APPLYING PREDICTIVE FINANCIAL RISK MODELS TO THE IDENTIFICATION OF LUNAR BASALT SPECTRA.

I. Antonenko, Planetary Institute of Toronto, 197 Fairview Ave. Toronto, ON M6P 3A6, Canada (PlanetaryInstituteofToronto@yahoo.ca).

Introduction: Identifying basalts in surface or subsurface deposits is fundamental for understanding the distribution of volcanic deposits on the Moon. Basalts are identified using spectra [e.g., 1] and various techniques are used to automate the process [e.g., 2, 3, 4].

In the finance industry, predictive models are used to automate the risk identification [5], for lending and regulatory compliance requirements. The models calculate the probability a customer will default based on past behavior. Existing records of both defaulted and non-defaulted accounts are used to create a predictive scorecard that is calibrated to differentiate risk levels.

Previous basalt identification work [3] in the Mare Humorum area of the Moon (Fig.1) produced a data set like those used in finance, with a set of manually predetermined basalt and non-basalt identifications associated with a series of characteristics, namely their spectra. A rudimentary financial risk modelling methodology was applied to this data set, in order to evaluate the method's applicability to planetary problems.

Method: Clementine multispectral data (Fig.1) was evaluated using R statistical software [6]. Variables were calculated from spectral bands, including band values, depths, slopes, ratios, etc., both normalized and not. The values of each variable were then binned using the smbinning package [7] to reduce the number of possible attributes for each variable and allow the information value (IV) to be calculated. Variables with a low IV are considered to be less predictive, so those with IV<1 were eliminated from consideration.

The data set was divided into training and validation partitions. A random sampling of 70% of the spectra were selected for training the model, with the remaining 30% set aside for validation.

The data variables were clustered using the ClustOfVar package [8] to reduce multicollinearity. The optimal number of clusters was determined by evaluating the gain in cohesion [8]. For our training data, 9 clusters were found to be reasonable, since additional clusters produced minimal gain in cohesion.

Clustering identifies groups of variables that are highly correlated to each other, but uncorrelated to other variables. Selecting a few representative variables from each cluster reduces the number of variables with little loss of information. Variables that are closest to their cluster's calculated centre (have a high square correlation [8]), best represent that cluster. However, clusters with high IV represent the cluster's strongest predictors. Thus, for each cluster both the variable with the highest IV and the variable with the highest square correlation were selected. These two conditions may coincide in one variable, so a total of 17 variables were chosen. Two variables, representing the 950nm and 1000nm normalized band depths, also added based on science considerations, resulting in 19 variables.

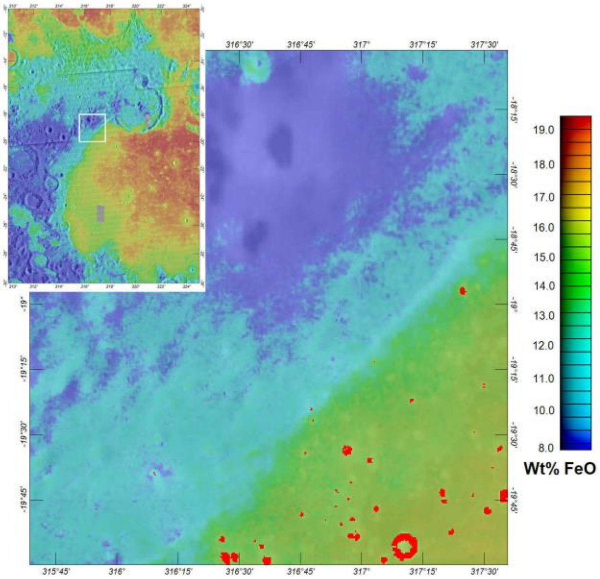

Figure 1: Study area with location north of Mare Humorum shown by square in the inset. Basemap, combined LO and LOLA; overlay, Clementine FeO; red squares are manually identified basalt spectra.

Stepwise logistic regression [6] was used to fit the 19 variables to a basalt/non-basalt indicator in the training data set. Stepwise logistic regression is an iterative process, where at each step the variable that most improves the model's fitness is added and any variable that reduce the model's fitness as a result is removed. Fitness is determined by Akaike's "An Information Criteria (AIC) [6], where lower AIC indicates a better model. The resulting scorecard, consisting of 7 final variables, is presented in Table 1.

Scorecard Validation and Calibration: The scorecard was applied to the training and validation partitions of the data set. Figure 2 shows the total scores distribution for the two partitions. Although some overlap exists, it is found to be minimal in both partitions.

