Math Geosci
DOI 10.1007/s11004-016-9661-y

Relative Abundances of Mineral Species: A Statistical Measure to Characterize Earth-like Planets Based on Earth's Mineralogy

Grethe Hystad[1] · Robert T. Downs[2] · Robert M. Hazen[3] · Joshua J. Golden[2]

Received: 7 December 2015 / Accepted: 15 October 2016

Abstract The mineral frequency distribution of Earth's crust provides a mineralogy-based statistical measure for characterizing Earth-like planets. It has previously been shown that this distribution conforms to a generalized inverse Gauss–Poisson large number of rare events model. However, there is no known analytic expression for the probability distribution of this model; therefore, the population probabilities do not exist in closed forms. Consequently, in this paper, the population probabilities are calculated numerically for all mineral species in Earth's crust, including the predicted undiscovered species. These population probabilities provide an estimate of the occurrence probabilities of species in a random sample of N mineral species–locality pairs. These estimates are used to characterize Earth in terms of its mineralogy. The study demonstrates that Earth is mineralogically unique in the cosmos. In spite of this uniqueness, the frequency distribution of minerals from Earth can be used to quantify the extent to which another planet is Earth-like. Quantitative criteria for characterizing Earth-like planets are given. An example, involving mineral species found on Mars by the CheMin instrument during the Mars Science Laboratory mission suggests that Mars is mineralogically similar to an Earth-like planet.

Keywords Statistical mineralogy · Mineral frequency distribution · Large number of rare events distribution · Mineral ecology · Earth-like planets · Mars mineralogy

✉ Grethe Hystad
ghystad@pnw.edu

1 Mathematics, Statistics, and Computer Science, Purdue University Northwest, 2200 169th Street, Hammond, IN 46323-2094, USA

2 Department of Geosciences, University of Arizona, Tucson, AZ 85721-0077, USA

3 Geophysical Laboratory, Carnegie Institution for Science, Washington, DC 20015, USA

Published online: 09 November 2016

1 Introduction

Humans have long asked questions about their place in the cosmos. How did Earth form? How did life emerge? Are we alone in the universe? What constitutes an Earth-like planet and its requisite life-generating processes is a pervasive theme in planetary science and astrobiology. However, a comprehensive definition of Earth-like planet does not exist, in part because each research group tends to emphasize their particular scientific specialty (Ward and Brownlee 2003; Hystad et al. 2015a). For instance, in the study of (Donahoe 1966), the definition of Earth-like planets was related to the composition of the atmosphere, whereas Seager (2003) emphasized the gravitational relationship between a star and its potential Earth-like planets and the importance of direct detection of planetary radiation. The Kepler spacecraft was launched by NASA in 2009 to search for Earth-size planets that occur near the habitable zone of solar-like stars by monitoring the brightness patterns of the stars during planetary transits (Borucki et al. 2003, 2008). Hystad et al. (2015a) suggested mineralogical criteria for classification as Earth-like, based on the relationship between diversity and distribution of near-surface beryllium mineral species. The distribution of beryllium minerals was chosen because a comprehensive study on these minerals had already been completed and the data were readily available (Grew and Hazen 2014). The probability distribution of Earth's beryllium minerals was found to exist in a closed form, for which simulations could easily be made. The paper of Hystad et al. (2015a) argued that, in spite of deterministic physical, chemical, and biological factors that control most of the planet's mineral diversity, Earth's mineralogy is unique in the cosmos.

The data in the present study consist of a list of mineral species and their localities as of February 2014 from the crowd-sourced web site Mindat.org. There are 135,415 distinct localities and, when counted over all mineral species, these data provide a sample size of 652,856 observations, where each observation is a unique mineral species–locality pair. The mineral species frequency distribution, which records the number of localities for each mineral species, is right skewed with a heavy tail (Hystad et al. 2015b). As of February 2014, there were 4831 approved mineral species reported from Earth's crust, where 22 % of the mineral species are found at only one locality, 12 % are found at only two localities, while more than half of all mineral species are found at five or fewer localities. Hazen et al. (2015a) suggested that the mineral frequency distribution in Earth's near-surface environment, as well as on other terrestrial planets and moons, is a consequence of both deterministic factors and chance events.

