



RAPID ANOMALY RECOGNITION AND RANKING
FOR MULTI-ELEMENT REGIONAL STREAM SEDIMENT SURVEYS

By P. Matysek, A. J. Sinclair, and W. K. Fletcher
Department of Geological Sciences
University of British Columbia

INTRODUCTION

As part of our continuing study of rapid, thorough evaluation procedures for multi-element stream sediment data (for example, Sinclair and Fletcher, 1979; Matysek, et al., 1980), we have developed a systematic, computer-oriented method of recognizing and ranking anomalous samples. Our detailed procedure utilizes the type and quality of data incorporated in various regional programs undertaken by the British Columbia Ministry of Energy, Mines and Petroleum Resources but can be adapted easily for data for other programs.

Regional multi-element stream sediment surveys of the type carried out in British Columbia under terms of the Uranium Reconnaissance Program contain coded information on the principal rock unit forming the provenance region of each sample. Consequently, the following procedure for determining multi-element background models is intended to be applied to sample subsets based on provenance (rock type). Rock-type coding for this purpose is never perfect: some basins may be underlain by two or more important rock types, other drainage basins may be miscoded, perhaps because of the scale of geological base maps available. In any case it is apparent that some apparently anomalous metal concentrations arise from incorrect assignment of the dominant rock type or from mixing of sediments derived from several rock types.

GENERAL METHODOLOGY

Our general approach to recognition and ranking of anomalous samples is summarized on Figure 1. In brief the method involves the following steps:

- (1) Sorting of data into provenance groups, that is, predominant rock type in drainage basin above the sample.
- (2) Evaluation of simple statistics and probability graphs for each element in each provenance group.
- (3) Threshold selection using the method of Sinclair (1976) to isolate anomalous samples from background samples.
- (4) Selection of one or more elements to serve as the focus of the study (for example, zinc).

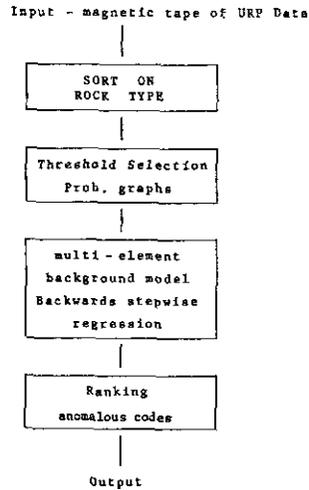


Figure 1. Sequential approach to anomaly recognition and ranking.

- (5) Backward stepwise regression of each provenance group to develop background models for zinc in terms of other elements.
- (6) Ranking individual samples in terms of (a) their contamination code and (b) the regression model and threshold.
- (7) Output of sample information in a manner convenient for practical use in follow-up examination.

SORTING INTO PROVENANCE GROUPS

Data for each provenance group should be dealt with separately. Means and standard deviations of all raw and log-transformed metal abundances provide insight into levels of abundance, dispersion, and general aspect of population densities (histogram). Correlation coefficients indicate metal associations of geological importance (for example, Sinclair and Tessari, 1980). If only background values are considered, these associations commonly reflect differences in background environments and are not related directly to anomalous samples.

THRESHOLD SELECTION

Separation of background and anomalous samples is essential to our method because it leads directly to statistical models for background metal abundances. Consequently, the method of threshold recognition is important. We have adopted the probability graph approach of Sinclair (1976) because this procedure is systematic and has been shown by numerous examples to provide effective thresholds for many types of geochemical data.

ELEMENT SELECTION

We must decide which element or elements are of direct concern to our search problem. Are we interested in silver-lead-zinc, copper-molybdenum, tungsten-uranium, or others? Of course, we may want to investigate many associations of the sort listed, but in our approach each association would be dealt with separately. Within a particular metal association it may not be necessary to deal thoroughly with all elements because some may be redundant, others may not show adequate geochemical contrast, and still others may present limitations resulting from analytical problems. In our case we will use zinc data as a basis for evaluating regional silt samples in terms of silver-lead-zinc and lead-zinc associations typical of our study area (map-area 82F).

