



Geographical
Association

A Pedagogic Application of Multiple Regression Analysis: Precipitation in California

Author(s): P. J. Taylor

Source: *Geography*, July 1980, Vol. 65, No. 3 (July 1980), pp. 203-212

Published by: Geographical Association

Stable URL: <https://www.jstor.org/stable/40569273>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



JSTOR

Geographical Association is collaborating with JSTOR to digitize, preserve and extend access to *Geography*

A Pedagogic Application of Multiple Regression Analysis

Precipitation in California

P. J. Taylor

ABSTRACT. Multiple regression analysis is illustrated in a familiar geographical problem for pedagogic purposes. Precipitation patterns in California are related to altitude, latitude and distance from the coast. This first model indicates a need to incorporate a shadow effect. The model is progressively improved using maps of residuals to the point where the third model is deemed to be highly successful.

Multiple regression analysis is a technique for describing how the variations in one phenomenon are related to variations in two or more other phenomena. It is usually assumed that these other variables do, in some sense, “determine”, “produce” or even “cause” the variations observed in the first variable. This way of thinking about the world has always been important in geography: for instance, areal variations in vegetation have normally been related to areal variation of climatic and soil phenomena. The advent of multiple regression analysis has brought to geography a more precise, numerical approach to the age-old problem of describing the inter-relationships we observe around us.

It is the purpose of this paper to describe multiple regression analysis in a completely non-technical manner. Descriptions of the technique at various degrees of sophistication are readily available in the literature;¹ this discussion is intended for those who wish to understand the rudiments of the technique but do not wish to follow statistical explanations of the various elements of a multiple regression model at this stage. On completing this paper the reader should have acquired the skill to understand the meanings and findings of other geographical articles that apply multiple regression analysis as a tool in substantive research. In order to use the technique in research the reader will have to go to references beyond this paper.

Although this paper does not have a substantive research purpose it is important to specify carefully a “typical” research problem which would be tackled using multiple regression analysis. The problem chosen is a simple one, incorporating well-known relationships. Nothing new will be discovered in the analyses below, but the technique will be illustrated in a situation that is familiar to all readers. An attempt will be made to explain the areal variations in average annual precipitation totals in California.² Thirty meteorological stations have been selected from all parts of the state (Table I and Fig. 1) to provide a data set. The large variations in precipitation recorded in column one of Table I reflect the wide variety of environments within California—literally from deserts to coastal mountain ranges. These

►Dr. P. J. Taylor is Lecturer in Geography at The University of Newcastle upon Tyne.

Table I
VARIABLES USED IN THE FIRST MODEL

Station	Average annual precipitation (inches)	Altitude (feet)	Latitude (degrees)	Distance from coast (miles)
1. Eureka	39.57	43	40.8	1
2. Red Bluff	23.27	341	40.2	97
3. Thermal	18.20	4152	33.8	70
4. Fort Bragg	37.48	74	39.4	1
5. Soda Springs	49.26	6752	39.3	150
6. San Francisco	21.82	52	37.8	5
7. Sacramento	18.07	25	38.5	80
8. San Jose	14.17	95	37.4	28
9. Giant Forest	42.63	6360	36.6	145
10. Salinas	13.85	74	36.7	12
11. Fresno	9.44	331	36.7	114
12. Pt. Piedras	19.33	57	35.7	1
13. Pasa Robles	15.67	740	35.7	31
14. Bakersfield	6.00	489	35.4	75
15. Bishop	5.73	4108	37.3	198
16. Mineral	47.82	4850	40.4	142
17. Santa Barbara	17.95	120	34.4	1
18. Susanville	18.20	4152	40.3	198
19. Tule Lake	10.03	4036	41.9	140
20. Needles	4.63	913	34.8	192
21. Burbank	14.74	699	34.2	47
22. Los Angeles	15.02	312	34.1	16
23. Long Beach	12.36	50	33.8	12
24. Los Banos	8.26	125	37.8	74
25. Blythe	4.05	268	33.6	155
26. San Diego	9.94	19	32.7	5
27. Daggett	4.25	2105	34.09	85
28. Death Valley	1.66	-178	36.5	194
29. Crescent City	74.87	35	41.7	1
30. Colusa	15.95	60	39.2	91

precipitation figures are located on Fig. 1 which may be regarded as the “problem map” (i.e. the pattern of the phenomena to be explained).

