WET CHEMISTRY DATA CLEANING FOR THE PHOENIX MISSION

D. Fang[1], W. Ding[2], E. Oberlin[3], S. Kounaves[3], [1] Predictive Analytics Center of Excellence, Zurich North America, dongping.fang@zurichna.com (dongping.fang@zurichna.com), [2] Department of Computer Science, University of Massachusetts Boston, 100 Morrissey Blvd., Boston, MA02125 (ding@cs.umb.edu), [3]Department of Chemistry, Tufts University, 62 Talbot Avenue, Medford, MA 02155 (elizabeth.oberlin, fsamuel.kounaves@tufts.edu).

Introduction: During the summer of 2008, the Wet Chemistry Laboratory (WCL) on board the Phoenix Lander performed the first comprehensive wet chemical analysis of the soil on Mars [2, 4]. The goal of the WCL was to analyze the chemistry of the soils at the surface and at depth in order to better understand, the history of the water, the biohabitability of the soil, the availability of chemical energy sources, and the general geochemistry of the site. The sensor array included ion selective electrodes (ISE) for K^+, Na^+, Mg^{2+}, Ca^{2+}, NH_4^+, Ba^{2+} (for SO_4^{2-}), Cl^-, Br^-, I^-, NO_3^-/ClO_4^-, H^+(pH), Li^+, and electrodes for conductivity, redox potential, cyclic voltammetry, chronopotentiometry, and an IrO2 pH electrode [1]. WCL data were collected over 17 Martian solar days (sols) during Phoenix 152-sol surface mission days: ~7 hours per day and ~2 seconds apart between measurements resulting over 3 million data points.

The WCL sensor data contains significantly degraded and noisy signals caused by unexpected electronic and thermal noise. So far only a very small portion of the data (< 1%) has been manually cleaned and analyzed (6 man-months) with 2,100 WCL sensor hours, around 3 million data points, still waiting to be interpreted. One of the challenges of de-noising the WCL data is that we cannot reproduce the chemical analyses exactly as what occurred on Mars. Thus it is impossible for us to explicitly identify the instrumental and environmental factors that in terfere with the true WCL sensor data. Previous WCL data analyses have processed the data one ISE measurement at a time [1-4]. The methods used to clean data include Fourier filtering to remove high frequencies [2], Kalman smoothing [3], and for noisier data, the data cleaning process was performed case by case with heavy reliance on human interactions. Still, much of the nosiest data remains uninterpretable. In this paper, we bridge the gap between human expertise and data intrinsic characteristics and propose a new common-factor removal method that utilizes multiple ISE measurements simultaneously to find the hidden shared factors that drive all measurements to vary simultaneously.

Methodology: We propose a new common-factor removal method that utilizes all ISE measurements simultaneously to find the hidden common factors that drive all measurements to vary simultaneously, but not as a result of the chemistry. These common factors represent the errors and variations caused by the combined and complicated influence of varying temperature, pressure, stirring motion, device malfunction, sensor locations, etc. We have cleaned the data by removing the effects of these common factors.

Let K denotes the number of common factors, F_{kt} the k^{th} common factor at time t. The observed data can be modeled as

$$E_t^{(i)} = \mu^{(i)} + \beta_1^{(i)} F_{1t} + \cdots + \beta_K^{(i)} F_{Kt} + \varepsilon_{t,}^{(i)}$$

