14th Congress of International Society for Photogrammetry

Hamburg, 1980

Commission VII

Presented Paper

Balter B.M., Egorov V.V.

Space Research Institute, Moscow, USSR

REMOTE SENSING OF OBJECT'S STATE AS A STATISTICAL ESTIMATION PROBLEM: GENERAL CONSIDERATIONS AND ALGORITHM

ABSTRACT: We argue that remote sensing of object's state is a promising field in itself, and that it could enhance the classification accuracy.We show that point-by-point methods fail,and so the problem is put as a statistical one of estimating the mean values (that is, the first moments) of state variables over an area.

The method is a modification of the maximum likelihood estimation procedure.It needs extensive information on the relation between the state and the brightness,which is derived by nearly the same method from test area observations. Once fully determined, this relation is valid everywhere.We are able to a priori and a posteriori assess the sensitivity of the method to data shortage, noise, and violation of some assumptions.Mathematically, the moments of a random variable are estimated from the observations of the summatory statistics of a non-linear function in that variable.It is a rather general problem.

Introduction

Remote sensing of a known object's state, e.g. of crop maturity or soil humidity, has been receiving little attention (particularly, in automatic data processing) as compared to remote sensing of object's cathegory, i.e. to classification.The state of objects is, however, the most promising field for remote sensing because of the need for this information and the difficulties of collecting in situ the large amounts of it in a timely and precise manner.

The obvious explanation is that even the object's type can be deduced from remotely sensed data with only a marginal accuracy so that the much more complicated problem of assessing the object's state by classification must seem altogether

hopeless. Some experience in the field has only supported this belief /1/. We deduce from this not that the problem should not be tackled, but that the method should be statistical estimation rather than classification. It is only natural since the problem is of a continuous rather than of a discrete nature.

We show that if the object's type is known and some rather common statistical assumptions are fulfilled, putting the problem as one of estimation immediately gives us a package of ready-to-use powerful statistical methods better developed and aestetically more pleasing than classification. We carry out in this paper what adaptation of classical estimation methods to our problem is necessary. In the end, we try to show that classification accuracy is limited mainly by not taking into account the fluctuations of the object's state. So combined classification and state estimation could enhance the precission of the former.

A number of more complicated mathematical points has been omitted. We intend to discuss them in a separate strictly mathematical paper. We feel that all that reasoning would be a bit out of place here. In this paper, all mathematics has been removed to Appendix.

1. Problem statement

In remote sensing, each resolution element (RE) usually consists of many ($n \gg 1$) identical microelements, as grass consists of individual blades with a portion of soil for each. The observed brightness Y_j ($j=1,\dots,N$) of the j-th RE is, then, the mean of the microelements' brightnesses Y_{ij}: $Y_j=\frac{1}{n}\sum_{i=1}^{n} Y_{ij}$, as most sensors do average just brightness RE's state vector $X_j=\frac{1}{n}\sum_{i=1}^{n} x_{ij}$, x_{ij} being state vectors of microelements. For grass, the components of x are projective coverage, chlorophyll and water content and so on. The size of a microelement and the number of state variables are supposed to be chosen in such a way that fully (i.e. within the admissible range of errors) determines Y_{ij}: $Y_{ij}=f(x_{ij})$. Then, the function f is the same for a given object regardless of all circumstances. In what follows, it is assumed known. For the method to obtain it, see chapter 2.

We suppose that the area H under study consists of N REs and is a statistically stationary one so that x's, though fluctuating, are all from the same probability distribution P(x). This implies that there is only one object in H. So one should first classify the frame and either pick out the safest parts of H by use of known non-parametric methods to discover non-stationarity, or use the simultaneous classification-state determination, the idea of which is given is chapter 4, to enhance classification accuracy.

Finally, we assume the independence of all x's, though close points are surely correlated. It is a common thing in statistics since the most powerful methods are derived under this assumption but work well even if it is violated. Besides, we could generalize our method by adopting the mixed autoreg-

ression-moving average model of the form $\sum_{i=K}^{K+m} \alpha_{i-K} x_i = \sum_{i=K}^{K+m} \beta_{i-K} \varepsilon_i$, where ε's are independent, so that our sums would remain sums. But this is yet to be done.

