1-1-1995

Information Theoretic Mask Analysis of Rainfall Time Series Data

Martin Zwick
Portland State University, zwick@pdx.edu

Hui Shu
Portland State University

Roy Koch
Portland State University

Citation Details

Zwick, M., Shu, H., and Koch, R. (1995). [Post-print]. "Information-Theoretic Mask Analysis of Rainfall Time-Series Data." Published in Advances in Systems Science and Applications, 1, pp.154-159.

Advances in Systems Science and Applications (1995) **Special Issue I**, 1–000

Information-Theoretic Mask Analysis of Rainfall Time Series Data

MARTIN ZWICK, HUI SHU, and ROY KOCH*

Systems Science Ph.D. Program and Dept. of Civil Engineering, Portland State University, U.S.A.*

(Received)

This study explores an information-theoretic/log-linear approach to multivariate time series analysis. The method is applied to daily rainfall data (4 sites, 9 years), originally quantitative but here treated as dichotomous. The analysis ascertains which lagged variables are most predictive of future rainfall and how season can be optimally defined as an auxiliary predicting parameter. Call the rainfall variables at the four sites A...D, and collectively, Z, the lagged site variables at t-1, E...H, at t-2, I...L, etc., and the seasonal parameter, S. The best model, reducing the Shannon uncertainty, u(Z), by 22%, is HGFSJK Z, where the independent variables, H through K, are given in the order of their predictive power and S is dichotomous with unequal winter and summer lengths.

1. Introduction

This study is an application of mask analysis to time series studies, more specifically, to rainfall forecasting. The term, "mask analysis," derives from the "general systems methodology" of Klir (1985), Krippendorff (1986) and their colleagues, which is extensively based in information theory. These methods, when applied to multivariate nominal statistical data, substantially overlap what in the social sciences are called "log-linear" techniques (Bishop Feinberg and Holland 1978, Knoke and Burke 1980). The primary purpose of the study is methodological: to test these information-theoretic techniques on a particular multivariate time series problem, and to enrich these methods with the statistical assessments common in log-linear analysis. A secondary purpose is to demonstrate the applicability of these methods to the water resources application area.

The data to be analyzed consist of daily rainfall measurements at four collecting sites over the nine-year period from 1982 to 1990, more specifically for 3104 days during this period. While the data were originally quantitative (inches of rainfall), they are here discretized into two states: rain or no rain. The methods used, however, are fully applicable to multichotomous data. No single program available to us could handle all the needed calculations, so various programs were used jointly: *SHANNON* (Hosseini, Zwick, and Anderson), *CONSTRUCT* (Krippendorff 1986), *GSPS* (Klir 1985, Elias 1988), *CHISQ* (Anderson).

ISSN 1078-6236

Table 1. Mask analysis framework

day:	t-3	t-2	t-1	t
site: 1	M	I	E	A
2	N	J	F	B
3	O	K	G	C
4	P	L	H	D
season:	S3	S2	S1	S

2. Models and Their Evaluation

The rainfall variables defined for the four sites at time t are called A, B, C, and D; at time t-1: E, F, G, and H; at time, t-2: I, J, K, and L; etc., as shown in Table 1. In addition, a seasonal variable, S, is used, as shown in the table. For convenience, we will refer to the dependent variables A...D, taken collectively, as the aggregate (16-state) variable, Z. An unspecified dependent variable will be called R. The aggregate variable, Y, will be the set of *possible* independent variables, e.g., if three lags are considered, E...P and S...S3 or, if two are considered, E...L and S...S2.

The question which mask analysis addresses is this: which subset, Y′, of the set, Y, maximally reduces the information-theoretic uncertainty (Shannon entropy), u(Z),

$$u(A, B, C, D) = u(Z) = -\sum p(Z)\, log_2\, p(Z).$$

We want to define Y′ such that

$$\triangle u = u(Z) - u(Z|Y') = u(Z) - u(Y', Z) + u(Y')$$

is statistically significant and as large as possible. $\triangle u$ will sometimes be expressed as a percentage, $\triangle u/u$(Z). It is useful also sometimes to calculate $\triangle u/u(Y')$, the "predictive power" of Y′, i.e., the uncertainty reduction in Z normalized by the amount of information in Y′ used to achieve that reduction. It should be realized that because of the logarithm in the definition of u, a small $\triangle u/u$(Z) can actually indicate high predictability. One could, for example, construct a hypothetical 2x2 contingency table where a rather small 8% reduction of uncertainty corresponds to a relatively large shift of probabilities for the dependent variable from (.5,.5) to (.67,.33) when the independent variable is known.