The multivariate context

Let us consider what variables should be included in the analysis in order to explain this variation in precipitation. We will start with the orographic component in the precipitation by specifying altitude above sea level in feet as our first explanatory variable. (In statistical terminology this is often referred to as the *independent* variable in contrast to the phenomena we are attempting to explain which is the *dependent* variable.) This new variable is added to Table I as column 2 and it can be seen that generally speaking, precipitation does increase with altitude among the stations. However, the relationship is by no means a simple direct one. For example, compare Bishop (station 15) and Mineral (16). Both are at approximately the same altitude but experience vastly different levels of precipitation. A glance at Fig. 1 will give a good clue as to why this is the case—the stations lie on very different latitudes. Clearly altitude alone does not determine precipitation; we must also consider latitude. Degrees of latitude north is our second independent variable and is shown on column 3 in Table I.

The influence of the new independent variable can be broadly seen by scanning columns 1 and 3 of Table I. The more northerly a station is, the more likely it will fall along or near to westerly-moving storm tracks and so usually it is found to receive more precipitation. Once again, however, this is not always the case. Consider Fort Bragg and Colusa, which are at similar latitudes and yet have very different levels of precipitation. This difference cannot be

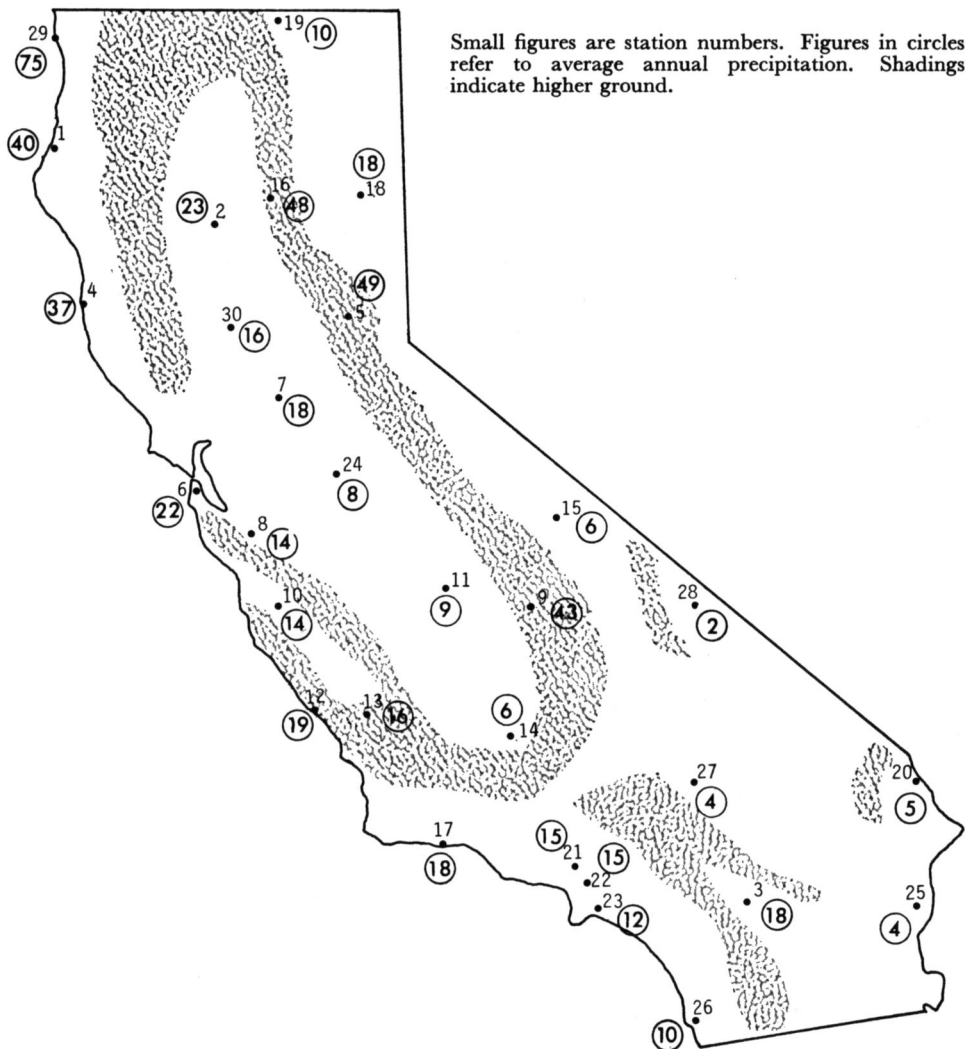


Fig. 1.—Average annual precipitation levels (inches) over thirty stations in California.