Hystad et al. (2015b) introduced a population model for the mineral species frequency distribution in Earth's crust. The mineral species coupled with their localities conforms to a large number of rare events (lnre) distribution since most of Earth's mineral species are rare, known from only a few localities worldwide. Lnre models formulated in terms of a type distribution allow the estimation of Earth's undiscovered mineralogical diversity and the prediction of the percentage of observed mineral species that would differ if Earth's history were replayed. The sample relative frequency of each mineral species tends to overestimate the corresponding population probability because there is no knowledge about the weight of the unobserved por-

tion of the distribution. A parametric model from the family of lnre models that takes into account the unobserved species was used to handle this problem in the study by Hystad et al. (2015b). The mineral frequency spectrum of all mineral species on Earth conforms to the generalized inverse Gauss–Poisson structural type (giGP) lnre model (Hystad et al. 2015b). Most subsets of the mineral species frequency spectrum fit into the giGP lnre model and/or the finite Zipf–Mandelbrot (fZM) lnre model (Hazen et al. 2015b). For example, the frequency spectrum of beryllium mineral species conforms to the fZM lnre model (Hystad et al. 2015a), which carries the advantage of a known analytic form for the probability distribution. The population probabilities can then be found in closed forms. There is no known analytic form reported for the probability distribution of the giGP model (Evert and Baroni 2008).

The objective of this paper is to numerically calculate the relative abundances for all mineral species in Earth's crust, including the undiscovered species. The paper builds on the results given by Hystad et al. (2015b), which included an estimate of the total number of undiscovered mineral species on Earth. Computing these population probabilities provides a statistical measure to characterize Earth-like planets based on their mineralogy. In the study by Hystad et al. (2015a), the argument that Earth is unique in the cosmos was based on a calculation involving the relationship between beryllium-containing minerals to all mineral species. In this paper, a more accurate estimate of the probability of duplicating Earth's mineralogy on an Earth-like planet is provided. Subsequently, Monte Carlo simulations with samples drawn from a multinomial distribution with the population probabilities as the marginal distributions can be made. These bootstraps samples are used to obtain standard errors for the estimated population size of all mineral species, as well as error estimates for the three free parameters of the giGP lnre model that were not included in the paper of Hystad et al. (2015b). Finally, using the population probabilities, samples from the population of mineral species from two Earth-like planets can be simulated to predict the number of mineral species that would be different if Earth's history were to be replayed. The simulation method leads to the same estimate as the method of extrapolating the species accumulation curve that was derived from the giGP lnre model, as given by Hystad et al. (2015b).

A replay of Earth's history requires knowledge of how the mineral frequency distribution changes with time. Hazen and coworkers (Grew and Hazen 2014; Hazen et al. 2008, 2011) demonstrated that the diversity and distribution of minerals on Earth has evolved through a combination of physical, chemical, and biological processes over a period of more than 4.5 billion years. For example, 4.56 billion years ago, the Solar nebula incorporated approximately 60 different mineral species, whereas 3 billion years ago an estimated 1500 mineral species occurred on Earth (Hazen et al. 2008). As a result of co-evolution of the geo- and biospheres, Hystad et al. (2015b) predicted that there are, as of February 2014, 1563 not-yet-observed species in addition to 4831 known species for a total of 6394 mineral species in Earth's crust. Clearly, the characterization of Earth-like planets should be based on the mineralogy of Earth at any time in its evolution. Unfortunately, the important variable of time cannot be addressed in this paper until age data are available for all mineral species.

2 A Brief Review of Mineral Species as LNRE Distributions

This section follows Baayen (2001) and Hystad et al. (2015b) closely. Let S denote the population size of distinct mineral species in Earth's crust and denote the kth mineral species by x_k for $k = 1, 2, \ldots, S$. Assume each mineral species x_k has a population probability π_k (relative abundance) of being sampled at an arbitrary locality, where $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_S$ defines the ordering schemes and $\sum_{k=1}^{S} \pi_k = 1$. The probability of a given mineral species being found is assumed to be constant over all localities. Let a sample of N mineral species–locality pairs be drawn randomly and independently from the total population of occurrences of S mineral species. Thus, the sample size N is the sum over all localities of the number of mineral species at each locality that has been sampled. Let $f_k(N)$ denote the frequency, which is the number of distinct localities for the kth mineral species x_k in the sample of N mineral species–locality pairs. Then $f_N = (f_1(N), f_2(N), \ldots, f_S(N))$ follows a multinomial distribution, where the marginal distribution of each frequency is binomial with N trials and success probability π_k. The probability that the kth mineral species x_k is found at exactly m localities is given by

