Exploiting the Mercury surface reflectance spectroscopy dataset from MESSENGER: Making sense of three million spectra.

Mario D'Amore[1], Jörn Helbert[1], Gregory M. Holsclaw[2], Noam R. Izenberg[3], William E. McClintock[2], James W. Head[4] and Sean C. Solomon[5], [1]Institute for Planetary Research, DLR, Rutherfordstrasse 2, Berlin, Germany; [2]Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, CO 80303, USA; [3]Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723, USA; [4]Department of Geological Sciences, Brown University, Providence, RI 02912, USA; [5]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, USA.

Introduction: The Mercury Atmospheric and Surface and Composition Spectrometer (MASCS) on the MErcury Surface, Space ENvironment, GEochemistry, and Ranging (MESSENGER) spacecraft has mapped the surface of Mercury on a global basis during its one-year primary orbital mission and the first third of its extended mission, producing more than three million spectra [1-3].

To analyze this large dataset we make use of our recently developed advanced database management system. This system allows the extraction and simultaneous analysis of large amounts of data, transparent to the underlying data structure. As a test case we analyze here the statistical distribution of MASCS normalized reflectance at a few selected wavelengths.

We obtain a separation of Mercury's surface into spectral classes that are coherent with the results from unsupervised cluster analysis on the same data [3].

Data collection and analysis: The primary challenge to analyzing this dataset is to cope with its large size [2]. In earlier studies of MASCS data, we combined several approaches, ranging from principal component analysis (PCA) to unsupervised cluster analysis and regridding to fixed global and local grids [3-6]. Each of those techniques provides insights into spectral variations for different aspects of the dataset, but the growing data volume quickly overcame each method. The most recent version of our data analysis procedure uses PostgreSQL and PostGIS, a type of database management that controls the creation, integrity, maintenance, and use of a spatially enabled database [7]. We extract all the ancillary data from data files [4], such as sensor temperature and spacecraft and instrument geometrical information, together with spectral parameters. All parameters can be searched in combination with customized ranges, and the search returns pointers to the relevant spectra.

Following the approach described earlier [3,4], the reflectance data are normalized at 700 nm as a first-order photometric correction, leaving (as of January 2013) a hyperspectral dataset of around three million data points characterized by spectral reflectance in the visible- and near infrared and several ancillary parameters. Here we have extracted the normalized reflectance values at 350 nm, 450 nm, and 650 nm from the database and calculated first-order Gaussian fits to histograms of the data (Fig. 1), yielding a measure of the statistical distribution of the reflectance at each wavelength.

The statistical distribution embeds several factors, such as the inherent reflectance distribution, residual geometrical effects, instrumental error, and random noise. The contributions of these factors change with increased data sampling rate on the ground: random noise can be reduced, geometrical effects will likely increase, and the reflectance distribution and instrumental error should be unaffected. Increasing the number of measurements analyzed might improve the global distribution of reflectance by reducing the noise, particularly if residual geometrical effects are addressed.

We fit the global data to a first-order normal distribution. Focusing on the 450 nm band (Fig. 2), we find that we can approximate the residuals on both tails of the main distribution with

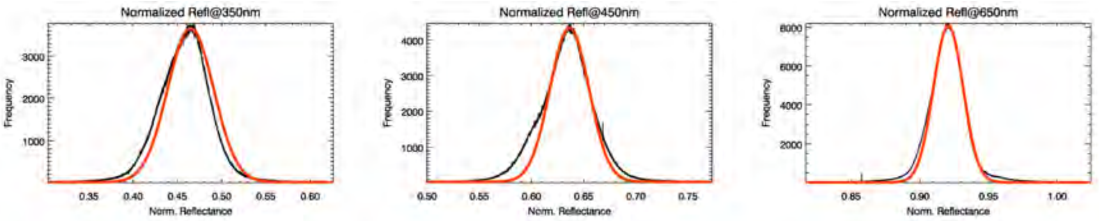

Fig. 1. MASCS normalized reflectance at three fixed wavelengths. From left: 350 nm, 450 nm, and 650 nm. The distribution of reflectance values at 450 nm shows a pronounced deviation from a normal distribution.