put down to altitude, since both stations are at similar heights above sea level. Once again a brief glance at Fig. 1 suggests an additional independent variable to account for this precipitation difference. Whereas Fort Bragg is on the coast, Colusa is nearly 100 miles inland. A third independent variable has therefore been added to Table 1 in column 4, namely distance from the coast in miles. It is expected, and generally borne out in comparing columns 1 and 4, that precipitation declines with distance from the coast as moisture-laden air has progressively less rain to deposit as it moves inland.

The above arguments have a two-sided purpose. First, we have selected in a quite logical manner a set of independent variables which would seem to be pertinent to our goal of explaining variations in California precipitation. We have, in fact, produced three variables describing the situation of each station which act as surrogates for the physical processes which underlie the actual precipitation. Second, we have been able to derive our set of explanatory variables by showing how one often confounds the effects of another. Put simply, to explain precipitation we need to know the altitude *and* latitude *and* distance from the coast

of a station. Knowledge of one of these variables without knowledge of the other two is not particularly helpful. Such a statement is to argue for a *multivariate analysis* in preference to a simple *bi-variate analysis*. The latter consists of pair-wise comparisons of variables as, for instance, in a simple correlation study. Our argument above suggests that such correlation coefficients will be of limited utility and may even be misleading since they cannot unravel the confounding effects of one independent variable on another. If we wish to relate precipitation to altitude, for instance, we cannot adequately do this without taking into account latitude and distance from the coast. A multivariate analysis allows us to do just this; variables are considered not in isolation from one another but jointly operating on precipitation together. The multiple regression technique is a common means for carrying out a multivariate analysis. By using this technique we are setting out our research problem realistically in a multivariate context.

The multiple regression technique

What, then, does the multiple regression technique do? It is that multivariate procedure which is used to relate one variable (known as the dependent) to a set of other variables (known as independent or explanatory). Hence our choice of the technique for the problem stated above, where precipitation becomes the dependent variable and altitude, latitude and distance from the coast are three independent variables. The purpose of the whole exercise is to derive an equation which precisely specifies the relationship between dependent and independent variables. The technique is designed to generate that equation which best describes the pattern of relationships to be found in the data.

The form which the equation takes must be decided upon beforehand. In most analyses, including those described below, a linear form is used. This assumes that changes in the dependent variable relating to changes in the independent variables are constant for all values of all variables. This simply means that if an increase in altitude from 500 to 1000 feet is associated with precipitation rising 10 inches, then similarly an identical increase of altitude from 2000 to 2500 feet will be associated with a 10 inch rise in precipitation. Given this assumption the equation for a multiple regression analysis will be of the form

$$X_1 = \pm a \pm b_2 X_2 \pm b_3 X_3, \dots, \pm b_n X_n \pm e$$

where the X 's refer to a set of n variables (one dependent and $n - 1$ independent) and a and the b 's are the parameters of the equation. The final term, e , includes those aspects of the dependent variable not accounted for by the independent variables. It is the X 's which constitute the data (i.e. Table I in our problem) and the parameters which are estimated by the technique. Hence the basic results of a multiple regression analysis are a set of estimated parameters (a and b 's) which can then be set out as an equation in the form shown above. They are interpreted as follows: (i) a is the base constant and is an estimate of the value of the dependent variable when all the independent variables are zero. In our example it is level of precipitation associated with zero altitude, latitude and distance from the coast: (ii) the b 's are regression coefficients which relate each independent variable to the dependent variable. They are simple gradients and are therefore also in units of the dependent variable. They tell how much change in the dependent variable is associated with a change of *one* unit of an independent variable. In our example the regression coefficient that relates altitude to precipitation is an estimate of how much precipitation increased in inches for an increase of one foot of altitude.