The algebraic uncertainty analysis which follows is based on the exposition of Krippendorff (1986). We evaluate only models of the form Y:Y′Z, where Y is the set of possible independent variables being considered and Y′ is the subset of Y actually used to predict Z. That is, each model considered here consists of two contingency tables (probability distributions), with overlapping variable sets: one for the variables in Y and the other for the variables in Y′ and in Z. When Y′=Y, we have the "saturated" model (m_0) for which Y:YZ is written simply as YZ. The constant Y component in these models reflects the fact that in systems where one distinguishes between independent and dependent variables, models must always have a component which groups together all independent variables. This assures that the models "cover" the same set of variables, so that they can be compared.

These Y:Y′Z models are only a subset of the full set of possible models. They have the virtue, however, of not having loops, which allows us to calculate statistical quantities easily in simple algebraic expressions. (A model Q:R:S:... has no loops if after repeatedly removing variables unique to individual components and also components embedded in other components one arrives finally at the null set of variables.)

In this paper, we always compare two models, a reference model of form Y:Y′Z, and a tentative model, Y:Y′QZ, which adds an additional predicting variable, Q, from the set Y to the predicting set Y′. We compute the increment of uncertainty reduction which Q produces, i.e., u(Z|Y′) - u(Z|Y′Q), and test the statistical significance of the deviation of this quantity from zero. If this deviation is significant, the addition of Q to the set of predicting variables is warranted.

We start with Y′ being the null set, choosing for a reference model the bottom model, Y:Z, which says that Z is independent of Y. We then increase the set of predicting variables one at a time, ascending via a sequence of Y:Y′Z models towards the top model, YZ, stopping at that particular Y:Y′Z where further ascent is statistically unjustified. At this point we simplify our notation by dropping the Y component and calling Y′Z our model. (The actual procedure is to try all single predictors, pick the best, try all pair predictors, pick the best, etc., but to make statistical comparisons between two models, one must be a descendent of the other.)

To compare two models, one computes $\triangle L^2$, the change in the likelihood-ratio Chi-square, and $\triangle$df, the change in the degrees of freedom, between the models as follows

$$\triangle L^2(Y\!:QY'Z \to Y\!:Y'Z) = 1.386NI(Y\!:QY'Z \to Y\!:Y'Z)$$

$$\triangle df(Y\!:QY'Z \to Y\!:Y'Z) = df(Y\!:QY'Z) - df(Y\!:Y'Z)$$

where N is the sample size and $I(m_i \to m_j)$ is the information distance between model i and model j. With these, we test the null hypothesis, H_0: Y:QY′Z = Y:Y′Z, i.e., $\triangle u =$ u(Z|Y′) - u(Z|Y′Q) = 0 by consulting the Chi-square table with a cut-off probability (α) of making a Type I error. If we reject the null hypothesis, it means that Y:QY′Z captures information in the data not captured by Y:Y′Z, i.e., that the nonzero $\triangle u$ is statistically signficant. If we cannot reject the null hypothesis, the nonzero $\triangle u$ is *not* statistically significant and the addition of Q to the predicting set cannot be justified.

The information distance, $I(Y:QY'Z \to Y:Y'Z)$, can be written as a difference of transmissions,

$$I(Y\!:QY'Z \to Y\!:Y'Z) = T(Y\!:Y'Z) - T(Y\!:QY'Z).$$

For a model, m_j, without loops, $T(m_j) = u(m_j) - u(m_0)$; thus

$$I(Y\!:QY'Z \to Y\!:Y'Z) = u(Y\!:Y'Z) - u(Y\!:QY'Z).$$

In general for composite variables Q and R, $u(Q:R) = u(Q) + u(R) - u(Q \cap R)$, where the intersect selects the variables common to Q and R. Hence,

$$u(Y\!:Y'Z) = u(Y) + u(Y'Z) - u(Y') = u(Y) + u(Z|Y')$$

$$u(Y\!:QY'Z) = u(Y) + u(QY'Z) - u(QY') = u(Y) + u(Z|QY')$$

and thus,

$$I(Y\!:QY'Z \to Y\!:Y'Z) = u(Z|Y') - u(Z|QY'),$$

that is, the information distance is the *additional* uncertainty reduction achieved by Q.