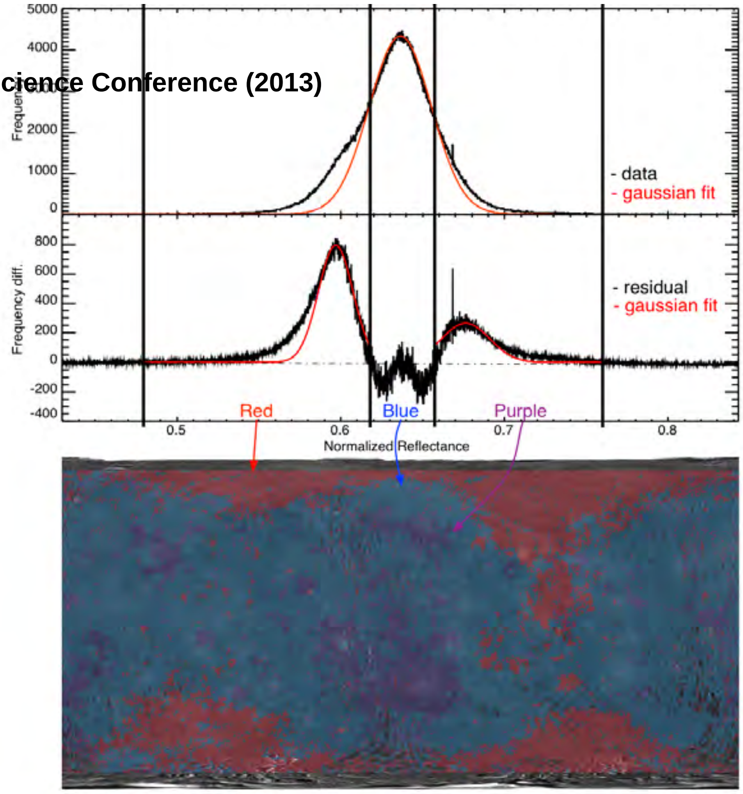

Fig. 2. (Top) Distribution of normalized reflectance at 450 nm and a single Gaussian fit to the data. (Middle) Distribution of the residual from the top panel and Gaussian fits to the tail residuals. (Bottom) Geographical distribution of ranges of values in the residual plot, grouped by color. The contributions (see text) to the observed reflectance could vary, in part due to the lower spatial resolution in the southern hemisphere.

Gaussian distributions. The central part of the residual distribution still shows the main Gaussian distribution (Fig. 2, middle).

The central fit has an estimated mean of 0.636 and a standard deviation (σ) of 0.018; the left-wing Gaussian has a mean value of 0.597 and a σ of 0.010; and the right-wing Gaussian has a mean of 0.676 and a σ of 0.014.

The locations where the residual data and model reach 0 and change sign are taken to define intervals in reflectance value. The spatial distribution of these intervals is shown on the map in Fig. 2.

The left (low-reflectance) wing of the distribution is defined by normalized reflectance values between 0.480 and 0.618 and coincides with the polar region (PR) found in the clustering analysis of MASCS data [3,4].

The central zone (average reflectance) is bound by normalized reflectance values between 0.618 and 0.658 and coincides with the equatorial region (ER) in the clustering analysis.

The right wing (high reflectance) is defined by normalized reflectance values ranging from 0.658 to 0.760. Those values come from an area in the central region of the ER. In the clustering analysis we observed this area to be a stable "core" of the ER even if we apply finer clustering.

Outlook: The data mining and exploration example presented here illustrates a powerful technique to extract hidden information in the MASCS dataset. Moreover, even if the data volume reaches a considerable size, meaningful subsets of the data can be separated for further investigation. In this application we show that some results derived with complex techniques such as unsupervised classification [3,4] can be reproduced by the analysis of the statistical distribution of one parameter in the MASCS dataset. Multi-dimensional analysis could be used to search further for scientifically meaningful relationships in the data. The statistical aggregation of this large dataset also provides a powerful technique for fast data validation, for example, to identify residual instrumental effects.

References: [1] D'Amore, M. et al. (2012), AGU Fall Meeting, abstract P33B-1940; [2] Izenberg, N. et al. (2013), *JGR*, in preparation; [3] Helbert, J. et al. (2013), *JGR*, in preparation; [4] D'Amore, M. et al. (2012), *LPS*, *43*, abstract 1413; [5] D'Amore, M. et al. (2013), *LPS, 44*, this meeting; [6] Helbert, J. et al. (2013), *LPS, 44*, this meeting; [7] Obe, R. O. and L. S. Hsu, *PostGIS in Action* (2011), Manning Publications.