Finally we should mention that these results are only "estimates". Our analysis is based upon a sample of 30 stations so that the results are only estimates of the true parameters (for all locations in California) based upon this partial evidence. The procedures of inferential statistics to cope with this situation are not described here. In fact our "simple" physical problem produces highly significant results even though based upon just thirty observations. We are now in a position to begin assessing these results.

The first model

Our first model of California precipitation is simply precipitation as a function of altitude, latitude and distance from the coast. This is calibrated as "Precipitation at a station equals $(0.004 \times \text{altitude of the station}) + (3.4536 \times \text{latitude of the station}) - (0.1426 \times \text{the distance of the station from the coast}) - 102.5314$ (which is the base level)." If we label our variables X_1 , X_2 , X_3 and X_4 respectively we can write the above formulation much more succinctly as

$$X_1 = 0.004X_2 + 3.4536X_3 - 0.1426X_4 - 102.5314.$$

For any particular station we can predict what the precipitation should be given the altitude, latitude and distance from the coast. At Eureka (1)

$$X_1 = (0.004 \times 43) + (3.4536 \times 40.8) - (0.1426 \times 1) - 102.5314 = 38.4076.$$

This prediction can be compared with the actual precipitation recorded for Eureka which was 39.57. The difference between these two values, the prediction and the actual, is the error or *residual*. In this case the residual is $(39.57 - 38.4076)$ which equals 1.1624. This is interpreted as our equation underpredicting precipitation at Eureka by just over 1 inch. Table II shows all 30 predictions and residuals plus the original values from Table I for comparison. Where the residuals are negative it means that precipitation at that station is over-predicted.

Table II
PREDICTIONS AND RESIDUALS FROM THE FIRST MODEL

Station	Actual precipitation	Predicted precipitation	Residuals
1	39.57	38.4076	1.1624
2	23.27	23.8550	-0.5850
3	18.20	21.1133	-2.9133
4	37.48	33.6988	3.7812
5	49.26	39.2784	9.9816
6	21.82	27.5129	-5.6929
7	18.07	19.1227	-1.0527
8	14.17	23.0258	-8.8558
9	42.63	29.0716	13.5584
10	13.85	22.8050	-8.9550
11	9.44	9.3018	0.1382
12	19.33	20.8513	-1.5213
13	15.67	19.3518	-3.6818
14	6.00	11.0181	-5.0181
15	5.73	14.7640	-9.0340
16	47.82	36.4777	11.3423
17	17.95	16.6180	1.3320
18	18.20	25.3039	-7.1039
19	10.03	38.6306	-28.6006
20	4.63	-6.0172	10.6472
21	14.74	11.7223	3.0177
22	15.02	14.2238	0.7962
23	12.36	12.6919	-0.3319
24	8.26	17.9680	-9.7080
25	4.05	-7.5089	11.5589
26	9.94	9.7653	0.1747
27	4.25	14.4417	-10.1917
28	1.66	-4.8715	6.5315
29	74.87	41.4833	33.3867
30	15.95	20.1136	-4.1636

These errors or residuals indicate that precipitation is not perfectly determined by our three independent variables. We do not have an exact deterministic relationship but instead we have an empirical statistical relationship where only part of the original problem map is explained. The amount that is explained is indicated by the *multiple correlation coefficient* which in the case of the perfect relationship (no errors) would equal unity. On the other hand, if

the multiple correlation coefficient is zero there is no linear relationship whatsoever between X_1 and X_2 , X_3 and X_4 . In this case the coefficient is a relatively high 0.7708. The square of this value, 0.5942, is termed the *coefficient of determination* which tells us how much of the variation in the original problem map is accounted for by our three explanatory variables. This is because this coefficient is measuring the ratio of the variation in predicted values to that of the actual values. In this example the coefficient is informing us that the variation in the values predicted by our equation is nearly 60 per cent of the original variation in the dependent variable. Hence we can state that 59 per cent of the variation of precipitation is accounted for by our three independent variables. Fully 41 per cent still remains (in the residuals) to be explained.