MULTIVARIATE MODELLING OF BACKGROUND VALUES

Multiple regression has been shown by many to be an effective method of demonstrating empirical relationships between a particular element (dependent variable) and a group of other elements (independent variables). In many cases a high proportion of the variability of the dependent variable is explained in terms of variations in the independent variables (Sinclair and Fletcher, 1980). Where such methods are applied to background samples only, the abundance of a dependent variable (for example, zinc) can be expressed as a linear combination of the abundances (or logarithms of abundances) of many other elements to provide a multivariate background model.

We have experimented with two approaches to the selection of samples used to establish a multiple regression model. In our first attempts sample selection was based on the dependent variable for a single provenance group with only those values below the threshold (based on probability graphs) being selected. In a later refinement we edited the data base for a single provenance group by omitting samples that were also obviously anomalous with respect to any of the independent variables.

The specific method we use for multivariate background modelling is backward, stepwise regression which starts with all independent variables in the data base and sequentially drops those that make no statistically significant contribution to explaining the variability of the dependent variable. Eventually a point is reached where all remaining variables are statistically significant (at the 0.05 level, for example) and an equation is obtained of the form

$$\text{Log (Zn)} = \beta_0 + \beta_3 \log (X_3) + \beta_4 \log (X_4) + \beta_9 \log (X_9) \text{ etc.}$$

where β 's are constant and X_i 's are abundances of metal i .

RECOGNITION AND RANKING OF ANOMALOUS SAMPLES

For each sample we determine a series of ranks from 0 to 3 by comparing the observed value of the dependent variable with the values calculated

by each of the provenance group multivariate models. Significance of the rank numbers is shown on Figure 2. We then calculate a 4-digit ranking code for each sample where the first digit is the number of rock types for which rank 3 was obtained, the second digit is the number of rock types for which rank 2 was obtained, and so on. If there are seven rock types all with very high zinc values (rank 3) the ranking code would be 7000; in another case rank might be (3) for two rock types, (2) for three rock types, (1) for two rock types, and (0) for one rock type to give a ranking code of 2321.

The main advantage of this procedure is as a refinement in the selection of anomalous values relating to the probability graph procedure and the assigning of relative priorities to anomalous samples. Values above t_1 (Figure 2) are recognized as being anomalous without the aid of multiple

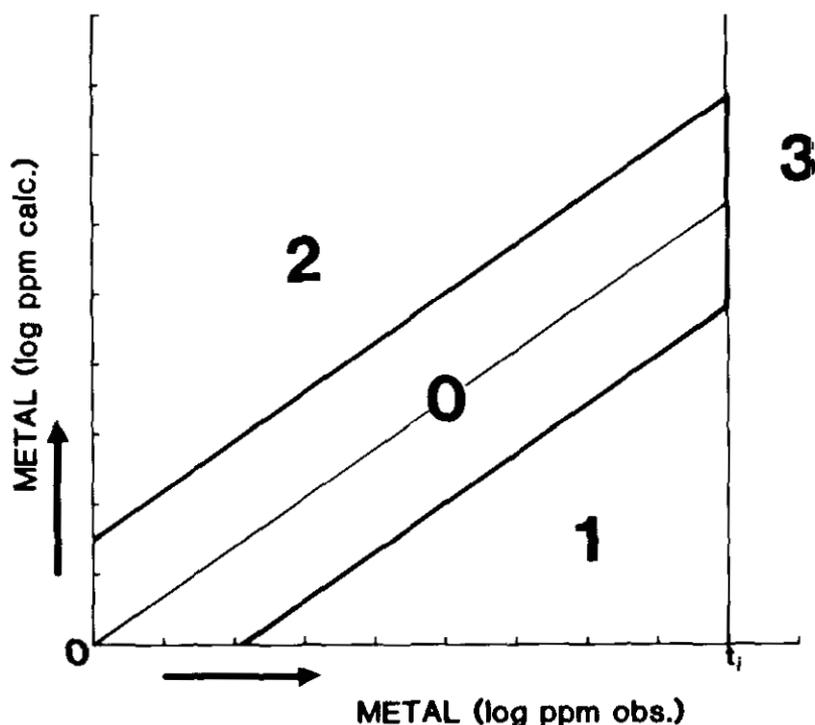


Figure 2. Sample ranking in relation to fields on a plot of observed value versus a value calculated from a multivariate model.

regression. In addition, however, values below t_1 that depart substantially from the expectation according to a multiple regression model (1 and 2 on Figure 2) are also out of the ordinary and warrant examination. In particular, we are interested in those values below t_1 that are much higher than the corresponding calculated values. Such samples are anomalous in one element, relative to a linear combination of other elements. On Figure 2 the suggestion is made graphically that samples are anomalous if observed values are more than two standard errors greater than values calculated according to the multiple regression model.

