

## CLASSIFICATION OF GEOLOGICAL MATERIAL UNITS IN THE GANIKI PLANITIA QUADRANGLE (V14) OF VENUS USING STATISTICAL CLUSTERING METHODS.

J. Richards<sup>1</sup>, J. Hardin<sup>1</sup> and E. B. Gross-fils<sup>2</sup>, <sup>1</sup>Department of Mathematics and Computer Science, Pomona College, Claremont, CA 91711 (joseph.richards@pomona.edu), <sup>2</sup>Department of Geology, Pomona College, Claremont, CA 91711.

**Introduction:** In an ongoing attempt to analyze volcanic and tectonic activity in the Ganiki Planitia quadrangle (V14) of Venus and to construct an accurate stratigraphy of the region, researchers have created a geologic map of the region [e.g., 1]. This geologic map is based on qualitative interpretation of V14 from the Magellan FMAP radar images (75m/pixel) and topography data. The mappers did not explicitly use the numerical information encoded in these images to guide their mapping, nor did they use the available physical property data sets of emissivity, reflectivity, and RMS slope quantitatively. Here we use statistical clustering techniques to analyze the existing map of V14 in order to (a) test whether the map is consistent with the numerical data and (b) identify units whose classification might be incorrect.

**Methods:** We begin our study by assuming that the units are well-drawn and that the geologists' classification of these units is correct. These assumptions are reasonable because geologists have spent three years creating the geologic map using standard planetary mapping techniques. We next used Arcview GIS software to extract the mean, standard deviation and median of pixel values within the boundaries of each unit for each physical property layer; the radar backscatter data employed were latitude-corrected [2]. We then clustered the  $n$  units into  $G$  unit types based on  $v$  numerical variables. By assuming that the existing map is correct, we ensure that any differences that arise between the map and the clustering result are due to properties of the data, and not any initial assumptions about the allocation of the units.

In total, there are 200 units in V14 distributed across 18 groups (unit types). However, five groups have only one member unit. Because the clustering method we use can only handle cases where group membership is two or greater, we assigned these five units to other groups based on their physical characteristics and interpretation. Consequently, we obtained 13 groups, with the smallest group having three units.

**Mixture Models.** We assume the data points (geologic units) are generated from a mixture of probability distributions, with each probability distribution corresponding to a different group (type of unit). We say that  $f_k(\mathbf{x}_i | \theta_k)$  is the probability density of data point  $\mathbf{x}_i$  being in group  $k$  given that  $\theta_k$  is the set of parameters that defines cluster  $k$  [3].

We model the 200 data points as being distributed multivariate normally for each cluster. Therefore,

$$f_k(\mathbf{x}_i | \mu_k, \Sigma_k) = \frac{\exp\{-0.5(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1}(\mathbf{x}_i - \mu_k)\}}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \quad (1)$$

where  $\mu_k$  is the vector of mean values of each variable for cluster  $k$  and  $\Sigma_k$  is the covariance matrix for cluster  $k$ . Equation (1) is the probability density function (pdf) for the multivariate normal distribution.

**Missing Data.** Notice that we do not have a variable that indicates the group membership of unit  $i$ , even though this is precisely what we want to find. Therefore for each unit we introduce  $\mathbf{z}_i = (z_{i1}, \dots, z_{iG})$  where  $z_{ik}$  is 1 if unit  $i$  is in group  $k$  and is 0 else. The  $z_i$ 's are distributed multinomially with probabilities  $\tau_1, \dots, \tau_G$  where  $\tau_k$  is the probability that a randomly-drawn unit is in group  $k$ . The pdf for  $\mathbf{z}_i$  is

$$f(\mathbf{z}_i | \tau) = \tau_1^{z_{i1}} \tau_2^{z_{i2}} \cdots \tau_G^{z_{iG}} \quad (2)$$

the pdf of a multinomial random variable .

As a result, the probability that point  $\mathbf{x}_i$  is observed based on the parameters that define the groups and the indicator variable for unit  $i$  is

$$f(\mathbf{x}_i | \mu_k, \Sigma_k, z_i) = \prod_{k=1}^G f_k(\mathbf{x}_i | \mu_k, \Sigma_k)^{z_{ik}} \quad (3)$$

where  $f_k(\cdot | \cdot)$  is defined in equation (1).

**Complete Data Likelihood.** We wish to find an equation for the probability that a specific allocation of units into a set of clusters is correct based on the data. Maximizing this function will give our optimal clustering based on the data [4]. Using Bayes' Rule and the definition of conditional probability, one can show that for the  $i^{th}$  unit,

$$L(\mu, \Sigma, \mathbf{z}_i, \tau | \mathbf{x}_i) = f(\mathbf{x}_i | \mu, \Sigma, \mathbf{z}_i, \tau) \cdot f(\mathbf{z}_i | \tau) \quad (4)$$

where  $L$  denotes the likelihood (probability) of observing those parameters based on the data.