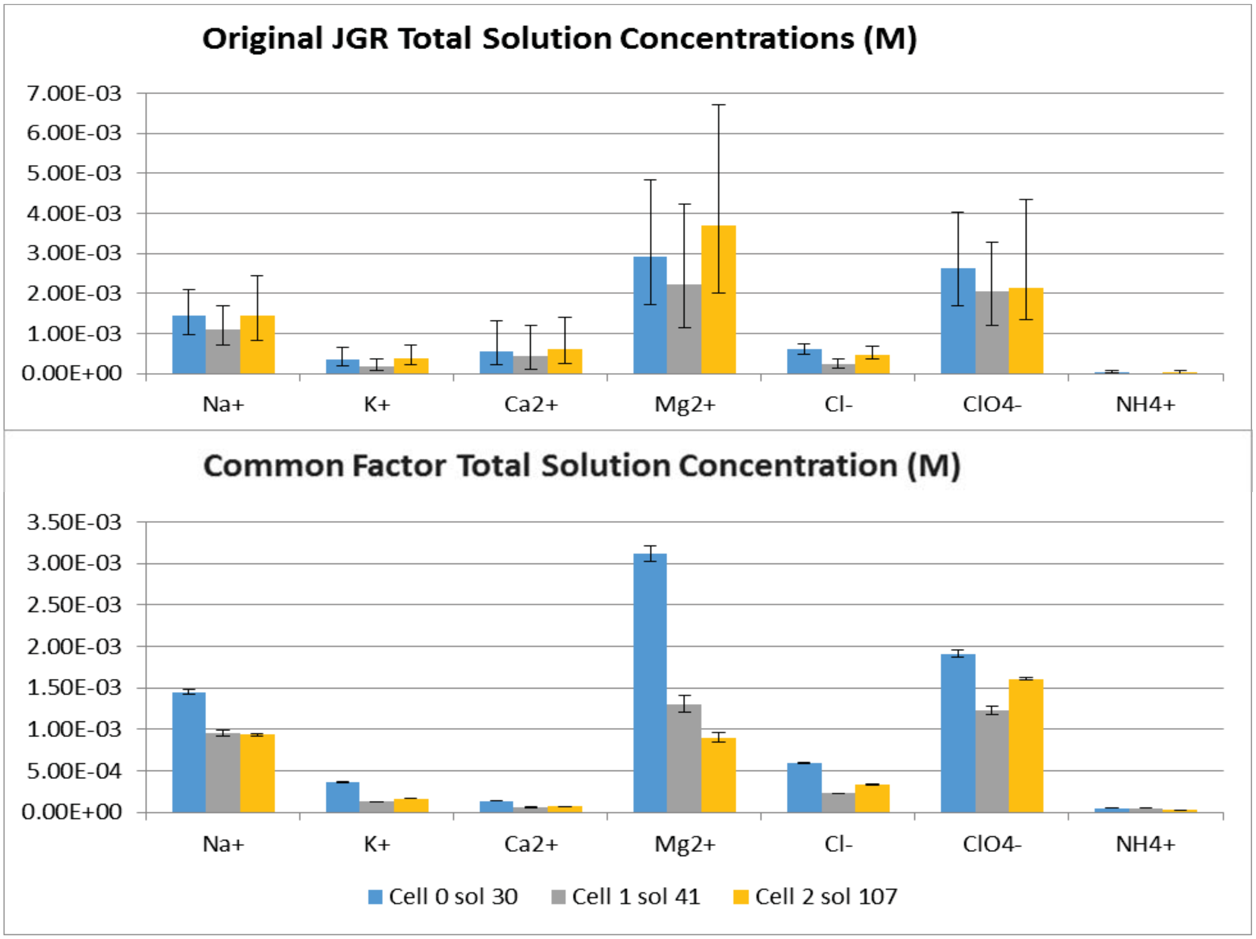
where $\beta_1^{(i)}, \dots, \beta_K^{(i)}$ are the coefficients of the K common factors for ISE signal i, and $\varepsilon_t^{(i)}$ are random noise as in Eq. (2). Notice that the common factors are the same for all ion ISEs, but their influences on each ion may be different due to its different physical properties, and this is reflected in the coefficient $\beta_K^{(i)}$ for that ion. We want to use common factors to help us reduce the variations in the data without changing the base mean level of the data. So we require the base mean of factors to be zero.

The cleaned data to be calculated are

$$E_t^{*(i)} = E_t^{(i)} - \beta_1^{(i)} F_{1t} - \cdots - \beta_K^{(i)} F_{Kt} = \mu^{(i)} + \varepsilon_t^{(i)}$$

We assume that all the ISE sensors mounted on the inside walls of the same beaker that demonstrated similar patterns, were impacted by the same set of common factors. These common factors represent the errors and variations caused by the combined and complicated influence of varying temperature, pressure, stirring motion, device malfunction, sensor locations, etc. We iteratively estimate the common factors of all ISE signals by minimizing the sum of squared errors of all the ions measured by the WCL. We then clean the data by removing the effects of these common factors.

Results: We used the two-common-factor model (K=2) to clean the data in the Calibrant and Sample Intervals. Kounaves et al. in JGR [2] used the mean and error bar of potential in the Calibrant and Sample Intervals to further calculate the total solution concentration and its error bar in their Table 7. The Original interpretation of the WCL chemistry experiments based on this data suggested a uniform distribution of the measured ions within the top 5 cm at the Phoenix landing site. While the mean concentration of ions in solution varied somewhat between samples, the large degree of uncertainty associated with each measurement resulted in an interpretation of uniformity. This



uncertainty in measurement was primarily due to an extremely noisy data set with unknown sources of systematic error. Therefore, the data analysis was often cleaned by hand with 'unlikely' data points being removed and resulting data points being averaged. This process of data cleaning introduced errors of up to 50% due to the inability to confidently discriminate between instrument noise and potentially relevant data.

The application of our common factor algorithm, designed to eliminate these unknown systematic errors in a bias-free way, enabled the reinterpretation of the WCL data to a much higher degree of certainty. By replacing their potential mean and error estimates with our common factor cleaned mean and standard error, we are able to reduce the uncertainty, and therefore increase the validity of the WCL data. Initial differences in the error between our common factor data compared to Kounaves et al. [2] are as follows:

- For Cell 0 sol 30, all differences are small (< 1%).
- For Cell 2 sol 107, when data become nosier, common-factor cleaned means are more than 10% different for K+, NH4+ in both calibrant and sample intervals;
- For Cell 2 sol 107, when data quality keep continues to decreased, common-factor cleaned means are more than 10% different for Na+, K+, NH4+, Mg2+ during their calibrant interval; and K+, Mg2+, NH4+ iduring then sample interval.

Our standard errors are much smaller and do not include common systematic interference.

Acknowledgements: This work was supported by NASA under Grant NNX13AJ69G.

References: [1] Kounaves, S. P.,(2009) J. Geophys. Res.,114, E00A19. [2] Kounaves, S. P.,(2010) J. Geophys. Res.,115, E00E10. [3] Toner, J.D., (2014) Geochimica et Cosmochimica Acta 136, 142-168. [4] Hecht, M. H.,(2009) Science, 325, 64-67.