OUTPUT PROCEDURES

We have designed an output system by which samples can be ordered in terms of decreasing priority for follow-up exploration. All anomalous samples recognized by the foregoing procedures are ranked according to the estimated likelihood of sample contamination from such factors as known mines, man-made metallic features, or fertilizer, on a scale of 0 to 3. Our first rank of anomalous samples is based on this coded parameter, zero contamination being of most interest. Within this group we code a sample for each background model as 3, 2, 1, or 0 as described previously and a 4-digit ranking code is used to list samples within each contamination group in order of decreasing ranking code. Locations for each sample are listed as is the observed abundance of the dependent variable and the sample number. These items are arranged in such a manner as to promote efficiency of evaluation of each sample. In addition, we use plot locations of anomalous samples with their identification number and ranking code.

CASE HISTORY (MAP-AREA 82F)

Multi-element data are available for sample sites in map-area 82F at an approximate sample density of one sample per 12 square kilometres. Samples are analysed for zinc, lead, nickel, cobalt, manganese, copper, mercury, tungsten, and molybdenum. Samples were grouped initially on the basis of coding as to dominant rock type in the provenance region. Data for each element in each provenance group were examined as a probability graph and a threshold selected separating two populations (presumably anomalous and background) using the method of Sinclair (1976). We chose to examine zinc as the dependent variable described here because of the association silver-lead-zinc in known vein deposits in the area. Background multivariate models for zinc in terms of other elements were obtained for each of the seven provenance groups for which we have adequate samples. Three of these models are summarized in Table 1 to illustrate the type of results obtained. Statistics for all seven provenance group models for zinc are given in Table 2 to illustrate the statistical quality of the background models.

All samples coded in one of the seven provenance groups for which we could calculate background models were treated by each of the background equations separately. The calculated zinc background according to a given model was then compared with expectations for that model so that for each background model a sample received a ranking from 0 to 3 inclusive (compare Figure 2). In our case each sample was ranked seven times, once for each model. These rankings were accumulated into a single ranking code. Samples recognized as anomalous or potentially anomalous were divided into three contamination classes with priority decreasing as certainty of contamination increases. For each contamination category samples are ranked according to decreasing numeric value of the ranking code. An example is shown in Table 3, where a small part of the 0-contamination category is listed.

GRANITE

$$\begin{aligned} \log (\text{Zn}) &= 0.4726 + 0.0713 \log (\text{Cu}) + 0.2420 \log (\text{Pb}) + 0.0529 \log (\text{Ni}) \\ &\quad + 0.3994 \log (\text{Mn}) + 0.4189 \log (\text{Fe}) - 0.2334 \log (\text{Co}) \\ R^2 &= 0.62 \\ S_e &= 0.1277 \\ n &= 393 \end{aligned}$$

QUARTZITE

$$\begin{aligned} \log (\text{Zn}) &= 1.1020 + 0.2721 \log (\text{Pb}) + 0.1316 \log (\text{Ni}) + 0.4891 \log (\text{Fe}) \\ &\quad + 0.1412 \log (\text{Mo}) + 0.0399 \log (\text{Hg}) \\ R^2 &= 0.74 \\ S_e &= 0.0915 \\ n &= 287 \end{aligned}$$

SCHIST

$$\begin{aligned} \log (\text{Zn}) &= 0.8392 + 0.4100 \log (\text{Pb}) + 0.2244 \log (\text{Ni}) + 0.5603 \log (\text{Fe}) \\ &\quad + 0.2412 \log (\text{W}) \\ R^2 &= 0.76 \\ S_e &= 0.1008 \\ n &= 27 \end{aligned}$$

TABLE 1. EXAMPLES OF MULTIVARIATE REGRESSION BACKGROUND MODELS FOR ZINC MAP-AREA 82F

	PROVENANCE GROUP						
	GRNT	QRTZ	SLTE	ANDS	ARGL	GNSS	SCST
n	393	287	100	57	56	53	27
R	.79	.86	.84	.84	.92	.85	.87
R ²	.62	.74	.70	.70	.85	.71	.76
S _e	.1277	.0915	.1184	.1031	.0812	.0940	.1008