Now consider the set of all X_j's and Y_j's in an area. As shown in Appendix, not all points (X_j, Y_j) lie on the regression line $M(Y/X)$: $D(Y/X) \neq 0$, so that, in contrast to microelements (unless they are chosen improperly: are too large or have too few state variables), neither X determines Y uniquely, nor an observation Y_0 of a RE determines that RE's state X, but a whole set (X_1, X_2) (see fig.1). So point-by-point methods dont's work. Their error is not even averaged out when taking into account many REs since there is no way to choose for each RE the right-on-the-average X_j lying on the unknown regression line. So we dismiss the point-by-point ideology. Its intrinsic error is assessed in Appendix. The a priori formulae are for computing at home when P(x) (i.e., its moments μ) are prescribed. The a posteriori formulae should instead be supplied by observations Y_j only. Note that, though $D(X/Y) \sim 1/n$ and vanishes at large n, general dispersions D(X) and D(Y) are likewise $\sim 1/n$ so that point-by-point methods are no better at large n, only the data cluster tightens since the fluctuations vanish.

But the fluctuations are necessary to secure the uniqueness of solution. Even if $D(Y/X)=0$ and $Y=F(X)$ - the best case for point-by-point methods, - a F with an extremum will give rize to two solutions X_1, X_2 for an observation Y. Likewise, if there are more state variables affecting the brightness than independent observation channels, we'll have for each RE more unknowns that equations. No consistent-on-the-average method could be provided since the system of equations for all X_j's would either be no better than an equation for a single RE (if X_j's are let to be different), or consist of the equations with the different left sides, i.e. observations, and the same right sides containing X (if all X's are assumed to be equal) and be unresolvable algebraically.

However, in both cases mentioned, the same P(X) placed in two otherwise indistinguishable points X_1, X_2 will produce the different distributions of observations P'Y), and so the mean values X_1 and X_2 could be distinguished. See, e.g., fig. 2: the distributions A_1, A_2, A_3 have different forms. In other words, now we have as many equations as the observable moments of P(Y), from which we may determine as many moments of P(X). (As there is no such function as F on fig. 2, but only a statistical relation, this reasoning should be slightly modified).

So we propose to limit our demands to only mean values of the state vector over H. We believe it is all one practically needs. Mathematically, a set of variables $Y_j = \frac{1}{n} \sum_{i=1}^{n} f(X_{ij})$ $(j=1,\ldots, N)$ is observed. The moment μ_1 (and, possibly, $\mu_2, \ldots$) of P(X) should be estimated. Now we'all expose the mathematical pitfalls of the problem and provide a method to solve it.

2. State estimation algorithm and obtaining a priori information

We'll use the maximum likelihood (ML) method of parameter estimation as it is suitable for analytical examination and is the most powerful one. ML equations are solved by some widely known numerical procedures /2/. The number of equations is limited by noise. If P(Y) is represented as the Edgeworth series (see Appendix), this number M is the number of terms above the noise level. Since f is non-linear, every term depends on all of the unknown parameters μ (and there is an infinite number of them). So unless the sequence of μs is in some way cut off, there will be always more unknowns that equations and the usual ML will be inapplicable.

It's one of the deepest questions of mathematical statistics: the estimation of some parameters in the presence of many other nuisance parameters, all of which it's impossible to estimate. There are some methods based on ML for this problem, but they need specific P(Y), and the only hope for them is the exponential approximation of P(Y) dealt with in chapter 4.

We propose instead to change the classical statement of ML estimation problem. We shall not demand classical consistency since there is no way to achieve it with the nuisance parameters. We'll prescribe fixed values to all the moments $\mu_{M+1},\dots$ above the number M of those we can consistently estimate. These values will be those characteristic to the closest-to-normal distribution and so will depend on unknown first M moments. Surely, the real moments $\mu_{M+1},\dots$ may deviate from the prescribed values and so produce an error in the estimates of $\mu_1,\dots,\mu_M$. We can neither establish an a priori or a posteriori upper limit to this error, nor even guarantee that it vanishes as $N\to\infty$. So the estimation is, strictly speaking, inconsistent.