To use the scorecard, a basalt cut-off point needs can be determined. The optimal cutoff point can be evaluated by considering the false positive rate, total accuracy, and predictive value of the model at various cut-offs. Table 2 shows that a cut-off point around 1.5 is optimal with respect to total accuracy. Applying this value to the validation partition gives results that are comparable to the training partition, supporting the validity of the model.

Table 1: Generated Model Scorecard

Variable Name	Variable Definition	Bin Range	Points	Std. Error	z-value
Intercept			-25.7829	2324.91	-0.0110
SNBC	Slope between 750nm and 900nm bands, each normalized to the 750nm band.	< -0.000152	-3.6960	0.4675	-7.9050
		-0.000152 to -0.000125	-5.3729	0.7119	-7.5470
		-0.000125 to -0.000104	-6.8240	1.0209	-6.6840
		-0.000104 to -0.000082	-25.0460	4766.98	-0.0050
		> -0.000082	-18.1676	2182.63	-0.0080
SRNABCD	Slope between 415nm and 750nm bands divided by slope between 900nm and 950nm bands, each band normalized to the 750nm band.	< -9.6879	3.3817	1.1501	2.9400
		-9.6879 to -6.1851	6.1664	1.1328	5.4430
		-6.1851 to -4.8605	7.0926	1.1871	5.9750
		-4.8605 to -3.9138	7.6633	1.1909	6.4350
		-3.9138 to -3.2637	8.6426	1.1900	7.2630
		-3.2637 to -2.6357	8.6527	1.1967	7.2300
		-2.6357 to -1.7823	4.5628	1.0940	4.1710
		-1.7823 to -1.075	0.8861	1.1083	0.7990
		-1.075 to 26.4179	-4.7202	1.7956	-2.6290
		> 26.4179	4.2077	1.5068	2.7920
		Is Null	1.9479	1.6397	1.1880
DNAC	Difference between 415nm and 900nm band values, each band normalized to the 750nm band.	< -0.1427	1.4891	1.2781	1.1650
		-0.1427 to -0.1394	0.6902	1.2746	0.5420
		-0.1394 to -0.1344	2.0846	1.1337	1.8390
		-0.1344 to -0.1306	2.2899	1.1570	1.9790
		-0.1306 to -0.1245	3.5056	1.1376	3.0820
		-0.1245 to -0.1108	4.6915	1.1411	4.1110
		-0.1108 to -0.0963	6.4626	1.2144	5.3220
		-0.0963 to -0.0863	7.1577	1.3665	5.2380
		-0.0863 to -0.0775	8.0279	1.5605	5.1440
		> -0.078	11.2042	2.1700	5.1630
SRNCDDE	Slope between 900nm and 950nm bands divided by slope between 950nm and 1000nm bands, each band normalized to the 750nm band.	< -1.75	-0.0836	0.4101	-0.2040
		-1.75 to -1.1837	0.0043	0.4599	0.0090
		-1.1837 to -0.7174	-0.4339	0.4792	-0.9050
		-0.7174 to -0.3651	-0.4415	0.5587	-0.7900
		-0.3651 to -0.234	0.7705	0.8251	0.9340
		-0.234 to -0.1333	3.1505	1.1715	2.6890
		-0.1333 to -0.0377	4.3882	1.9019	2.3070
		-0.0377 to 0.2344	-1.9787	1.6656	-1.1880
		0.2344 to 0.9744	-26.5440	4876.95	-0.0050
		0.9744 to 2.0556	-10.7869	1.0435	-10.3370
		> 2.0556	0.2378	0.7400	0.3210
		Is Null	-1.2423	1.1269	-1.1020
DNBD	Difference between 750nm and 950nm band values, each band normalized to the 750nm band.	< 0.0197	-12.9852	4913.1	-0.0030
		0.0197 to 0.0237	2.2682	2172.5	0.0010
		0.0237 to 0.029	5.2885	2172.5	0.0020
		0.029 to 0.0347	6.7210	2172.5	0.0030
		0.0347 to 0.1054	7.1768	2172.5	0.0030
DBC	Difference between 750nm and 900nm band values.	0.1054 to 0.0074	9.7642	827.7	0.0120
		0.0074 to 0.0105	11.5597	827.7	0.0140
		0.0105 to 0.017	12.0611	827.7	0.0150
		0.017 to 0.0215	12.3737	827.7	0.0150
		0.0215 to 0.0273	13.2345	827.7	0.0160
		> 0.0273	13.7044	827.7	0.0170
SNAB	Slope between 415nm and 750nm bands, each band normalized to the 750nm band.	< 0.000349	2.5889	1.5701	1.6490
		0.000349 to 0.000369	0.1831	1.7014	0.1080
		0.000369 to 0.000382	-0.4347	1.7384	-0.2500
		0.000382 to 0.000393	-0.2926	1.7076	-0.1710
		0.000393 to 0.000431	-0.1903	1.7647	-0.1080
		0.000431 to 0.000439	-0.1659	1.8624	-0.0890
		> 0.000439	0.3652	1.8469	0.1980

Table 2 also shows that the new model significantly outperforms the author's old estimation method [3].