TABLE 2. SUMMARY STATISTICS FOR MULTIVARIATE BACKGROUND ZINC MODELS SEVEN PROVENANCE GROUPS MAP-AREA 82F

	UTME	UTMN	GNSS	QRTZ	GRNT	SLTE	ANDS	SCST	ARGL	TOTALS (3210)	ZN (PPM)	ROCK TYPE	CONTAM	URPS (82F) ID #
27	536212	5457758	3	3	3	3	3	3	3	7000	354	QRTZ	0	773128
28	507056	5518882	3	3	3	3	3	3	3	7000	359	SCST	0	777314
29	476055	5450823	3	3	3	3	3	3	3	7000	345	ANDS	0	771076
30	482931	5461513	3	3	3	3	3	3	3	7000	314	ANDS	0	779126
31	501399	5526605	3	3	3	3	3	3	3	7000	309	SLTE	0	777274
32	489602	5443727	3	3	3	3	3	3	3	7000	304	SLTE	0	773056
33	484807	5528468	3	3	3	3	3	3	3	7000	295	GRNT	0	771198
34	487496	5456216	3	3	3	3	3	3	3	7000	295	SLTE	0	773086
35	476166	5487364	3	3	3	3	3	3	3	7000	299	GRNT	0	777198
36	480507	5487605	3	3	3	3	3	3	3	7000	299	GRNT	0	777236
37	476597	5458297	3	3	3	3	3	3	3	7000	274	ANDS	0	771103
38	475075	5488439	3	3	3	3	3	3	3	7000	279	GRNT	0	777302
39	502653	5508315	3	3	3	3	3	3	3	7000	279	SLTE	0	777259
40	479646	5455841	3	3	3	3	3	3	3	7000	264	ARGL	0	771100
41	494963	5528152	3	3	3	3	3	3	3	7000	255	SLTE	0	777306
42	474792	5452528	3	3	3	3	3	3	3	7000	244	ANDS	0	771077
43	489916	5458581	3	3	3	3	3	3	3	7000	239	SLTE	0	775116
44	509423	5500800	3	3	3	3	3	3	2	6100	224	SCST	0	773237
45	524924	5447471	3	3	3	3	3	3	2	6100	214	QRTZ	0	775100
46	448185	5533845	3	3	3	3	3	3	2	6100	214	SCST	0	779163
47	462164	5538222	3	3	3	3	3	3	2	6100	209	GRNT	0	775328
48	487885	5451555	3	3	3	3	3	3	2	6100	191	GRNT	0	777112
49	489443	5499269	3	3	3	3	3	3	2	6100	198	GRNT	0	777212
50	549914	5512200	3	3	3	3	3	3	0	6001	191	QRTZ	0	777182
51	482276	5463771	3	3	3	2	2	3	2	4300	186	ARGL	0	771084
52	484487	5484729	3	3	3	2	2	3	2	4300	187	GRNT	0	777203
53	441779	5442706	3	3	3	0	0	3	0	4003	186	GNSS	0	775015

TABLE 3. PART OF A TABLE LISTING ANOMALOUS SAMPLES IN ORDER OF DECREASED RANKING CODE*

From a total of 1 259 samples, this procedure produced 115 anomalous samples in the 0-contamination category. Our procedure is to list these samples in tabular form in Table 3 and to produce computer-drawn plots of anomalous sample locations as illustrated on Figure 3.

In addition to ranking information, original raw data, and coordinates, the output table contains a simple consecutive numeric identifier used for clarity on the map output and permitting easy combined use of the tabulated data and the output map. The output map is of particular use because it identifies the most obvious anomalous samples (for example, 7000) from those that might escape detection (for example, 0520). The scale of the location plot should be identical to geological base maps of the area so the two can be studied together without ambiguity. We tested sensitivity of the regression procedure for determining a multivariate background for zinc by establishing such models based on two training