The multiple correlation measure tells us how important our variables are; an alternative statistic is the *standard error of estimate* which tells us how good a predictor our equation is. It is literally a measure (standard deviation) of the spread of the residuals in Table II. It is interpreted as indicating the range of error you can expect when using the equation to predict: approximately two-thirds of predictions will have a residual less than one standard error away from the true value and about 95 per cent will be within two standard errors of the correct answer. For this model the standard error is 11.1825; in other words, we expect two-thirds of our predictions to be within 11 inches of the true level of precipitation. In Table II 25 out of 30 residuals are less than one standard error, which is slightly better than expected. In general we would consider this model a poor predictor, for this standard error is far too large to place much reliance on the predictions. This point is emphasised by three stations (20, 25 and 28) where the model predicts the impossible—negative rainfall.³

The second model

Table III
THE SHADOW EFFECT, PREDICTIONS AND RESIDUALS FROM THE SECOND MODEL

Station	Shadow effect	Actual precipitation	Predicted precipitation	Residuals
1	0	39.57	43.2121	-3.6421
2	1	23.27	20.6225	2.6475
3	1	18.20	7.8397	10.3603
4	0	37.48	38.3934	-0.9134
5	0	49.26	44.6183	4.6417
6	0	21.82	32.5565	-10.7365
7	1	18.07	14.8945	3.1755
8	1	14.17	13.8967	0.2733
9	0	42.63	34.6175	8.0125
10	1	13.85	12.2372	1.6128
11	1	9.44	7.5089	1.9311
12	0	19.33	25.4467	-6.1167
13	1	15.67	9.1893	6.4807
14	1	6.00	5.3289	0.6711
15	1	5.73	13.3347	-7.6047
16	0	47.82	44.8022	3.0178
17	0	17.95	21.0454	-3.0954
18	1	18.20	23.8966	-5.6966
19	1	10.03	32.2326	-22.2026
20	1	4.63	-1.9120	6.5420
21	0	14.74	19.2057	-4.4657
22	0	15.02	19.6331	-4.6131
23	0	12.36	18.2329	-5.8729
24	1	8.26	12.9763	-4.7163
25	1	4.05	-5.5642	9.6142
26	0	9.94	14.6906	-4.7506
27	1	4.25	6.5233	-2.2733
28	1	1.66	1.5827	0.0773
29	0	74.87	46.3354	28.5346
30	1	15.95	16.8426	-0.8926