$$P(f_k(N) = m) = \binom{N}{m} \pi_k^m (1 - \pi_k)^{N-m} \approx \frac{(N\pi_k)^m}{m!} \exp(-N\pi_k), \tag{1}$$

where the binomial probabilities can be approximated with the Poisson probabilities with mean $N\pi_k$ because N is large and π_k is small for all k. Denote the number of distinct mineral species in a sample of N species–locality pairs by $V(N)$ and the number of distinct mineral species with exactly m localities by $V_m(N)$. The sequence $(V_1(N), V_2(N), \ldots, V_{V(N)}(N))$ is called the observed frequency spectrum. As of February 2014, the total number of observed mineral species is $V(N) = 4831$, while for example, the number of distinct mineral species found at only one and two localities are $V_1(N) = 1062$ and $V_2(N) = 569$ for $N = 652{,}856$. Using Eq. (1), expected values of $V_m(N)$ and $V(N)$ are given by

$$E(V_m(N)) = \sum_{k=1}^{S} \frac{(N\pi_k)^m}{m!} \exp(-N\pi_k), \tag{2}$$

and

$$E(V(N)) = \sum_{k=1}^{S} (1 - \exp(-N\pi_k)), \tag{3}$$

respectively (Baayen 2001).

Let $I_{[\pi_k \geq \rho]}$ be the indicator function, which is 1 if $\pi_k \geq \rho$ and 0 otherwise. The structural type distribution is defined by

$$G(\rho) = \sum_{k=1}^{S} I_{[\pi_k \geq \rho]}, \tag{4}$$

which is the number of mineral species in the population that have probability greater than or equal to ρ (Baayen 2001). The structural type distribution $G(\rho)$ will be approximated by a continuous function (Evert 2004)

$$G(\rho) = \int_{\rho}^{\infty} g(\pi)\mathrm{d}\pi, \tag{5}$$

where $g(\pi)$ is a type density function that satisfies $g \geq 0$ and

$$\int_{0}^{\infty} \pi g(\pi)\mathrm{d}\pi = 1. \tag{6}$$

The population size is given by $S = \int_0^{\infty} g(\pi)\mathrm{d}\pi$. Since G is of bounded variation, the expressions in Eqs. (2) and (3) can be written in terms of the Stieltjes integrals (Baayen 2001)

$$E(V_m(N)) = \int_{0}^{\infty} \frac{(N\pi)^m}{m!} \exp(-N\pi) g(\pi)\mathrm{d}\pi, \tag{7}$$

and

$$E(V(N)) = \int_{0}^{\infty} (1 - \exp(-N\pi)) g(\pi)\mathrm{d}\pi. \tag{8}$$

Note that the marginal probability density function, $E(V_m(N))/S$, of m is an overdispersed Poisson count distribution with mixing density $g(\pi)/S$. Note also that the integrals in Eqs. (5) and (6) were extended to infinity but the integral over [0, 1] will lead to identical results. The use of the larger interval is the convention in the literature, because a change of variable in Eqs. (7) and (8) leads to considering the parameter of a Poisson-distributed random variable as a rate of the number of occurrence for a particular mineral species (Baayen 2001).

The generalized inverse Gauss–Poisson (giGP) structural type distribution was used as a model for $G(\rho)$ in Hystad et al. (2015b). This model was introduced by Sichel (1971, 1975, 1986), where the type density function is given by

$$g(\pi) = \frac{\left(\frac{2}{bc}\right)^{\gamma+1}}{2K_{\gamma+1}(b)} \pi^{\gamma-1} \exp\left(-\frac{\pi}{c} - \frac{b^2 c}{4\pi}\right), \tag{9}$$

with parameters in the range $-1 < \gamma < 0$, $b \geq 0$, and $c \geq 0$, and where $K_{\gamma}(b)$ is the modified Bessel function of the second kind of order γ and argument b (Baayen 2001). The closed forms for $E(V_m(N))$ and $E(V(N))$ resulting from integration of Eqs. (7) and (8) are given in Baayen (2001).