But we'll show that wr are able to assess the sensitivity of the estimates $\mu_1,\dots,\mu_M$ to the deviations of $\mu_{M+1},\dots$ and, moreover, to nearly all other sources of error. This is a rare possibility. If the sensitivity is not too great, one may hope that as, in general, $\mu_{M+1},\dots$ won't drastically depart from the closest-to-normal values, on the average, the method will work well. But if the a posteriori assessed sensitivity is great, the estimates of $\mu_1,\dots,\mu_M$ are suspicious. Moreover, a priori assessing the sensitivity as a function of unknowns $\mu_1,\dots,\mu_M$, the sensor characteristics, object type and so on, we could outline the more and less perspective fields for remote sensing.

Now, deriving f from the simultaneous observations of X_i's and Y_j's at a test area follows approximately the same line, only we now write the common distribution P(X,Y) as a function of unknown μ's and a's ($f(x)=\sum_{i=0}^{k} a_i x^i$). The increased number of unknowns (plus a's) is compensated by the increased dimensionality of observations (plus X's).

Sometimes, there are some reliable analytical models for an object's interaction with radiation /3/. They give us f explicitly and eliminate the need for test area observations.

3. Sensitivity estimation. Exponential approximation

As shown in Appendix, the sensitivity of ML estimate μ_k (k=1,...,M) to some parameter η involved in ML equations is simply the ratio of the determinants of the matrices of Fisher information about the set ($\mu_1,\ldots,\mu_M$) and ($\mu_1,\ldots,\mu_{k-1},\eta,\ldots,\mu_M$). Fisher information on ($\mu_1,\ldots,\mu_M$) is extremely important in itself as it determines the lower bound of the error due to the limited supply N of data.

Computing a priori Fisher information essentially reduces to $\int^{**}_{**} P(Y)\log P(Y)\,dY$. The logarithm is a stumbling block. It should be approximated to perform the integration analytically. Above that, the Edgeworth expansion now can't be used since it may have negative probabilities at tails. Besides, the logarithm in ML equations may give rise to prohibitive calculations.

As $P(Y)\geq 0$, if may be written as $P(Y)=\exp T(Y)$. T(Y) may be, naturally, approximated as a series in (1/n). Expanding the exponent in Taylor series and equating its terms with the terms of the same order in (1/n) of the Edgeworth series, we find $P(Y)\exp\sum^{M}\theta_j U_j(Y)n^{-j/2}$, which is an exponential type distribution (/4/, chapter 19). Now we may perform the integration.

Besides, distributions of the exponential family had been throughly analyzed in statistics. Approximating P(Y) by them, we may simply look up for efficiently estimable parameters and for their sufficient statistics or whether P(Y) decomposes into the informative and uniformative (about some parameter) parts as stipulated by some methods to resolve the nuisance parameters difficulty.

4. Better classification by taking into account state fluctuations

Mean classification accuracy in remote sensing is 80-85% and is limited mainly by significant cluster overlap. However, as shown in /5/, small homogeneous areas produce much narrower clusters (since the close points' state is correlated) whose classification accuracy could be nearly 100%. This figure can't be achieved straightforwardly for neither may training data be collected at many small areas, nor, in clusterization, may clusters by reliably constructed from small data sets and identified in situ for each area. That's the reason why clusters are usually defined in such a way as to represent the whole frame under study. Then, they are superpositions of many small area clusters shifted because of state fluctuations at a large frame. So they are much wider and the classification accuracy is poor. To improve it, we could decompose a wide cluster into the set of narrow ones if the objects' states were known. But to estimate the state, one must know the object's type. So we are driven towards combined classification - state estimation. Its idea is illustrated by fig.2.

Conclusions

In the way of applying our methods, one should first extract the state-brightness relation f from extensive in situ and remote observations of a test area. Then f could be applied to any other area to assess its mean state from remote observations only (see chapter 2). By the method of chapter 3, one may compute the sensitivity of the estimate to mean error sources and the error resulting from the limited amount of data. One may also compute the a priori error estimate for statistical and point-by-point methods, adopting some possible state probability distribution. If f is not available from test areas, it may in some cases be computed analytically.