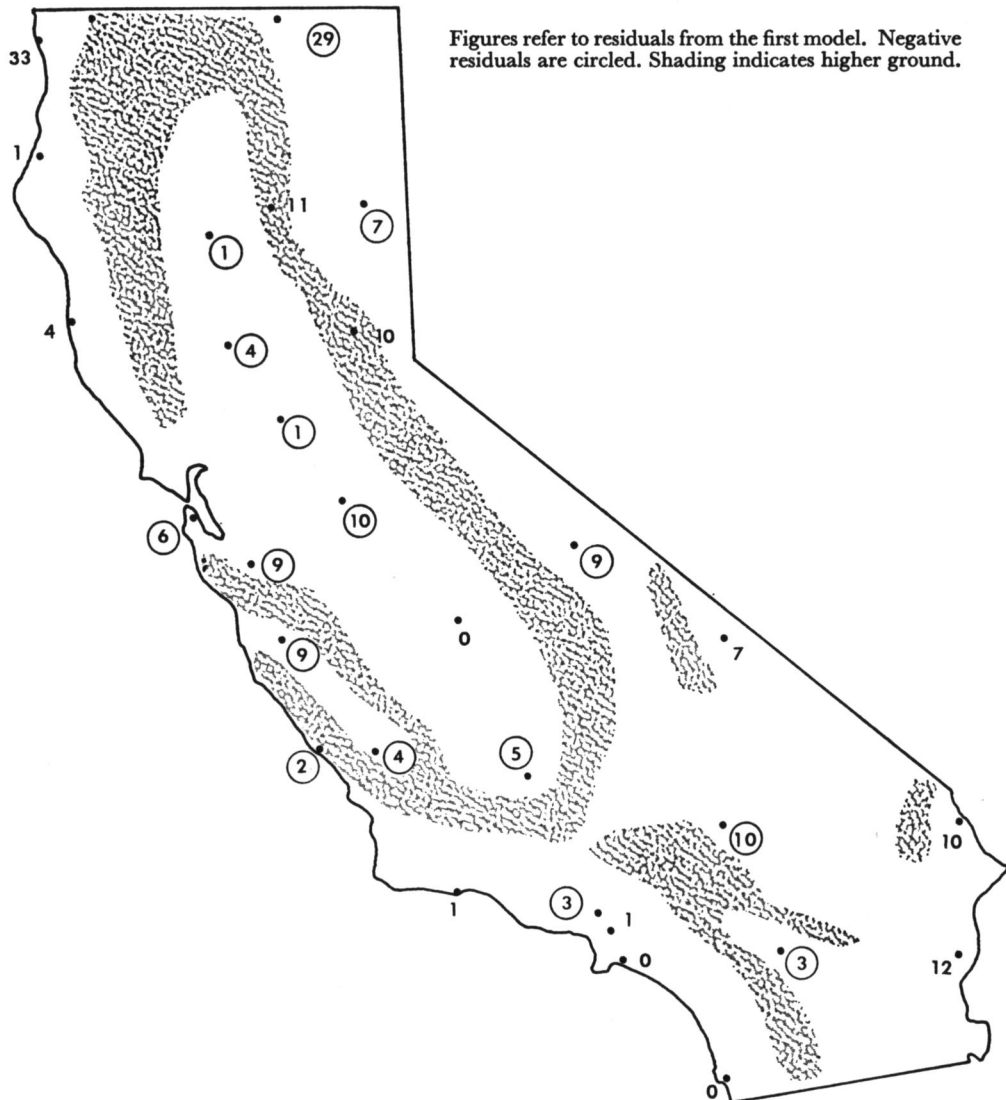


Fig. 2.—Residuals from the first model.

Our first model does indicate that we have made a reasonable start to explaining precipitation variation in California, but it also shows that we have not yet produced a satisfactory model. The first place to look in order to improve our model is at the residuals. These are plotted in Fig. 2. This map shows us where our predictions are good and where they are poor, where we under-predict and where we over-predict. It is from this map of residuals that we hope to find clues as to other possible explanatory variables to add to our analysis. In fact Fig. 2 shows a fairly consistent pattern. On the westward-facing slopes the residuals are invariably positive, while to the leeward side the residuals are typically negative. In other words, to the lee of the mountains our first model over-predicts precipitation, while on the other side of the mountains it under-predicts. This suggests a very clear shadow effect of the mountains, for which California is well known. We can add this to the model by incorporating a further variable which we will term shadow effect. This will be what statisticians refer to as a "dummy variable" taking only the values 0 and 1. All stations in the lee of mountains will score 1, other stations score 0. This new variable is listed in the first column of Table III.

When this is added to our multiple regression analysis as X_5 we obtain the following equation

$$X_1 = 0.0021X_2 + 3.4893X_3 - 0.6518X_4 - 16.1660X_5 - 99.1909.$$

This reads that precipitation at a station will be equal to $(0.0021 \times \text{altitude of the station}) + (3.4893 \times \text{the latitude of the station}) - (0.6518 \times \text{distance of the station from the coast}) - (16.1660 \text{ if the station is in the lee of mountains}) - 99.1909$ (the base level). The addition of our extra variable has been quite successful, because the multiple correlation coefficient has risen to 0.8587 which indicates that we can now account for 0.7374 of the variation in precipitation in the multiple coefficient of determination. Although the level of explanation has risen appreciably the standard error of estimate is still relatively high at 9.1732. The residuals from this new equation are shown in Table III. It is instructive to compare these with our original residuals to see what effect our new variable has had on predictions. Normally the effect is beneficial; for instance adding a shadow effect to San Jose (8) reduces the residual from -8.8558 to $+0.2733$ inches. On some occasions, however, predictions have become worse. San Francisco (6), for example, had a negative residual in the first model (-5.6929) but is not in the lee of mountains, so that the addition of the shadow effect (for which San Francisco scores 0) increases the residual to -10.7365 .

To summarise the second model, we can say that our level of explaining is much better than the first model but that the improvement in predictions is much more slight.