The difference in degrees of freedom between two models is

$$\triangle df(m_i \to m_j) = df(m_i) - df(m_j).$$

Table 2. Determination of number, starting month, and length of seasons

n_S	mo_1	△u/u(R)	△u/u(S)	
2	10	2.90%	2.83%	equal
2	11	4.44%	4.35%	
2	12	2.51%	2.53%	
2	1	1.07%	1.05%	
2	2	0.02%	0.02%	
2	3	0.70%	0.69%	
4	11	4.74%	2.32%	
6	11	5.45%	2.06%	
12	11	6.52%	1.77%	
2	11	5.00%	5.12%	unequal

The degrees of freedom of a model of form A:B,

$$df(A\!:\!B) = df(A) + df(B) - df(A \cap B),$$

where the intersect operator selects the variables common to A and B (e.g., for the model $Y\!:\!QY'Z$, the intersect selects QY'). This yields, in the present case,

$$\triangle df(Y\!:\!QY'Z \to Y\!:\!Y'Z) = df(QY'Z) - df(QY') - df(Y'Z) + df(Y').$$

3. Preliminary Analysis

First, we estimate the number of usable predicting variables. For 3104 data points and assuming about 5 data points per cell (a Chi-square rule of thumb) we have a limit of about 600 cells in our contingency table for both independent and dependent variables. This means 9 to 10 site variables, i.e., 2^9=512 to 2^{10}=1024 cells, hence 5 to 6 predicting dichotomous variables. Since using t-2 lags, even without season, would involve 8 predicting variables, we can probably safely ignore all t-3 lags.

We now consider the seasonal variable, S, and decide how many seasonal states to allow and for what temporal periods. Let $\triangle u = u(R) - u(R|S)$, for $R =$ A...D. Table 2 gives the average, over the 4 individual sites, of both $\triangle u/u(R)$, the % reduction in uncertainty, and $\triangle u/u(S)$, the predictive power (efficiency) of S, as a function of the number of seasons, n_S, and the starting month, mo_1. The table indicates the following. (1) For 2 seasons, the optimum month to begin winter is November (mo_1=11). (2) Although greater uncertainty reduction is achievable with additional seasons, predictive power is better for 2 than for 4, 6, or 12. Using season in the model requires giving up degrees of freedom of lagged site variables for degrees of freedom of S, and it seems unlikely that a multichotomous season would offer any predictive advantage. (3) Additional improvement of uncertainty reduction and predictive power is gained by making the 2 seasons unequal (7 months winter + 5 months summer). Greater inequality of season length (calculations not shown), however, does not improve uncertainty reduction. The conclusions drawn from these calculations are not actually definitive, as these ratios are not statistical measures with error probabilities we can calculate. However, a mask analysis reported briefly in the next section provides some supporting evidence.

Table 3. Mask analysis for lags EFGH JKL and season, S (key values are dotted)

		$\triangle u_{cum}$	$\triangle u_{incr}$	$\triangle L^2$	$\triangle$df	α
0. u(Z|EFGHJKLS)	2.08	36.2	10.5	1003	1920	1.00
1. u(Z|FGHJKLS)	2.32	28.8	8.2	852	960	0.99
2. u(Z|FGHJK S)	·2.52	·22.4	6.0	668	480	·0.00
3. u(Z|FGHJ S)	2.69	17.5	3.6	409	240	0.00
4. u(Z|FGH S)	2.78	14.4	2.7	322	120	0.00
5. u(Z|FGH)	2.86	12.4	2.5	302	60	0.00
6. u(Z|GH)	2.93	9.8	3.1	395	30	0.00
7. u(Z|H)	3.03	6.9	6.9	926	15	0.00
8. u(Z)	3.25	–	–	–	–	–

4. Mask Analysis

With the seasonal variable defined, we do mask analysis as shown in Table 3. We do a bottom up stepwise analysis to select a subset of predicting variables from the set Y = EFGH JKL S. This starting set was chosen on the basis of preliminary calculations, not shown here, where mask analysis was done *without* using the seasonal variable. In this earlier analysis for which $u(Z|Y) = 2.61$, the optimum predictors were EFGH KL and the next best predictor was J. We include these predictors in Y, along with S; it is unlikely, because of this earlier run, that we need to include I. (A simpler, one-step, procedure would have been to define S and just start with Y = EGH IJKL S.)