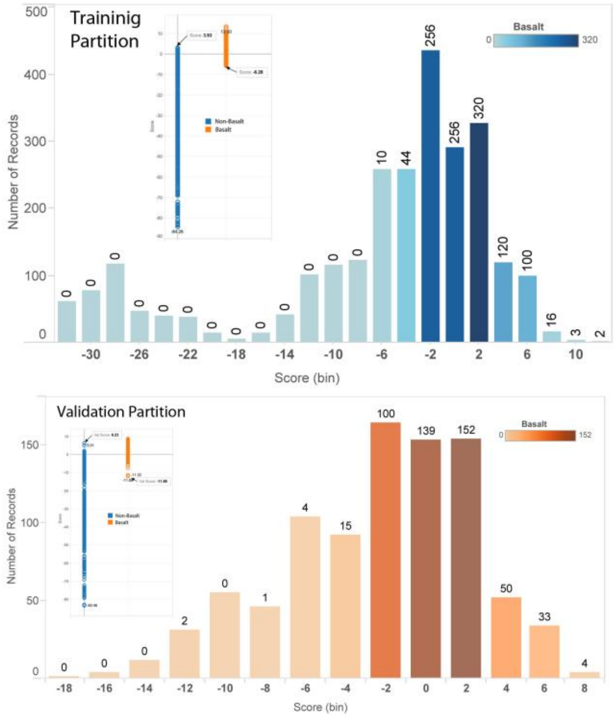

Figure 2: Histogram of score totals from Table 1. Colours and bar labels represent the number of identified basalt spectra in each bin. Insets show the overlap between basalt and non-basalt scores.

Conclusions: The simple scorecard developed in this study demonstrates that financial risk scoring methods can be successfully applied to the problem of spectral identification, when a data set of known classifications is available. Future work will build on this study, in order to how data mining methods in other fields can be applied to planetary data sets.

Table 1: Generated Model Scorecard

Training	False -ve	False +ve	True -ve	True +ve	False +ve Rate	True +ve Rate	Total Accuracy	Predictive Value
Model Score >-3	31	307	22569	1096	1.3%	97.2%	98.6%	78.1%
Model Score >-2	54	221	22655	1073	1.0%	95.2%	98.9%	82.9%
Model Score >-1.5	78	161	22715	1049	0.7%	93.1%	99.0%	86.7%
Model Score >-1	135	124	22752	992	0.5%	88.0%	98.9%	88.9%
Model Score >0	310	42	22834	817	0.2%	72.5%	98.5%	95.1%
Old Estimate	124	781	22095	1003	3.4%	89.0%	96.2%	56.2%
Validation	**False -ve**	**False +ve**	**True -ve**	**True +ve**	**False +ve Rate**	**True +ve Rate**	**Total Accuracy**	**Predictive Value**
Model Score >-1.5	33	54	9734	467	0.6%	93.4%	99.2%	89.6%
Old Estimate	48	297	9491	452	3.0%	90.4%	96.6%	60.3%

References: [1] Antonenko I. (1999) *Volumes of Cryptomafic Deposits on the W. Limb of the Moon: Implications for Lunar Volcanism* (Thesis), Brown U. Providence, R.I. p305. [2] Tompkins S. and Pieters C.M. (1999) *MaPS.*, *34*, 25-41. [3] Antonenko I. and Osinski G.R. (2011) *PSS* ***59***, 715-721. [4] Cheek L.C., *et al.* (2011) *JGR* **116**, E00G02. [5] Siddiqi N. (2005) Credit Risk Scorecards, Wiley, USA, p208. [6] R Core Team (2013). ISBN 3-900051-07-0, http://www.R-project.org/. [7] R Package 'smbinning' Version 0.2 (2015), http://www.scoringmodeling.com/rpackage/smbinning/. [8] Chavent M. *et al.* (2012) *JSS* **50**(13).