In view of all said, we think the efforts to fit our problem into a good theoretical frame were justified.

Appendix

Y to X: not a one-to-one mapping. Let $Y=\frac{1}{n}\sum_{i=1}^{n}y_i$, $X=\frac{1}{n}\sum_{i=1}^{n}x_i$ $y=f(x)$, μ_k^a - moments of x with respect to the point a, m_k^a - the corresponding sampled moments. Regression curve of Y onto X is

$$M(Y/X)=M\left(\sum_{i=1}^{n}\frac{f(x_i)}{n}\,\middle|\,X\right)=M\left(\sum_{i=1}^{n}\sum_{k=0}^{\infty}\frac{f^{(k)}(X)}{K!}\,\frac{(x_i-X)^K}{n}\right)=\sum_{k=0}^{\infty}\frac{f^{(k)}(X)}{k!}\mu_k^X \qquad (1)$$

$$D(Y/X)=D\left(\sum_{k=0}^{\infty}\frac{f^{(k)}(X)}{k}m_k^X\,\middle|\,X\right)=\sum_{k,l=0}^{\infty}\frac{f^{(k)}(X)f^{(l)}(X)}{nk!\,l!}\left(\mu_{k+l}^X-\mu_l^X\,\mu_k^X\right) \qquad (2)$$

$\left(\mu_k^X=M\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-X)^k\,\middle|\,\frac{1}{n}\sum_{i=1}^{n}x_i\equiv X\right)\right.$ may be expressed as a function of $\mu_1,\ldots,\mu_K$ and X). It is clear that, generally (even for normal x), $D(Y/X)\neq 0$, unless f is linear. As seen from fig. 1, $D(X/Y)\neq 0$ as well, and X may not be uniquely determined from an observed Y. The a priori error estimate for point-by-point methods is $D(X/Y)=D(Y/X)/\left[\partial M(Y/X)/\partial X\right]^2$ (see fig.1). The a posteriori estimate is as (2) with f^{ν}, ν, Y instead of f, μ, X, provided that f is a one-to-one mapping (ν are the moments of y and may be computed through the observed moments of Y). If this assumption is violated, the a posteriori estimate may be obtained from (2) where μ's should be substituted with their estimates computed by the chapter 2 method.

Edgeworth series. P(Y), the probability distribution of $Y=\frac{1}{n}\sum_{i=1}^{n}y_i$, may be represented as Edgeworth series (/6/, chapter 17):

$$P(Y)=\frac{1}{\sqrt{2\pi\lambda_2}}\exp\left[-\frac{1}{2}\tilde{Y}^2\right]\left\{1+\sum_{j=3}^{M}\frac{Q(\lambda,\tilde{Y})}{j!}\right\}+R_M;\quad \tilde{Y}\equiv\frac{Y-\lambda_1}{\sqrt{\lambda_2}} \qquad (3)$$

where λ are the central moments of Y; $Q_3(\lambda,\tilde{Y})=\lambda_3 H_3(\tilde{Y})/\lambda_2^{3/2}$; $Q_4=(\frac{\lambda_4}{\lambda_2^2}-3)H_4(\tilde{Y})+\frac{\lambda_3^2}{3\lambda_2^3}H_6(\tilde{Y}),\ldots$, or, ν being the central moments of Y, $\lambda_1=\nu_1$, $\lambda_2=\nu_2/\sqrt{n}$; $Q_3(\nu,n,\tilde{Y})=\nu_3 H_3(\tilde{Y}/\sqrt{n})\nu_2^{3/2}$; $Q_4=\nu_4 H_4(\tilde{Y})/n)\nu_2^2+\nu_3^2 H_6(\tilde{Y})/3n)\nu_3^2$, so that $Q_j(\nu,n,\tilde{Y})$ is the $(-j/2+1)$-th power of n.