The giGP lnre model was found to fit well to the frequency spectrum of all mineral species by Hystad et al. (2015b), where the total number of mineral species in the population was estimated (with standard error) to be $S = 6394$ (111). This number is an underestimate, as explained in Hystad et al. (2015b). The R-package ZipfR (Evert

and Baroni 2007, 2008) was employed to fit the model, where the parameters were estimated by minimizing, through the Nelder–Mead algorithm, the simplified version of the multivariate Chi-squared test for goodness-of-fit using the first 11 spectrum elements. The parameters (with standard errors) are $\gamma = -0.42$ (0.01), $b = 0.0132$ (0.0003), and $c = 0.014$ (0.001), with $\chi^2 = 10.36$, $df = 13$, and p value = 0.66 (Hystad et al. 2015b).

3 Relative Abundances of Mineral Species in Earth's Crust

In this section, the relative abundances (population probabilities) for all mineral species in Earth's crust are calculated numerically. Knowing the population probabilities will allow the simulation of samples taken from the estimated population of mineral species. Furthermore, it enables the calculation of standard errors of the estimates of the parameters in the giGP lnre model and for the population size S using multinomial bootstrap sampling.

Recall from the previous section that the number of mineral species in the population with probably greater than or equal to ρ is approximated by the continuous function

$$G(\rho) = \int_{\rho}^{1} g(\pi)\mathrm{d}\pi, \tag{10}$$

where $g(\pi)$ is the type density function given in Eq. (9). Here, the upper limit of the integral was restricted to 1 since the type probabilities are in the interval $0 < \pi < 1$. The structural type distribution given in Eq. (4) is a step function with $G(\pi_k) = k$ since there are k mineral species with probability greater than or equal to π_k, namely $x_1, x_2, \ldots, x_k$ (Evert 2004). Using the continuous approximation to G in Eq. (10), the equation of interest is $G(\rho_k) = k$, where the population probabilities π_k asymptotically are equal to ρ_k for $k = 1, 2, \ldots, 6394 = S$.

Define the function $f : (0, 1) \to \mathbb{R}$, where $f(\rho_k) = G(\rho_k) - k$. Notice that $f'(\rho_k) = G'(\rho_k) = -g(\rho_k)$. To solve the equation $f(\rho_k) = 0$, the Newton–Raphson method is employed. Let $\rho_k^{(l)}$ denote the lth iterated value of ρ_k in the Newton–Raphson algorithm. Then

$$\rho_k^{(l+1)} = \rho_k^{(l)} + \frac{G(\rho_k^{(l)}) - k}{g(\rho_k^{(l)})},$$

for $l = 0, 1, 2, \ldots, L$, where $\rho_k = \rho_k^{(L+1)}$ for some positive integer L.

Define the interval $I = [\rho_k - \epsilon, \rho_k + \epsilon]$ for some $\epsilon \geq |\rho_k - \rho_k^{(0)}|$ with $\rho_k > \epsilon > 0$, where $\rho_k^{(0)}$ is the initial value in the Newton–Raphson algorithm. It is easy to check that $f'(\rho) = -g(\rho)$ is different from zero and $f''(\rho) = -g'(\rho)$ is continuous and bounded for ρ in I. The initial value $\rho_k^{(0)}$ can be chosen sufficiently close to ρ_k such that the convergence is quadratic (Ryaben'kii and Tsynkov 2006) (in this paper, the bisection method was used to find the initial values $\rho_k^{(0)}$). Using Eq. (6), define for $0 \leq \rho \leq 1$

$$F(\rho) = \int_{\rho}^{1} \pi g(\pi) \mathrm{d}\pi.$$

Then the population probabilities π_k are given by

$$\pi_1 = F(\rho_1) = \int_{\rho_1}^{1} \pi g(\pi) \mathrm{d}\pi, \tag{11}$$

$$\pi_k = F(\rho_k) - F(\rho_{k-1}) = \int_{\rho_k}^{\rho_{k-1}} \pi g(\pi) \mathrm{d}\pi, \tag{12}$$

for $k = 2, 3, \ldots, 6394 = S$. The values of the integrals were found numerically using the integrate function in *R*. Figure 1 shows the graph of the population probabilities as a function of rank, where the species with the highest probability is ranked number 1. The mineral species with rank number 1 has a population probability of 0.0423; that is, if a mineral species is randomly sampled at an arbitrary locality, the probability of observing species ranked number 1 is 0.0423. Keep in mind that the observed ranking of the 4831 mineral species is different from the population ranking of the 6394 mineral species. However, it is likely that, for example, quartz, which is ranked number 1 in the observed ranking, is also ranked number 1 in the population ranking. Notice that the sample proportion of quartz of $44{,}973/652{,}856 = 0.0689$ is an overestimate of the population probability of 0.0423, as expected, since the calculation of the sample proportion did not take into account the unobserved species in the population. The mineral species with the highest ranking (ranking number 6394) has population probability of 8.94×10^{-8}.

