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

I. Antonenko

Planetary Institute of Toronto, PlanetaryInstituteofToronto@yahoo.ca

Introduction

- Various techniques have been used to automate the identification of basalt spectra in lunar data [e.g., 1, 2, 3].
- The finance industry routinely uses automated methods to model the probability of an account defaulting [4].
- Previous work [2] in the Humorum region of the Moon (Figure 1) produced a large data set of manually identified basalt spectra.
- The data set of [2] is suitable for financial modeling methods, which were applied to evaluate their usefulness in identifying basalts.

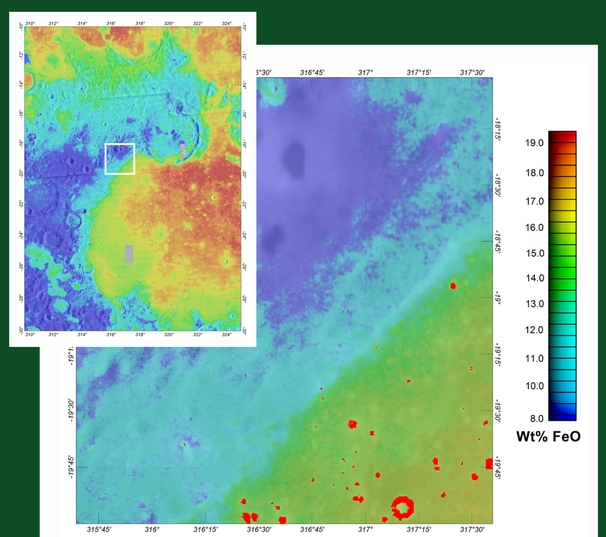

Figure 1: Study area, north-west shore of Mare Humorum. The exact location is shown by the white square on the inset map. The basemap underlay consists of combined LO and LOLA image data. The colour overlay gives Clementine FeO concentrations. Red squares indicate locations of manually identified basalt spectra.

Table 1: Champion 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

Figure 2: Histogram of scores within the training data partition for the champion model. Colours show the percentage of manually identified basalt spectra in each bin. Labels represent the number of basalt spectra per bin. The histogram is truncated to highlight high scores, but the full score range is shown in the inset, which also illustrates the overlap between basalt and non-basalt scores.

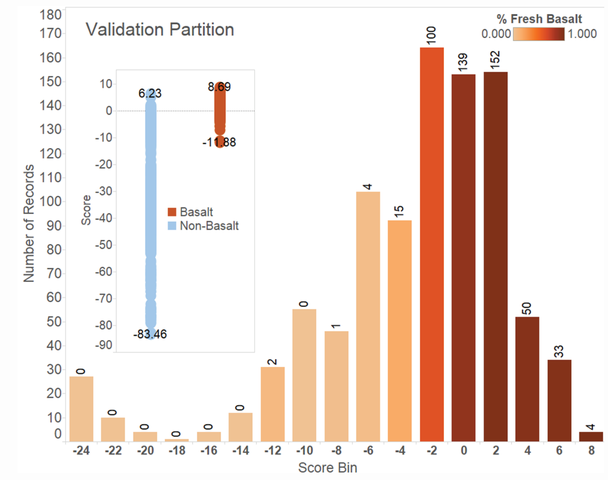

Figure 3: Histogram of scores in the validation partition for the champion model. The % and number of basalt spectra per bin is shown by colours and labels, respectively. The histogram is truncated for high scores, but the full range and basalt/non-basalt overlap is shown in the inset. The score distribution and % of basalt spectra is fairly consistent with that of the training partition (Figure 2).

Method

- Clementine multispectral data (Figure 1) was evaluated using the **R** statistical software package [5].
- **Variables** were calculated from spectral bands, including band values, depths, slopes, ratios, etc., both normalized and not.
- The attributes for each variable were **binned** using the smbinning package [6] and their information value (IV) calculated.
- Variables with low **IV** are considered to be less predictive and so those with low IV were eliminated from consideration.
- The variables were **clustered** using the ClustOfVar package [7] to reduce multicollinearity. The optimal number of clusters was determined by evaluating the gain in cohesion [7].
- Two representative variables were selected from each cluster using the criteria:
  - closest to cluster centre (high square correlation [7]),
  - high IV.
- Some variables were specifically included/excluded based on logical considerations.
- A random **sampling** of the data (70%) was selected for training the model, with the remainder (30%) set aside for validation.
- Iterative stepwise logistic **regression** [5] was used on the training partition to fit the selected variables to a basalt indicator.
- **Multiple models** were generated to allow the most fit to be selected.

Benchmarks

Unselected models are compared as benchmarks to the champion model. Two such benchmark models are presented below.

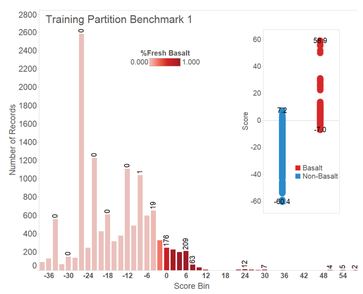

Figure 4: Histogram of scores in the training partition for Benchmark 1. Details are as for Figure 2. A number of exceptionally high scores point to the unsuitability of this model.