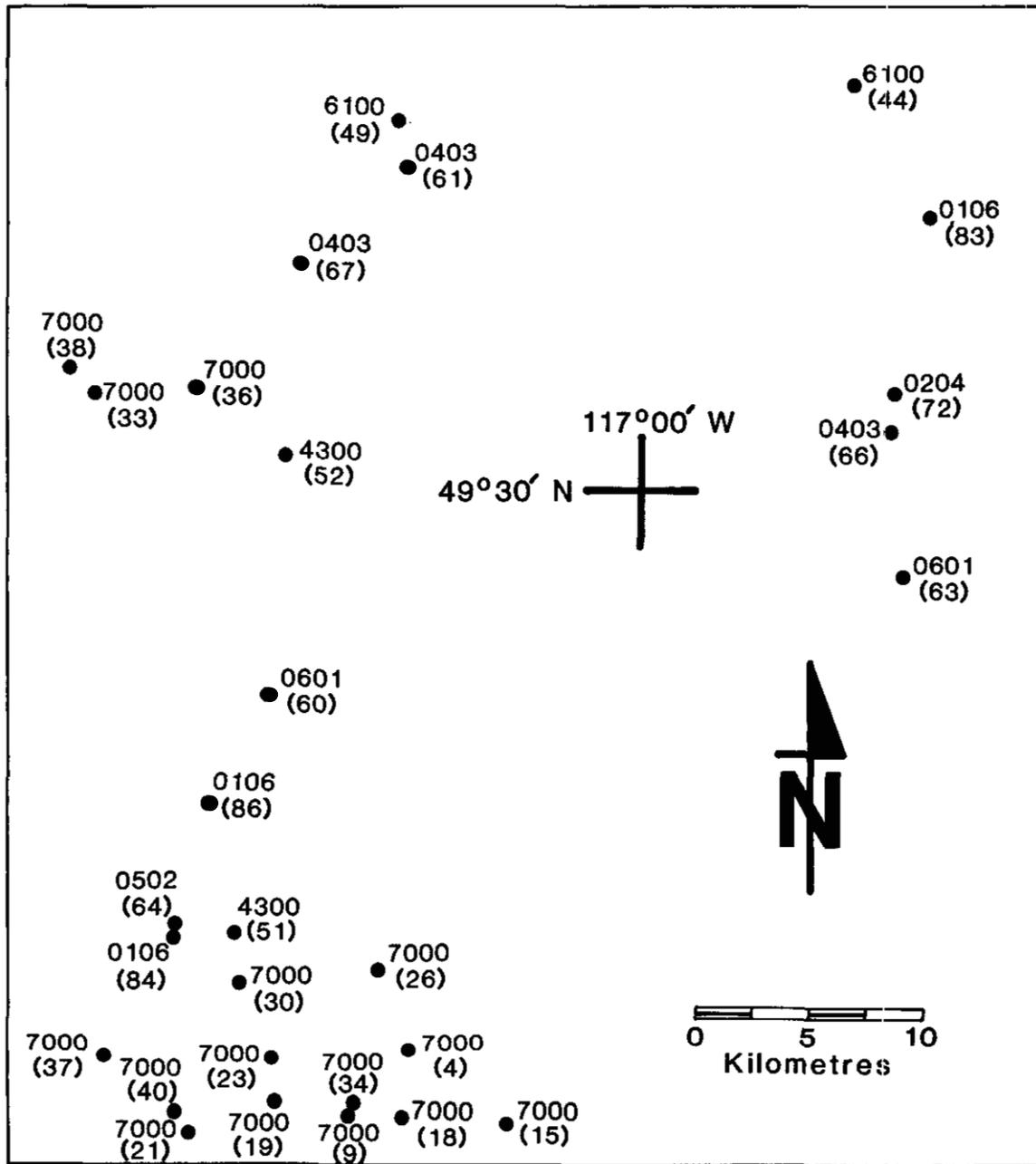


Figure 3. Plot of an area of anomalous samples redrafted from computer output.

sets: (1) all samples indicated as having background zinc values, and (2) the same data set minus any samples that appeared to be anomalous in any element other than zinc. Tables 1 and 2 are based entirely on the second training set. Figures 4 and 5 illustrate the contrasting results obtained in background definition. It is clear that the 'cleaner' data set (number 2 previously) leads to a better multiple regression relationship, that is, with less scatter of calculated and observed values. The problem with using the second training set is that more work is required to set it up and more samples will be included in the anomalous category.