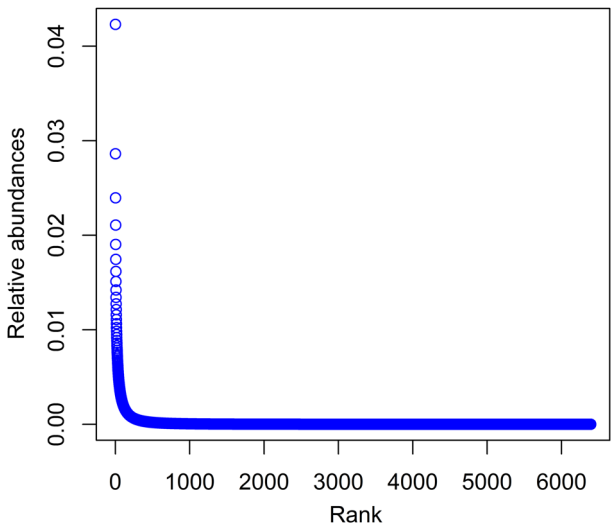

Fig. 1 Population probabilities versus rank for the 6394 mineral species in Earth's crust

4 Differences in the Number of Mineral Species from Two Modeled Earth-like Planets

In the study by Hystad et al. (2015b), the expected number of mineral species that will be different in two random samples of the same size from two modeled Earth-like planets was estimated by extrapolating from the expected species accumulation curve using the giGP lnre model from size N to $2N$. The resulting value was multiplied by 2 to estimate the number of different mineral species distributed over the two samples. Because the population probabilities for finding mineral species at an arbitrary locality are now estimated, the differences between randomly generated samples from the estimated population of mineral species can be modeled. One measure of the difference between two randomly generated samples is the number of minerals in one sample that are not in the other. Thus, two random samples of size $N = 652,856$ are generated from the multinomial distribution with the probabilities computed in Sect. 3 as the marginal distributions. Two thousand pairs of simulations, called A and B, are generated. The mineral species of simulation A are compared with those of simulation B; the difference is defined as the number of species in A that are not in B plus the number of species in B that are not in A; that is, the number of species that are not common to A and B or the symmetric difference of A and B. The average value of the differences between these two samples is then calculated. This value is the estimated true value of the expected number of different species in two random samples from the population of mineral species. This number was found to be 1324 with a standard error of 28, which is the same number as computed by Hystad et al. (2015b) using extrapolation from the species accumulation curve. The standard errors given in Sect. 2 of the estimated parameters in the giGP lnre model, as well as for the estimated population size S, are also computed using the 2000 bootstrap samples from the multinomial distribution.

5 How to Characterize Earth-like Planets

In this section, probabilities for the occurrence of each of the 6394 mineral species in a random sample of $N = 652,856$ mineral species–locality pairs from Earth's crust are estimated using the relative abundances outlined in Sect. 3. Subsequently, the number of mineral species that should be present in a random sample of size N mineral species–locality pairs is determined for varying values of N. These results provide a statistical measure to characterize Earth-like planets in terms of their mineralogy using a snapshot of the surface of today's Earth. Finally, it will be shown that even though Earth is mineralogically unique in the Universe, there are mineralogical criteria that can be used to quantify how Earth-like a given planet might be.