Maximum likelihood method. ML equations are

$$\sum_{i=1}^{N}\frac{\partial \log P(Y_i)}{\partial c_k}=0\quad(k=1,\ldots,M) \tag{4}$$

where c are the unknown parameters of P(x) and may be expressed as function of $\mu_1,\ldots,\mu_k$. C_k is defined as $M(H_k(\tilde{x}))$ where $\tilde{x}=(x-\mu_1)/\sigma$ and σ is the standard error of x. Then, $C_k=\sum_{i=0}^{k}\alpha_{1k}\mu_1$ where α are the coefficients of the power series expansion of the k-th Hermite polynomial. As P(x) may by expanded into Gram-Charlier series of the form

$$P(\tilde{x})=\exp(-\frac{\tilde{x}^2}{2})\sum_{k=0}^{\infty}C_k H_k(\tilde{x}),$$

a distribution with fixed first M moments (of c's) will be closest to normal when $C_{M+1}=C_{M+2}=,,,0$. If we write μ_K as a function of $C_1\ldots C_K$, then express ν's as the functions of μ's (.i. c(s)) and of the coefficients a of the power series expansion of f: $y=F(x)=\sum_{i=0}^{\mu}a_i x^i$; $\nu_1=\sum_{i=0}^{\mu}a_i\mu_i$; $\nu_2=\sum_{i,j=0}^{\mu}a_i a_j\mu_{i+j}$, and then substitute the ν's into (3), P(Y) will become a series with coefficients depending on a's and C's. Now, as set forth in chapter 2, C_{M+1}, $C_{M+2},\ldots$ are assumed equal to zero, and, substituting P(Y) into ML equations (4), we get a system of M equations with M unknowns (assuming a's known). If we should estimate both μ's and a's (when determining f from test area observations), P(X,Y) must be substituted into (4) in the same Edgeworth form. Numerically solving ML equations will be much easier if P(Y) or P(X,Y) is expressed in the exponential form (as shown below) so that the logarithm vanishes.

Fisher information. From Kramer-Rao theorem it follows that the error of the estimate $\hat{\mu}$ of a parameter μ has the lower limit:

$$D(\hat{\mu}-\mu)\geqslant\frac{-1}{NM(\frac{\partial^2\log P(Y)}{\partial\mu^2})}=\frac{1}{NM(\frac{\partial\log P(Y)}{\partial\mu})^2}$$

The denominator is the Fisher information I_μ on μ. For several parameters, one should consider the information matrix with the elements $I_{\mu_i,\mu_j}=-NM(\partial^2\log P(Y)/\partial\mu_i\partial\mu_j)$:

$$D(t(\hat{\mu}_1,\ldots,\hat{\mu}_M)-t(\mu_1,\ldots,\mu_M))\geqslant\sum_{k,l=1}^{M}\frac{\partial t}{\partial\mu_k}\frac{\partial t}{\partial\mu_l}I^{-1}_{\mu_k,\mu_l}$$

When $N\to\infty$, the error of a ML estimate approaches the Kramer-Rao limit. Assigning some values to μ's, one could thus a priori estimate the lower bound of error. A posteriori estimate

may be obtained (by substituting the expectation by the sample mean and μ's by their ML estimates) as $-N\sum_{i=1}^{N} \partial^2 \log P(Y_i)/\partial\mu^2$. As shown in /7/, this is even closer to ML error than the classical Kramer-Rao expression.

Now, as stated in chapter 3, P(Y) may be approximated by an exponential form distribution, e.g. up to terms 1/n:

$$P(\tilde{Y})=\exp\left(-\frac{\tilde{Y}^2}{2}\right)\exp\left\{\frac{\lambda_3}{\lambda_2^{3/2}}H_3(\tilde{Y})+\left(\frac{\lambda_4}{24\lambda_2^2}-\frac{\lambda_3^2}{4\lambda_2^3}-\frac{1}{8}\right)H_4(\tilde{Y})-\right.$$
$$\left.-\frac{\lambda_3^2}{72\lambda_2^3}H_6(\tilde{Y})-\frac{\lambda_3^2}{36\lambda_2^3}\left[H_2(\tilde{Y})+6\right]\right\}. \qquad (5)$$

As a matter of fact, the positivity of P(Y) is achieved by a slight alteration of its tails as compared to Edgeworth representation, so that the change is less than the first neglected Edgeworth term. So the precision of the exponential approximation is essentially the same as of Edgeworth series. It is enough for the a posteriori formula above, but a priori formula involves integration which may diverge if the tails are not approximated accurately (which is not the case for Edgerworth series). But , though exponential approximation proceeds from Edgeworth series, it may correct the latter in the right sense by making the tails positive and so have greater accuracy at tails. We'll abandon the question at that point.