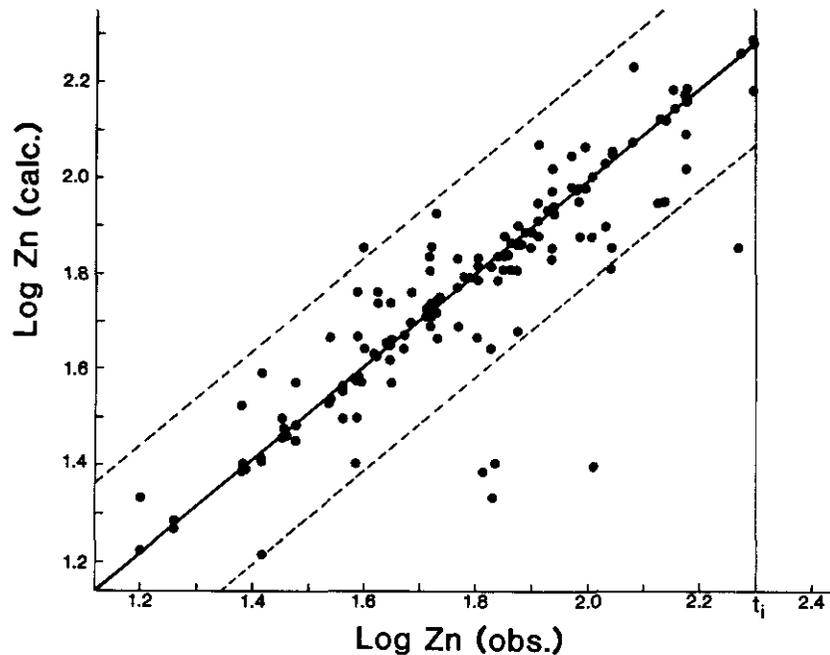


Figure 4. Observed versus calculated zinc values for provenance group, 'ARGL,' map-area 82F; calculated values based on a model determined from all samples with background zinc values.

DISCUSSION

The methodology described here would appear to have a wide range of applications to geochemical data evaluation, perhaps with minor modifications to suit particular data sets. For example, many geochemical surveys may not record the likelihood that a sample is contaminated, and this level of ranking might have to be omitted. The precise limits to the coding regions illustrated on Figure 2 can be changed to suit a particular bias to anomaly selection, resulting in a slightly different listing of anomalous samples.

One of the serious problems is the question of initial grouping of data on the basis of dominant rock type that underlies the drainage basin of each sample, a classification which is fundamental to our procedure. A substantial amount of effort is required to code this rock-type information even if the data are available. If rock type has not been coded it may be necessary to use some less satisfactory method of grouping data, such as the use of factor analysis to provide an approximation of background geology for each sample. In some environments, of course, some other parameter may be more useful than rock type for grouping data.