Using the binomial distribution in Eq. (1) and the estimated relative abundances π_k outlined in Eqs. (11) and (12), the probabilities were calculated that each of the mineral species in the population will occur at least once in a random sample of $N = 652,856$ mineral species–locality pairs from Earth's crust. The results are illustrated in Fig. 2. Rounded to three decimal places, there are 1962 mineral species with greater than 0.999 probability of occurring on all Earth-like planets, whereas 2908 mineral species have at least a 0.950 probability of occurrence. These mineral species represent the

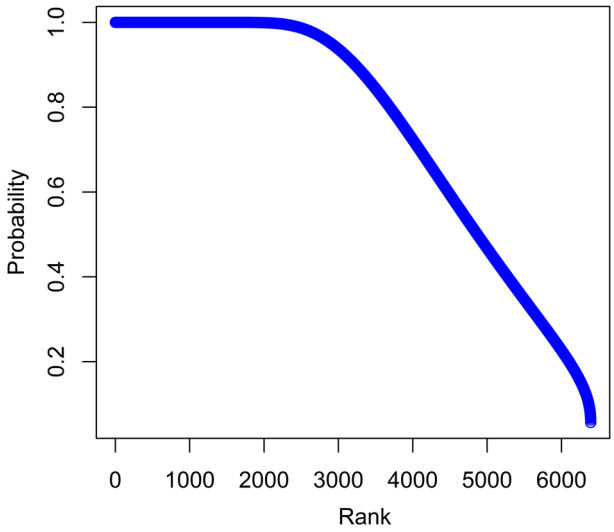

Fig. 2 The probabilities that mineral species are present at least one time in a random sample of $N = 652{,}856$ mineral species–locality pairs from Earth's crust versus population rank

suite of species that is most likely to be observed on Earth today. Furthermore, 4870 mineral species have at least 0.500 probability of occurrence. The mineral species with low probability of occurrence represent species that occur by chance; for example, 27 mineral species have less than 0.100 probability of occurrence. Thus, if a random sample of $N = 652{,}856$ mineral species–locality pairs is taken from an extraterrestrial planet, it should contain at least 1962 out of the 4831 suite of minerals species observed in Earth's crust today to be considered mineralogically Earth-like. These observed mineral species are likely to be among the ones with the highest observed frequencies on Earth, such as, quartz, pyrite, and albite.

Shen et al. (2003) made the assumption that species with equal frequencies in the sample also have equal relative abundances in the population for predicting the number of new species in further taxonomic sampling. It is reasonable to assume, in particular for the observed high-frequency species, that if a species A has lower observed ranking than species B, species A will also have lower ranking than species B in the population. In that case, the 1962 observed species with the highest frequencies will be the species that will characterize an Earth-like planet, based on a snapshot of today's Earth.

To obtain an upper bound for the probability of duplicating exactly all of the 4831 mineral species observed on Earth on another Earth-like planet, the occurrence probabilities illustrated in Fig. 2 were multiplied for the first 4831 ranked mineral species in the population. The resulting probability of discovering the exact same 4831 mineral species on another Earth-like planet is thus less than 1.4×10^{-263}. According to Williams et al. (1996), there may be 10^{22} terrestrial (if not Earth-like) worlds. Therefore, it can be concluded that Earth is mineralogically unique in the visible universe. Nevertheless, many planets may be mineralogically Earth-like if they satisfy mineralogical criteria based on Earth's mineralogy. To identify an Earth-like planet from the

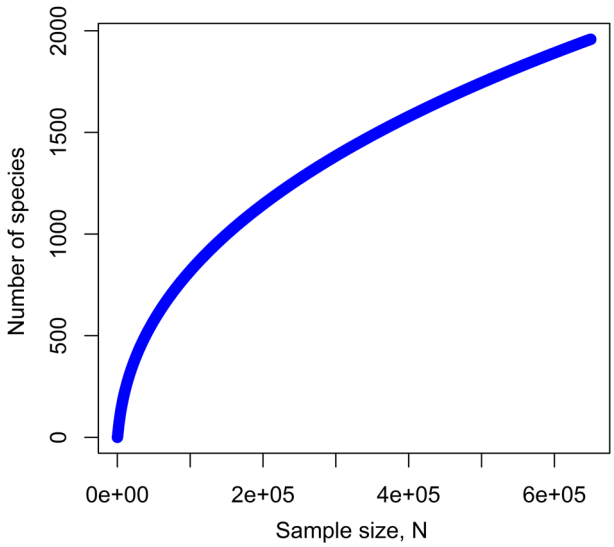

Fig. 3 Number of mineral species with occurrence probability greater than 0.999 versus sample size N on an Earth-like planet based on the mineralogy of today's Earth

perspective of mineralogy, it is crucial to consider the sample size N (number of mineral species–locality pairs) for which the species were collected. It would be difficult to obtain a sample size as large as the current sample size from Earth of $N = 652{,}856$ mineral species–locality pairs on another world.