The third model

Let us look at the distribution of residuals from the second model in Fig. 3 to see whether any further explanatory variables suggest themselves. Unlike Fig. 2 there is no clear-cut pattern among this set of residuals. This suggests that the unexplained variation in our precipitation data does not relate to any broad structural variables of the sort dealt with so far. It is more likely that the remaining unexplained variation (the residuals) relate to micro-climatic processes associated with the particular sites of the meteorological stations. It will be difficult to incorporate such details into a broad state-wide modelling exercise such as is attempted here.

There are two stations that do stand out, however, as truly exceptional. In the far north of the state, Crescent City (29) has had by far the largest positive residual and Tule Lake (19) has had by far the largest negative residual for both models developed so far. These residuals are so much larger than all other residuals recorded in this study that we suspect they indicate a different mixture of effects from our explanatory variables. They clearly affect precipitation differently in the far north than elsewhere. This warrants further investigation not necessarily using multiple regression analysis. In the present context, however, we can proceed as follows. We argue that our variables behave differently in the extreme north and so we cannot expect to model that region within the same analysis as the rest of the state. Hence we will omit these two stations and proceed with a third model incorporating just 28 stations. Such a strategy is certainly controversial, for a researcher should not pick and choose what observations to include in this way other than in exceptional circumstances. The residuals for stations 19 and 29 are exceptional. (A useful next stage would be to collect data for Washington and Oregon and see whether these two stations fit consistently into a new "north-west region" multiple regression equation.)

Our third model is highly successful. The multiple correlation coefficient is very high at 0.9392, which means that we now account for 88 per cent of the variation of precipitation (coefficient of determination is 0.8821). More significant is the new, relatively low, standard error of estimate of only 4.9661. This dramatic improvement indicates that the poor predictive abilities of our previous models were largely due to the two northern stations, as in fact the residuals indicated. Our new final equation is

$$X_1 = 0.0033X_2 + 3.0655X_3 - 0.0546X_4 - 11.5771X_5 - 87.7331$$

which by now can be left to the reader to translate into English.