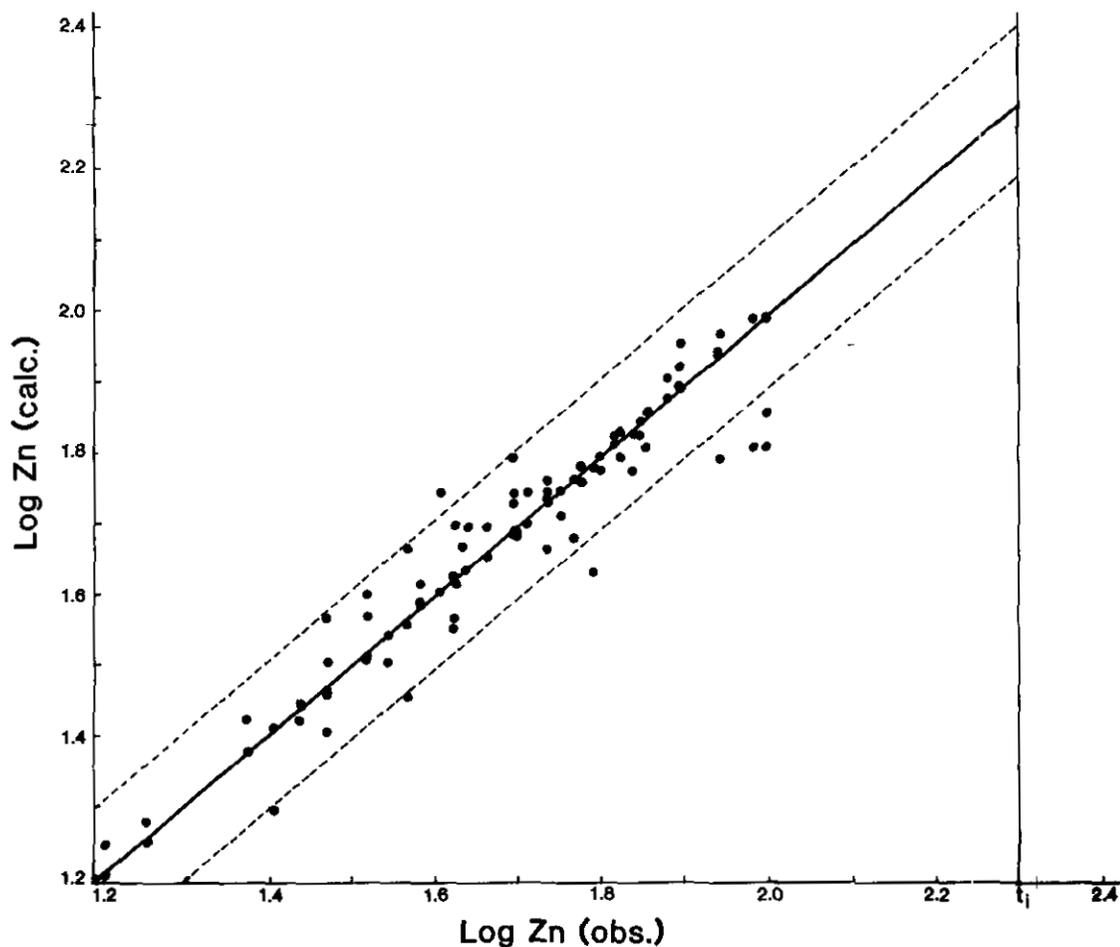


Figure 5. Observed versus calculated zinc values for provenance group, 'ARGL,' map-area 82F; calculated values based on a model determined from those samples with zinc background values that also are not anomalous in any other element (that is, a 'cleaner' subset of the data used for Figure 4).

CONCLUSIONS

A method of anomaly selection and ranking for multi-element regional stream sediment data has been described. The procedure offers the following advantages:

- (1) The method is rigorous in making use of established statistical methods for treating geochemical data such as a probability graph analysis and backward stepwise regression.
- (2) The procedure is computer based and is rapid and thorough.
- (3) The methodology ensures that some anomalous values which are not obvious (that is, are not higher than a simple threshold) will be recognized.

- (4) A novel ranking procedure is described that assigns relative priorities to samples for further investigation. Details of the ranking procedure are subjective but a system of ranking codes clearly describes the manner in which a sample is anomalous.
- (5) Because samples are tested against every rock type, the procedure incorporates an evaluation as to whether other rock types might be contributing to the provenance area of a particular sample. Possible additional rock types are identified and can be compared with available geological maps.

ACKNOWLEDGMENTS

This project has been funded by the British Columbia Ministry of Energy, Mines and Petroleum Resources.

REFERENCES

- Matysek, P., Fletcher, W.K., Sinclair, A.J., and Bentzen, A. (1980): A Preliminary Evaluation of Categorical Field Observations for Regional Stream Sediment Samples, B.C. Ministry of Energy, Mines & Pet. Res., Geological Fieldwork, 1980, Paper 1981-1, pp. 148-158.
- Sinclair, A.J. (1976): Applications of Probability Graphs in Mineral Exploration, Assoc. Explor. Geochem., Special Vol. No. 4, 95 pp.
- Sinclair, A.J. and Fletcher, W.K. (1979): Evaluation Procedure for Geochemical Data, Uranium Reconnaissance Program, B.C. Ministry of Energy, Mines & Pet. Res., Geological Fieldwork, 1979, Paper 1980-1, pp. 130-141.
- Sinclair, A.J. and Tessari, O.J. (1981): Vein Geochemistry, An Exploration Tool in Keno Hill Camp, Yukon Territory, Canada, Jour. Geochem. Expl., Vol. 14, pp. 1-14.