WORM WEATHER: A PRELIMINARY LOOK AT A LIMITED SAMPLE OF DATA

R. L. Winkler

January 1974

WP-74-1

The data available represented 100 years at one location, and each year was summarized as a "1" (good weather), a "2" (intermediate weather), or a "3" (bad weather). The objective was to investigate this limited sample of data while waiting for more extensive data to arrive at IIASA.

At one point it was thought that it might be possible for the purposes of the worm study to drop the "2"'s from consideration (on the grounds that a "1" helped the worms, a "3" hurt the worms, and a "2" didn't have much of an effect at all). Looking just at the "1"s and "3"s, there are 22 runs in the data. Under independence, the expected number of runs is 33.9 and the variance is 14.6. Thus, the observed number of runs represents a standardized value of -3.12. This suggests that there might be some dependence in the process. By way of comparison, if the "1"s are dropped, there are 34 runs of "2"s and "3"s (mean = 35.6, variance = 15.67), and the standardized value is -0.40. If the "3"s are dropped, there are 25 runs of "1"s and "2"s (mean = 27.0, variance = 12.71), and the standardized value is -0.56. Thus, the only pairwise sequence showing reasonably strong non-independence is the sequence involving the "1"s and "3"s.

Looking at the entire sequence (i.e., not looking at the different types of weather in a pairwise fashion) yields the following data:

Number of Runs of a Given Length (# of Years)

Length	"1"	"2"	"3"	Total
1	5	15	6	26
2	7	6	6	19
3	0	0	10	10
4	0	0	0	0
5	0	0	0	0
6	1	0	0	1
Total runs	13	21	21	55
Total observations	(25)	(27)	(48)	(100)
Mean obs./run	1.92	1.29	2.29	1.82

observations per run, whereas these averages were 1.92 for the "1"s and 2.29 for the "3".

Another way to investigate the data is to look at transition matrices. In terms of one-step transitions, the data are as follows:

		Weather in Year t		
		"1"	"2"	"3"
Weather in Year t-1	"1"	12	5	8
	"2"	7	6	14
	"3"	5	16	26

This yields the following matrix of estimated transition probabilities:

$$\begin{array}{ll} & \begin{array}{ccc} & \text{To} & \\ \text{"1"} & \text{"2"} & \text{"3"} \end{array} \\ \begin{array}{ll} & \text{"1"} \\ \text{From} & \text{"2"} \\ & \text{"3"} \end{array} & \begin{bmatrix} .48 & .20 & .32 \\ .26 & .22 & .52 \\ .11 & .34 & .55 \end{bmatrix} \end{array}$$

If one ignored transitions and just estimated marginal probabilities, the estimates would be .25 for "1", .27 for "2", and .48 for "3". Thus, part of the apparent persistence in the "3"s appears to be caused by the large number of "3"s, although the estimate of p_{33} (.55) is slightly greater than the estimate of p_3 (.48). Of the diagonal elements of the estimated transition matrix, the estimate of p_{11} deviates the most from the corresponding marginal probability. However, this is probably due to the fact that the one long run in the data (a run of 6 years) was a run of "1"s.

In terms of two-step transitions, the data are as follows:

"1"	"1"	4	5	3
	"2"	0	4	1
	"3"	0	5	2
"2"	"1"	7	0	0
	"2"	4	0	2
	"3"	0	0	14
"3"	"1"	0	0	5
	"2"	3	2	11
	"3"	5	11	10

These data provide some unusual results. For example, all seven "2""1" sequences were followed by another "1", and all 14 "2""3" sequences were followed by another "3". The three-step transitions also provide some unusual results. However, the amount of data is so limited that little faith can be placed in the two-step and three-step results.

Another possible approach is to consider an autoregression, and this was done with a constant term and two lagged terms. The estimated regression line (with standard errors of coefficients in parentheses) is

$$y_t = \underset{(.27)}{1.91} + \underset{(.10)}{.33}y_{t-1} - \underset{(.10)}{.18}y_{t-2}.$$

The simple autocorrelation in the sequence is .28, so an autoregresssion with just one lagged variable would explain about 8 percent of the variation (sample variance = .66) in the data. The autoregression with two lagged variables explains just over 10 percent of the variation. Neither of these results is very impressive, but that's not too surprising considering the nature of the data (i.e., the data can only take on the values "1", "2", and "3"). Without the restriction imposed by the linear autoregression model, an estimate of w^2, which is the proportion of variance in y_t that can be accounted for by knowledge of y_{t-1} and y_{t-2}, is .44. Thus, while only 10 percent of the variance can be accounted for by a linear autoregression with two lagged terms, another 34 percent can be accounted for by a nonlinear relationship involving two lagged terms.

immediately preceding years, although the relationship is probably not too strong. In order to investigate this in more detail, more data and better data are needed, and apparently they are on the way. Ideally, it would be nice to have data going back more than 100 years, but that appears not to be possible. Data for approximately the past 60 years are available for 10 different locations, however. Moreover, the information in this data is much greater than the information in the current set of data. The new set of data will include several summary statistics for each year, and these statistics should provide more information than the "1"s, "2"s, and "3"s currently available.

Of course, the analysis of the weather data should not proceed in isolation. In particular, the relationship of the weather to the budworm population should be considered carefully in order to attempt to pinpoint what sorts of "weather events" are of special interest. This will hopefully increase the efficiency and usefulness of the analysis of the new set of data that will arrive shortly.