By substituting equations (2) and (3) into equation (4) and solving for complete data likelihood, we get

$$L(\mu, \Sigma, \mathbf{z}, \tau | \mathbf{x}) = \prod_{i=1}^n \prod_{k=1}^G f_k(\mathbf{x}_i | \mu_k, \Sigma_k)^{z_{ik}} \tau_k^{z_{ik}} \quad (5)$$

**EM Algorithm.** We want to find the values of the cluster parameters and our missing data  $z$  that maximize equation (5). The general approach to problems

of maximum likelihood with missing data is the expectation-maximization (EM) algorithm [5].

The EM algorithm begins with some initial values for our missing data. Because we are assuming that the existing geologic map is correct, we initiate the algorithm with  $z_{ik}$ 's that are 1 if unit  $i$  is in group  $k$  in the geologic map, and 0 if it is not.

The algorithm then proceeds to the M-step. Here, the values of  $z_{ik}$  are held constant, and equation (5) is maximized with respect to the parameters  $\mu$ ,  $\Sigma$ , and  $\tau$  [5]. In essence, the M-step fits a cluster to each group where group membership is held constant.

Next, the algorithm performs the E-step, which calculates the  $z_{ik}$ 's based on the parameters found in the M-step and the pdf of the multivariate normal distribution using an equation for posterior probability.

The algorithm continues until a tolerance value in the increase in the likelihood equation in successive iterations is reached. Note that the final  $z_{ik}$ 's represent the posterior probability that unit  $i$  is in group  $k$ . These values can be converted into a classification of the units by assigning each unit to the group of its maximum posterior probability.

**Results:** The variables we have available are the mean, standard deviation and median of pixel values corresponding to five different data sets. However, emissivity and reflectivity are both measurements of the same property, and are inversely related (with some variation from true inverses in our data). Also, mean and median are both measures of center. Therefore it is wise to cluster using different combinations of variables and then to compare the results.

We cluster using the package mclust in the statistical program R. When we cluster with the mean, standard deviation and median of all the properties except emissivity, 41 units (20.5%) are allocated differently than what is specified by the geologic map. This is a large number of units being reclassified. Since we are maintaining a high level of confidence in the original geologic map, it is advantageous to look at trends across different clustering results. When we look at four different clustering results based on different combinations of variables, only 14 units (7%) are allocated differently every time (Table 1).

**Discussion:** The most commonly reclassified units are those identified as volcanic edifice flows (fe) in the original geologic map. Of these five units, three are classified as intermediate or dark plains units by our clustering methods; in these instances it was readily apparent that failure to recognize the edifice geometry led to the reclassification as plains. The other two, units 6 and 10, are classified as either tessera or intermediate plains; the tessera classification may be a result of a higher than normal concentration of tectonic

lineaments within these volcanic units. Clearly there is something anomalous in the data for these units and they should be further examined.

Two units originally identified as tessera (t) are classified differently by statistical clustering: unit 3 is classified as Lehevhev lineated plains and unit 5 is classified as a volcanic edifice flow unit. The classification of unit 3 is explained by the large density of lineaments on Lehevhev plains units; failure to observe the geometry of the lineaments and relationship with surrounding units led to the reclassification. The classification of unit 5 is unexpected, and the unit should be further examined. Three small dark plains units (prc), an intermediate plains unit (prb), and two large edifice plains units (pe) are reclassified by clustering as well (Table 1), and it is our recommendation that they should be further studied.

Future directions for this project are to incorporate unit area and variable importance into the likelihood equation. Incorporating unit area will allow us to give bigger units more importance in determining the parameters for each cluster. Integrating variable weights will allow us to give more weight to backscatter, which has higher resolution than the physical property variables.

**References:** [1] Grosfils et al., this volume, *LPSC XXXVI*, 2005. [2] Long S. M. and Grosfils E. B., this volume, *LPSC XXXVI*, 2005 [3] McLachlan G. J. and Basford K. E., Mixture Models: Inference and Applications to Clustering, 9-15, 1997. [4] Fraley C. and Raftery A. E., *JASA*, 97, 611-631, 2002. [5] McLachlan, G. J. and Krishnan T., The EM Algorithm and Extensions, 82-109, 1997.

Unit	Unit Area (m <sup>2</sup> )	Original Class	Clustering Classifications
1	$3.087 \times 10^{10}$	pe	plb (2), prb, prc
2	$8.812 \times 10^{10}$	pe	fe (2), pla, prb
3	$7.151 \times 10^9$	t	pLlb (4)
4	$2.432 \times 10^{10}$	prb	t(3), fe
5	$3.662 \times 10^8$	prc	t (3), pe
6	$3.414 \times 10^{10}$	fe	t (3), prb
7	$1.245 \times 10^{10}$	plb	prb (2), c, fe
8	$2.373 \times 10^{11}$	fe	prc (3), plb
9	$2.068 \times 10^9$	fe	prb (2), prc, pl
10	$1.915 \times 10^{10}$	fe	prb (2), t (2)
11	$1.809 \times 10^9$	t	fe (4)
12	$6.462 \times 10^9$	fe	prc (2), prb, pra
13	$1.917 \times 10^9$	prc	pe (4)
14	$6.677 \times 10^8$	prc	fe (3), plb

**Table 1:** Fourteen units were classified differently by statistical clustering than the original geomorphological efforts. Classification system based on [1].