.: Use of regression methods for control of socio-economic confounding. Journal of Epidemiology and Community Health. 49 (1995) Suppl 2, S15-S19.
4. Bithell, J.F. and Dutton, S.J.: Optimal frequentist procedures for detecting raised disease risk near point sources. Epidemiology Proceedings of the American Statistical Association, Joint Meetings 1995 (1996) 1-10.
5. Dempster, A.P. and Schatzoff, M.: Expected Significance Levels as a Sensitivity Index for Test Statistics. Journal of the American Statistical Association, 60 (1965) 420-436.
6. Lumley, T.: Efficient execution of Stone's likelihood ratio tests for disease clustering. Computational Statistics and Data Analysis 20 (1995) 5, 499-510.
7. Robertson T:, Wright F.T. and Dykstra R.L.: Order Restricted Statistical Inference, (1988) London: Wiley.
8. Sharp, L., Black, R.J., Harkness, E.F. and McKinney, P.A.: Incidence of childhood leukaemia and non-Hodgkin's lymphoma in the vicinity of nuclear sites in Scotland, 1968-93. Occup. Environ. Med., 53 (1996) 823-31.
9. Stone, R.A.: Statistical methodology and causal inference in studies of the health effects of radiation. D.Phil. thesis (1986) University of Oxford.
10. Stone, R.A.: Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. Statistics in Medicine, 7 (1988) 649-660.

Figure 1.
Expected Significance Levels for Stone's tests estimated by 1000 simulations (points) and for three LRS tests using a normal approximation (smooth curves). Reciprocal RRF.

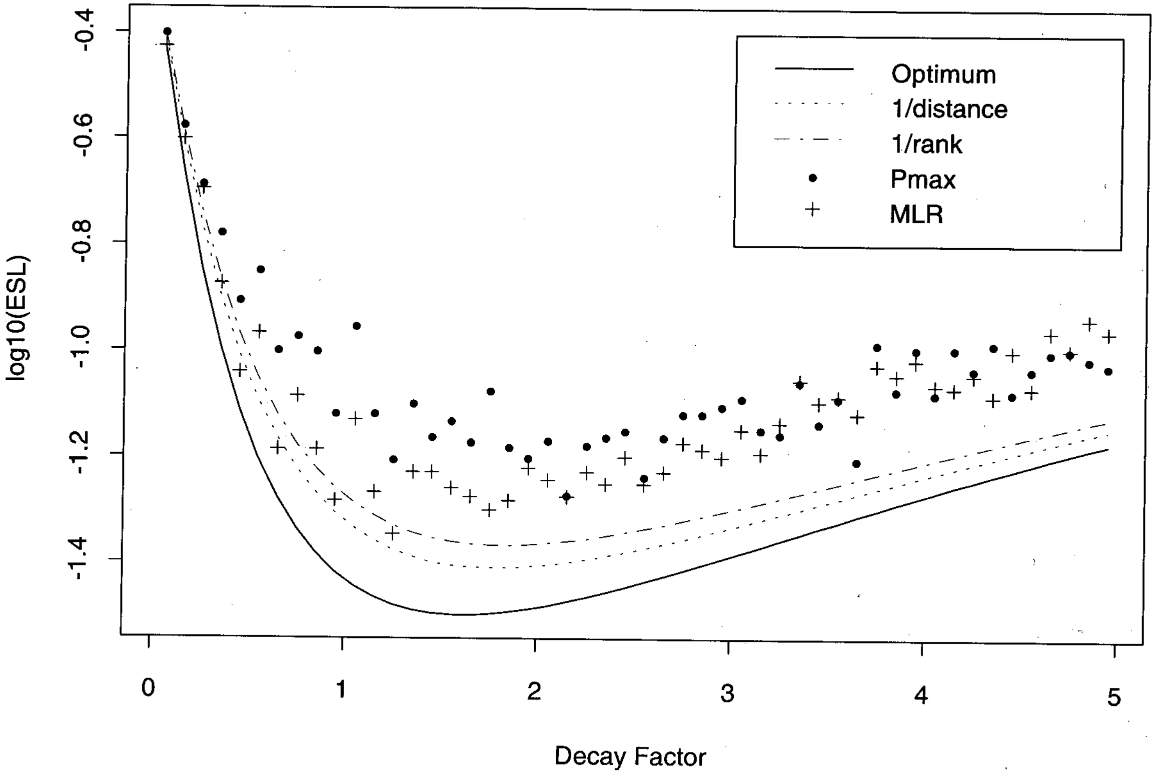


Figure 2.
Expected Significance Levels for Stone's tests estimated by 1000 simulations (points) and for three LRS tests using a normal approximation (smooth curves). Negative exponential RRF.