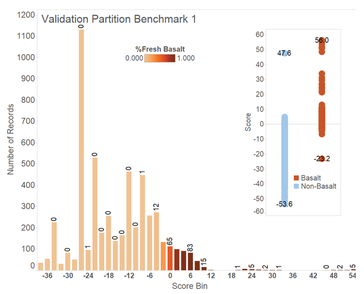

Figure 5: Histogram of scores in the validation partition for Benchmark 1. Details are as in Figure 4. Performance for the validation partition is consistent with that of the training partition (Figure 4).

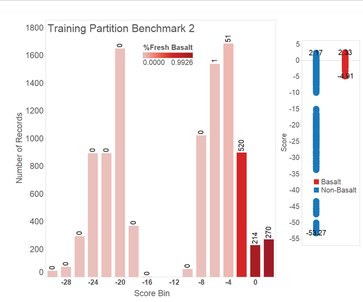

Figure 6: Histogram of scores in the training partition for Benchmark 2. Details are as for Figure 2. The complete overlap of basalt and non-basalt scores illustrates the unsuitability of this model.

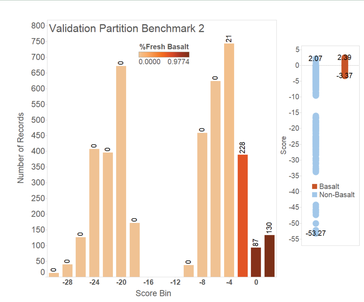

Figure 7: Histogram of scores in the validation partition for Benchmark 2. Details are as in Figure 6. Performance for the validation partition is consistent with that of the training partition (Figure 6).

Table 2: Champion Scorecard Calibration

	False -ve	False +ve	True -ve	True +ve	False +ve Rate	True +ve Rate	Total Accuracy
Model Score >-3	31	307	22569	1096	1.3%	97.2%	98.6%
Model Score >-2	54	221	22655	1073	1.0%	95.2%	98.9%
Model Score >-1.6	70	183	22,693	1,057	0.8%	93.8%	98.9%
Model Score >-1.5	78	161	22715	1049	0.7%	93.1%	99.0%
Model Score >-1.4	97	151	22,725	1,030	0.7%	91.4%	98.9%
Model Score >-1	135	124	22752	992	0.5%	88.0%	98.9%
Model Score >0	310	42	22834	817	0.2%	72.5%	98.5%

Model Assessment

- Model performance was evaluated (Table 3).
- Training and validation data sets were compared to ensure robustness.
- Models were compared to select a champion model.
- Score behaviour for the manually identified basalt spectra was evaluated (Figures 2 through 7).
- Fitness is generally determined by Akaike's "An Information Criteria" (AIC) [5], but here low AIC produced unsuitable models (Figures 4 through 7).
- Generally, models performed fairly comparably in training and validation data partitions.
- All models, even unsuitable ones, performed better than the model from previous work [2].

Conclusions

The scorecards developed in this study demonstrate 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 determine how data mining methods in other fields can be applied to planetary data sets.

Calibration

- The regression procedure [5] produces a scorecard (Table 1) to evaluate pixel data.
  - Scorecard points were assigned for each variable based on pixel values.
  - Points were summed over all variables to give a pixel's score.
- Scores for the training data were used to calibrate basalt cutoffs (Table 2).
  - Cutoffs were selected based on Total Accuracy.

Table 3: Scorecard Assessment

	Model	AIC	False +ve Rate	True +ve Rate	Total Accuracy	Predictive Value
Training	Champion	767.55	0.7%	93.1%	99.00%	86.7%
	Benchmark 1	559.84	0.7%	90.1%	98.64%	91.7%
	Benchmark 2	756.28	0.8%	88.7%	98.37%	89.5%
	Previous Work [2]		3.4%	89.0%	96.23%	56.2%
Validation	Champion	767.55	0.6%	93.4%	99.15%	89.6%
	Benchmark 1	559.84	2.3%	96.4%	97.57%	77.7%
	Benchmark 2	756.28	0.9%	86.9%	98.14%	88.8%
	Previous Work [2]		3.0%	90.4%	96.65%	60.3%

References

[1] Tompkins S. and Pieters C.M. (1999) MaPS., 34, 25-41.
[2] Antonenko I. and Osinski G.R. (2011) PSS 59, 715-721.
[3] Cheek L.C., et al. (2011) JGR 116, E00G02.
[4] Siddiqi N. (2005) Credit Risk Scorecards, Wiley, USA, p208.
[5] R Core Team (2013). ISBN 3-900051-07-0, http://www.R-project.org/.
[6] R Package 'smbinning' Version 0.2 (2015), http://www.scoringmodeling.com/rpackage/smbinning/.
[7] Chavent M. et al. (2012) JSS 50(13).