Essentially, our objective here is to select 6 of the possible 8 independent variables (four t-1 and three t-2 lags + season), i.e., to determine which of the 8!/(6!2!)=28 models should be used.

Table 3 shows that the best predictors are FGH JK S (model 2). Compare this to the best predictors obtained when season was not used, namely EFGH KL. The lagged site variables E and L are replaced by J and S. The final model is FGHJKS Z, and the predictive order for the independent variables, from most predictive to least, is: (most) H G F S J K (least). The model achieves a $u(Z|Y) = 2.52$ compared to $u(Z|Y) = 2.61$ for the earlier model where season was not considered. Measuring from the reference level of $u(Z) = 3.25$, consideration of season improves the predictive model from a 19.7% to a 22.4% reduction in uncertainty. It is interesting that E, the t-1 lagged value of site 1, does not appear in the model, i.e., is a weaker predictor than the t-2 lags (J and K) of sites 2 and 3. This is consistent with I being the weakest predictor of the t-2 lags. Site 1 also had much smaller values of $\triangle u/u(R)$ and $\triangle u/u(S)$ than the other 3 sites.

Season is less predictive than each of the t-1 lagged site variables, HGF, but more predictive than each of the t-2 lagged site variables, JK. It may be surprising that season is not a stronger predictor, but the lagged variables (especially the t-1 lags) intrinsically capture seasonal information in the sense that runs of rain or no-rain are more likely in the winter and summer, respectively. The uncertainty reduction is not very sensitive to the choice of the last two predictors. For example, a model, FGHKLS Z, has $u(Z|Y) = 2.55$, which is only slightly worse than the 2.52 of the best model. The previous determination that S should be dichotomous was also checked with a mask analysis using a 4-state seasonal variable, which is a composite of an s_1, specifying two equal seasons, and an s_2, splitting winter and summer into early and late halves. Consistent with the results shown in Table 3, the final model was FGH s_1 JK, and s_2 was omitted.

Table 4. Final model, FGHJKS Z; superscripts give predictive order

time	t-2	t-1	t
site: 1	I	E	A
2	J^5	F^3	B
3	K^6	G^2	C
4	L	H^1	D
season:	S2	S1	S^4

5. Discussion

In summary, our best predictors are Y = FGH JK S with $u(Z|Y) = 2.52$, an overall uncertainty reduction of 22% from the initial $u(Z) = 3.25$. This final model is shown in Table 4. Improvements might be possible, however, by further information-theoretic analyses, as only a small subset of the possible models have been considered. The model is used for predictive purposes simply by computing the conditional probabilities, p(Z|Y), from which one can calculate the probabilities of rainfall one day ahead.

Nothing in the present approach is intrinsically dependent upon variables being dichtomous. We could have "binned" the rainfall data into more than two states and thus approximated a treatment of rainfall as a quantitative variable. Optimal binning, however, is non-trivial. Alternatively, the FGHJKS Z model could be taken as a starting point for full quantitative modeling by other techniques. More generally, the approach used here is broadly applicable to multivariate time series analysis of nominal variables or quantitative variables with unknown non-linear relations when large data sets are available.

6. Acknowledgements

This work was supported in part by funds provided (to RK) by the Water Supply Forecasting Center of the U.S. Department of Agriculture Soil Conservation Service. One of the authors (MZ) thanks Jamshid Hosseini and Doug Anderson for introducing him to the log-linear literature and for many stimulating discussions. We also thank Klaus Krippendorff for the use of *CONSTRUCT*, Doug Elias for help with *GSPS*, and Doug Anderson for the use of *CHISQ*.

References

Bishop, Y.M.M. Feinberg, S. and Holland, P. (1978). *Discrete Multivariate Analysis.* Cambridge: MIT Press.

Elias, D. (1988). *General Systems Problem Solver: A Framework for Integrating Systems Methodologies. Ph.D. Dissertation.* Systems Science Dept., State University of New York-Binghamton.

Klir, G. (1985). *The Architecture of Systems Problem Solving.* New York: Plenum Press.

Knoke, D. and Burke, P. (1980). *Log-Linear Models (Quantitative Applications in the Social Sciences Monograph No. 20).* New York: Sage Publications.

Krippendorff, K. (1986). *Information Theory. Structural Models for Qualitative Data (Quantitative Applications in the Social Sciences Monograph No. 62.* New York: Sage Publications.