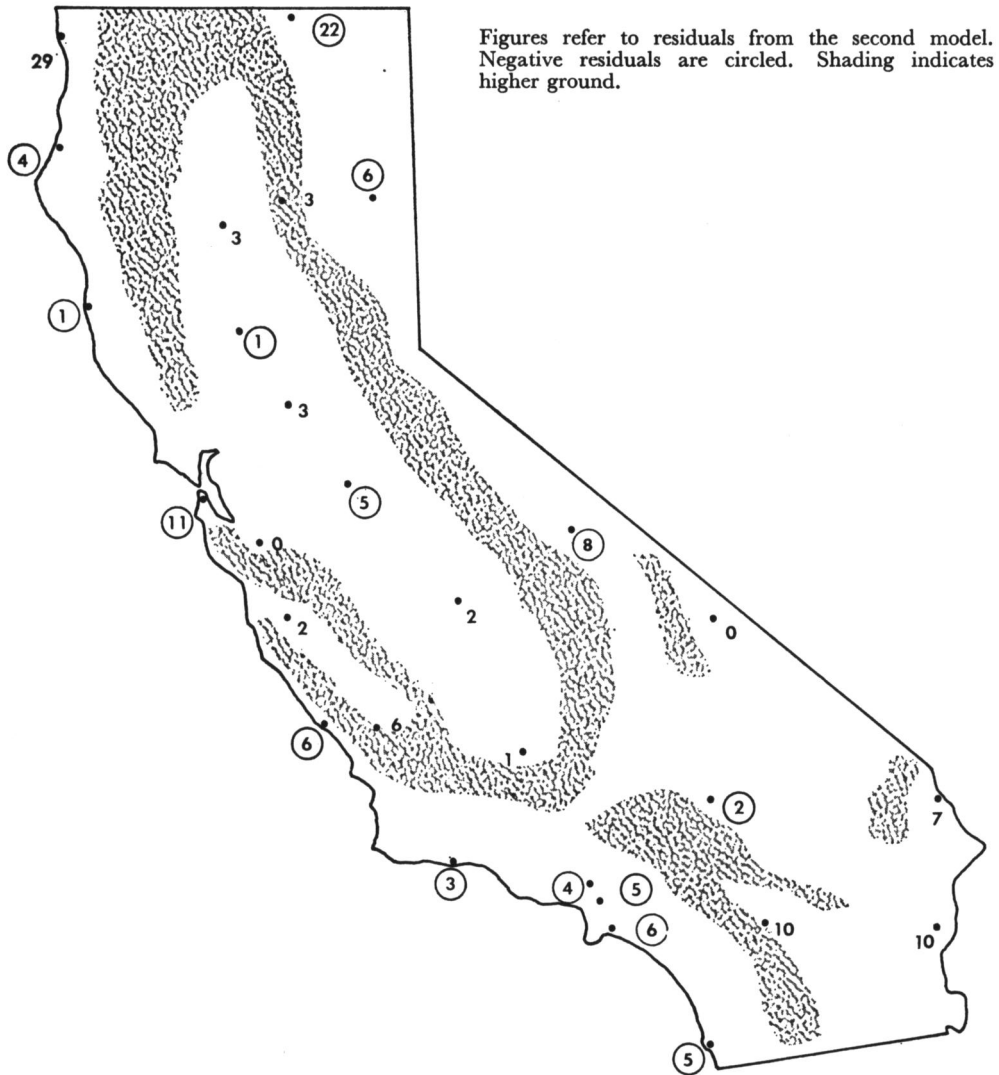


Fig. 3.—Residuals from the second model.

Conclusions

Our models have become increasingly better as we have proceeded through the analyses. This is because we have used the method of residuals which cumulatively attempts to incorporate part of the unexplained variation in each succeeding model. In this case the strategy has been very successful, leaving only 12 per cent of the variation unaccounted for in the final model. All the results are summarised in Table IV.

The first three rows of Table IV show the cumulative improvement of our models. The second three rows show the changing magnitude of the regression coefficients which are interesting in the way they reflect the modelling sequence. The addition of the shadow effect variable in the second model reduces by about one half the regression coefficients for altitude and distance from the coast. This means that in the first model the shadow effect, although not explicitly modelled, was implicitly confounded with the altitude and distance variables.

Table IV
SUMMARY OF RESULTS

	First model	Second model	Third model
Multiple correlation	0.7708	0.8587	0.9392
Multiple determination	0.5942	0.7374	0.8821
Standard error of estimate	11.1825	9.1732	4.9661
Regression coefficient for altitude	0.0041	0.0021	0.0033
Regression coefficient for latitude	3.4536	3.4893	3.0655
Regression coefficient for distance from coast	-0.1426	-0.0518	-0.0546
Regression coefficient for shadow effect	—	-16.1660	-11.5771
Base constant	-102.5314	-99.1909	-87.7331

As we might expect it was not confounded with latitude. In the second model this confounding is solved by allowing the shadow effect to have its own separate regression coefficient. Latitude, on the other hand, is affected by omitting the two northern stations in creating the third model. It is the regression coefficients in this third model which produce the high level of explanation and relatively accurate predictions.

The discussion presented above has not added anything substantively new to our knowledge of California climates. It has illustrated the multiple regression model, however, in a situation relatively well known to most geographers. By observing use of the technique to replicate the familiar, the reader should now be able to understand the technique when it is used to unravel the less well known.

One final point is worthy of note. The above presentation has concentrated upon the explanatory role of multiple regression (i.e. explaining *known* precipitation levels). Suppose that we wish to know the level of precipitation at an ungauged site to see whether there is sufficient precipitation to build a dam. In this situation we could use our regression equation in an applied predictive role (i.e. estimating *unknown* precipitation). This would merely require the input of the values of the independent variables for the new site. This example gives a brief glimpse of the potential application of what has become a standard statistical technique in modern geography.

REFERENCES

¹ See, for example, P. J. Taylor, *Quantitative Methods in Geography*, Houghton Mifflin, Boston, 1976.

R. J. Johnston, *Multivariate Statistical Analysis in Geography*, Longmans, London, 1978.

R. Ferguson, "Linear regression in geography", *CATMOG*, 15, Geo-Abstracts, Norwich, 1978.

² The data for the analyses are from E. L. Felton, *California's Many Climates*, Pacific Books, Palo Alto, 1965.

³ We can "force" the model to predict only positive precipitation values by using the logarithm of precipitation rather than the raw precipitation totals in our model. Logarithms, and therefore predictions of logarithms, can only be positive. This strategy is not employed here but could be considered a logical next step.