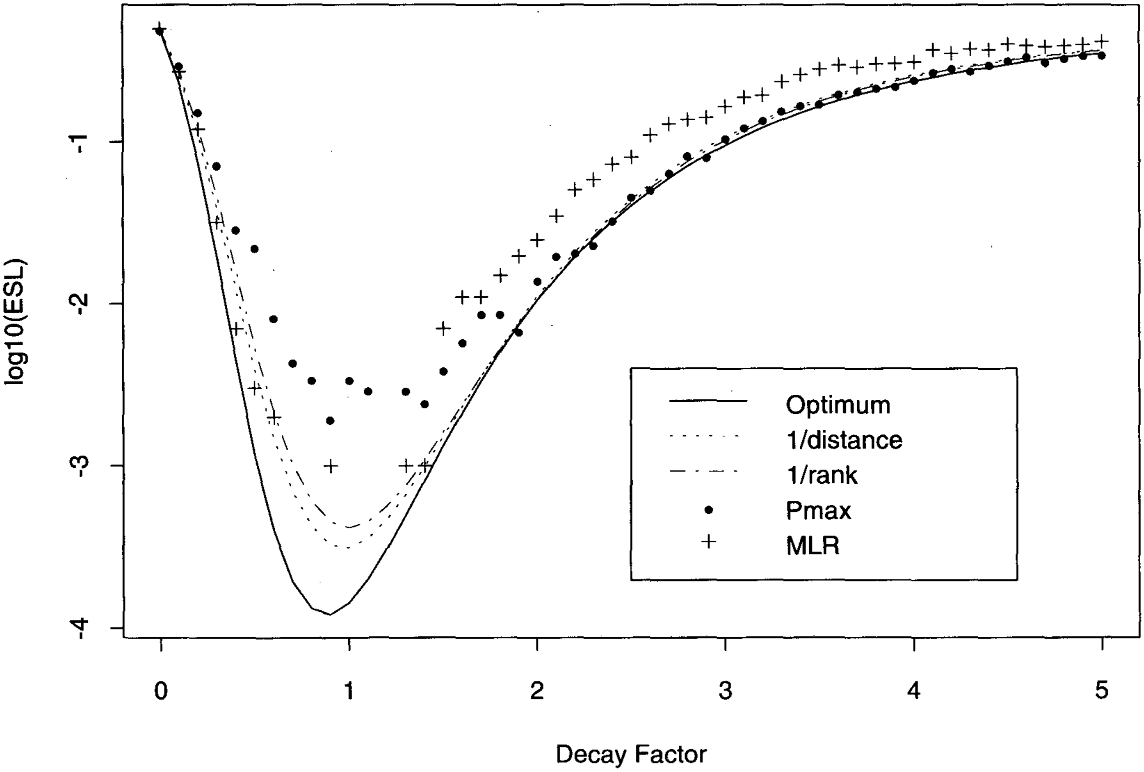
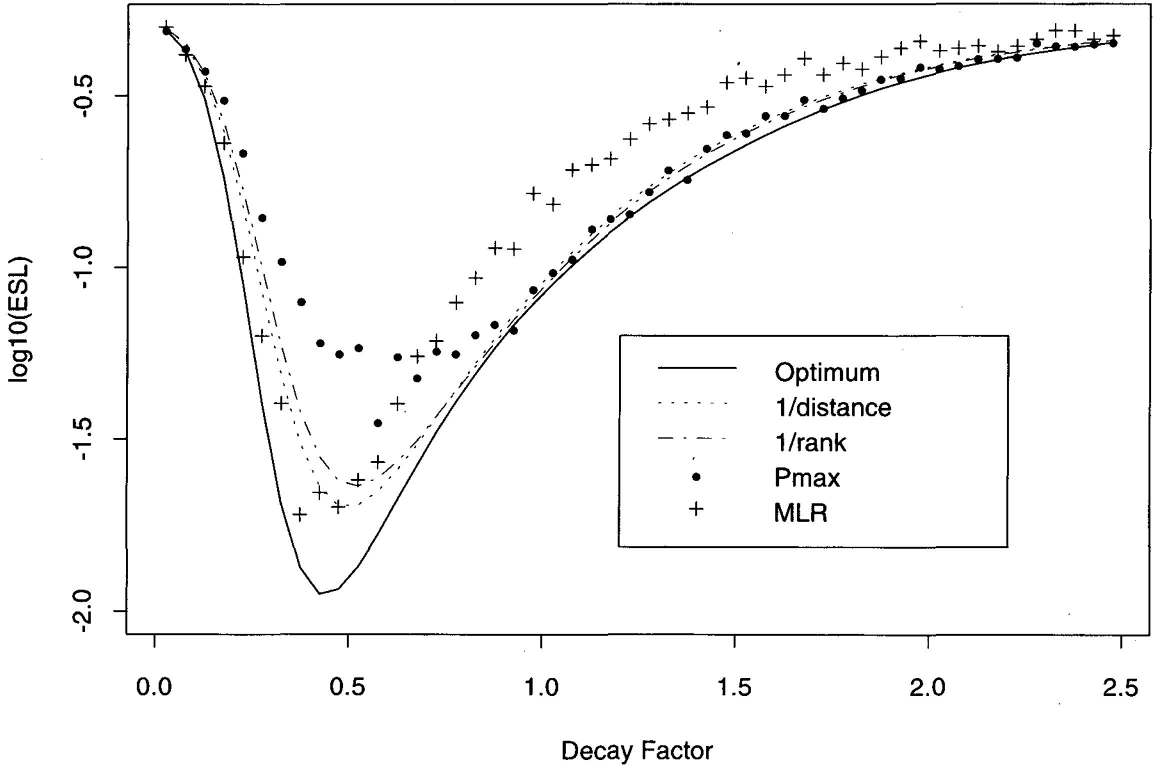


Figure 3.
Expected Significance Levels for Stone's tests estimated by 1000 simulations (points) and for three LRS tests using a normal approximation (smooth curves). Gaussian RRF.


log10(ESL)
Decay Factor
Optimum
1/distance
1/rank
● Pmax
+ MLR