However, using the statistical model of the mineral frequency distribution and its species accumulation curve on today's Earth, the number of mineral species that should be observed on an potential Earth-like planet as a function of sampling size can be calculated. Figure 3 illustrates the number of distinct observed mineral species that are present in a random sample of N mineral species–locality pairs, where N varies from 100 to 650,000 in increments of 100. Thus, at a fixed sampling size N, the number of distinct mineral species that should be present in a random sample can be determined. Figure 4 illustrates the number of mineral species that have a probability of greater than or equal to 0.500, 0.700, 0.950, and 0.999 of being observed in a sample from an Earth-like planet for sampling sizes ranging from 10 to 300. For example, if a random sample of 300 mineral species–locality pairs is taken from an Earth-like planet based on the mineralogy of today's Earth, three distinct mineral species will have a probability of at least 0.999 and 16 distinct species will have a probability of at least 0.950 of being observed in the sample.

5.1 How to Determine if an Extraterrestrial Planet is Earth-like in Terms of Mineralogy

Suppose a sample of size $N = M$ mineral species–locality pairs is taken from an extraterrestrial planet of which there are v observed distinct species. Suppose

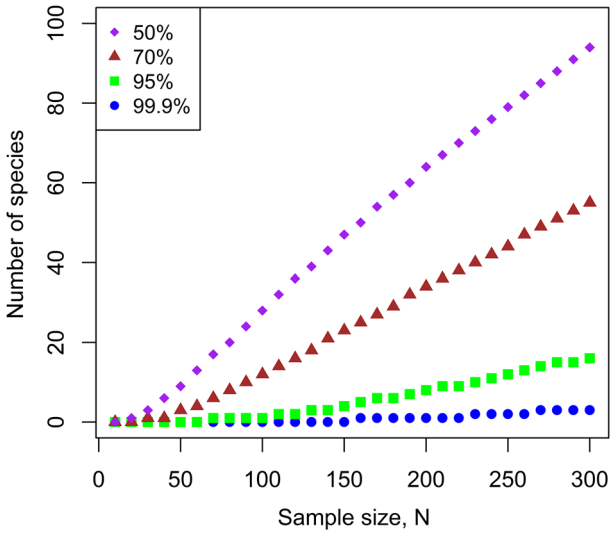

Fig. 4 Number of mineral species with occurrence probability greater than or equal to 0.500, 0.700, 0.950, and 0.999 versus sample size N, for N ranging from 10 to 300 on an Earth-like planet based on the mineralogy of today's Earth

all of the v species are also observed on Earth. In this section, a method will be described to determine if such an extraterrestrial planet can be characterized as Earth-like.

The mineral species observed on an extraterrestrial planet will be assumed to be among the most common species observed on Earth as discussed in Sect. 5. The population rankings of these common species on Earth are assumed to be the same as the observed rankings as discussed earlier. The marginal frequency distribution of each species x_i on Earth in a sample of M species–locality pairs follows the binomial distribution with mean $M\pi_i$ and variance $M\pi_i(1-\pi_i)$, where π_i is the population probability described in Sect. 3. For $i = 1, 2, \ldots, v$, let C_i denote a confidence statement about the value of the frequency for the ith mineral species on Earth with $P(C_i \text{ is true}) = 1-\alpha/v$ and $0 \le \alpha \le 1$. Simultaneous confidence intervals using the Bonferroni method are developed such that $P(C_i \text{ is true for all } i) \ge 1 - \sum_{i=1}^{v}(\alpha/v) = 1-\alpha$ (Johnson and Wichern 2007). Therefore, with an overall confidence greater than or equal to $1-\alpha$, we have for each, $i = 1, 2, \ldots, v$, the intervals $M\pi_i \pm z_{1-\alpha/2v}\sqrt{M\pi_i(1-\pi_i)}$, where $z_{1-\alpha/2v}$ is the critical value on the standard normal distribution with $P(Z \ge z_{1-\alpha/2v}) = \frac{\alpha}{2v}$. The normal approximation to the binomial distribution is here applied for sufficiently large sample size. As an example, 95 % confidence intervals for the $v = 100$ most frequent species in a sample of size $N = M = 3000$ observations are illustrated in Fig. 5. The middle curve represents the expected frequencies for each of the v species at sample size $N = 3000$, while the lower and upper curves represent the boundary of the 95 % confidence intervals for all 100 mineral species. Note that we can also compute the $(1-\alpha) \times 100$ % confidence intervals as: $M\pi_i \pm q_{1-\alpha/2}\sqrt{M\pi_i(1-\pi_i)}$, where $q_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of