Sensitivity of estimates to errors. If $\hat{\mu}$ is determined from the equation $F(\hat{\mu},\theta)=0$, by the implicit function theorem $\partial\hat{\mu}/\partial\theta=-(\partial F/\partial\theta)/(\partial F/\partial\hat{\mu})$. As in ML estimation F is simply $\sum\partial P(\hat{\mu},\theta,Y_i)/\partial\hat{\mu}$, both the numerator and the denominator are the above-mentioned sample mean formulae for Fisher information on θ and μ, respectively. For several parameters, we'll have the determinants of Fisher information matrix for $(\mu_1,\ldots\mu_M)$ and $(\mu_1,\ldots,\mu_{k-1},\theta,\ldots,\mu_M)$. So to estimate the sensitivity to some neglected moment, or error term, or noice term, P(Y) should be represented so as to include this term, and then the formulae above should be applied.

Bibliography

1. M.E.Bauer, P.H.Swain, R.P.Mroczynski, P.E.Anuta, R.B.MacDonald. Detection of southern corn leaf blight by remote sensing techniques. Proc. 7th Intern.Symp. on Remote Sensing, Ann Arbor, 1971, 693.
2. T.Bard.Nonlinear parameter estimation.Academic Press,1974.
3. A.J.Richardson,C.L.Wiegand,H.W.Gausman,J.A.Cuellar, A.H.Gerbermann. Plant,soil and shadow reflectance components of row crops.Photogramm.eng. and remote sensing, 1975, 41, 11, 1401.
4. M.Kendall, A.Stuart. The advanced theory of statistics. Charles Griffin & Co., London, 1958-1966.
5. M.J.Duggin. On the natural limitations of target differentiation by means of spectral discrimination techniques. Proc. 9th Symp. on Rem. Sens., Ann Arbor, 1974, 499.
6. H.Cramer.Mathematical methods of statistics.Princeton,1958.
7. D.V.Hinkley. Likelihood inference about location and scale parameters. Biometrika, 1978, 65, 2, 253-261.

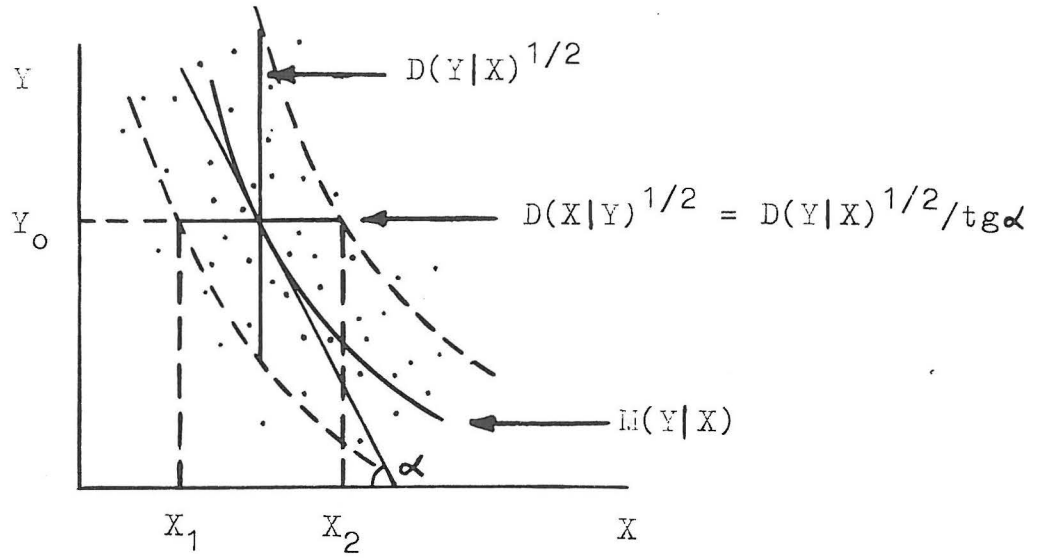

Fig. 1. Regression of the observations Y of the REs on the RE's state variables X.X to Y is not a one-to-one mapping.

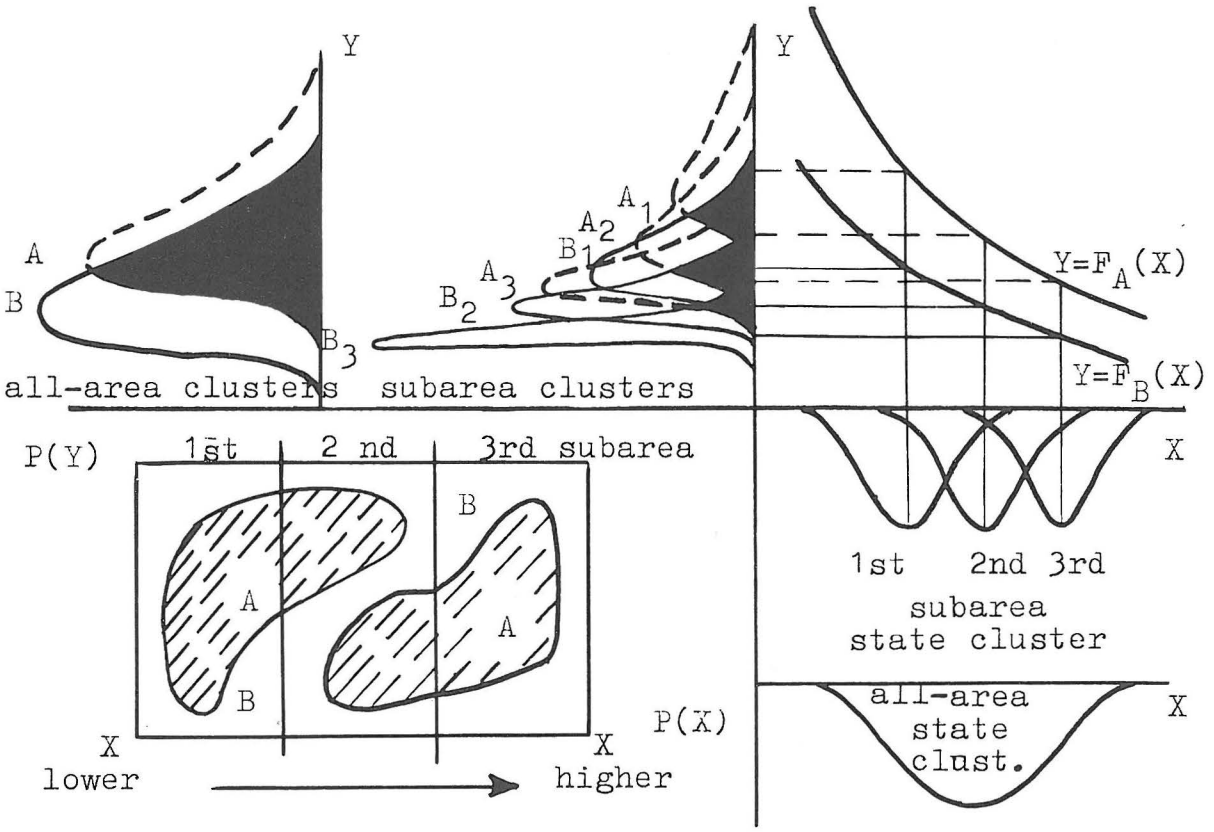

Fig.2. Better classification through taking into account the dependence of Y on X. Error areas are blackened. All-area clusters are significantly overlapping. If f is known for both objects, clusters may be decomposed into pairs with almost no overlap. There is only one possible X cluster for any Y cluster pairs if both objects should have the same X.