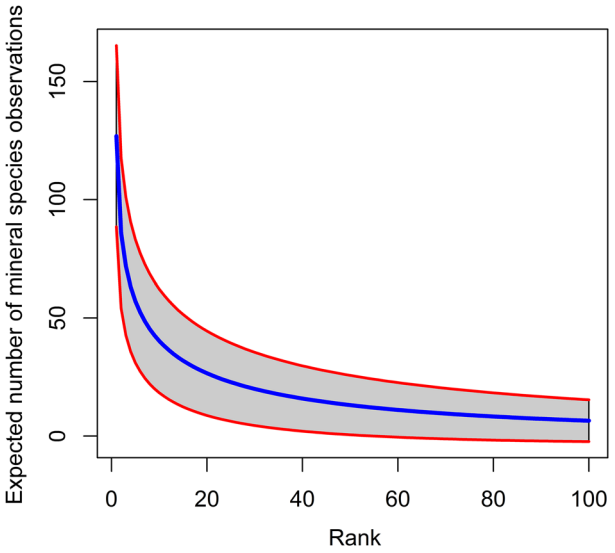

Fig. 5 Expected frequencies and 95 % simultaneous confidence intervals for the frequencies of the 100 species that are ranked 1–100 in a sample of $N = 3000$ observations from Earth

$$\max_{1\le i\le v} \frac{f_i(M) - M\pi_i}{\sqrt{M\pi_i(1-\pi_i)}}, \tag{13}$$

where $f_i(M)$ is the frequency of the ith species at sample size M. A random variable is simulated from the normal distribution with mean 0 and standard deviation $\sqrt{M\pi_i(1-\pi_i)}$, and the value of Eq. (13) is computed. The method is repeated many times. In this paper, $K = 10{,}000$ was used. The value of $q_{1-\alpha/2}$ is the $(1-\alpha/2)K$th element of the K simulated values, ordered from the smallest to the largest. Using this method for the critical value will give about the same result as using Bonferroni confidence intervals. To test whether an extraterrestrial planet is Earth-like, the following null hypothesis is formed.

H_0 The mineral species frequency for each species from an extraterrestrial planet in a sample of size M species–locality pairs is the same as the frequency of the corresponding mineral species x_i on Earth in a sample of size M.

If the frequencies of one or more mineral species of an extraterrestrial planet fail to be contained in the $(1-\alpha)\times 100$ % confidence band, the null hypothesis that the planet is mineralogically Earth-like, is rejected at the α significance level.

If all of the mineral species from an extraterrestrial planet have frequencies contained in the $(1-\alpha)\times 100$ % confidence band, the conclusion is that there are no significant differences at the α significance level between these mineral frequencies and the corresponding frequencies on Earth. It still does not mean that the planet is Earth-like; rather there is not sufficient evidence to reject the hypothesis that it could be Earth-like at the α significance level. In these calculations, the snapshot of the

current Earth is used to obtain the confidence intervals. Note that when data on Earth's mineral diversity through deep time are organized we may have information about the abundances of mineral species in its earlier time.

Baayen (2001) introduced the concept of an lnre-zone, which is the range of values of the sample size N for which the expected species accumulation curve is still increasing, while the expected number of rare species is non-negligible. The distribution of the most common mineral species found on Earth is located in the late lnre-zone or even outside the lnre-zone, where there are no species with only a few localities as well as no predicted new species. The mineral frequency distributions on an extraterrestrial planet with no bio-mineral species present will likely be located in the late lnre-zone or outside the zone.

5.2 Is Mars Earth-like?

On Mars, 33 mineral species have been found by the CheMin instrument during the Mars Science Laboratory mission as of July 2015 (Downs 2015). All of the species found on Mars appear in Earth's crust; furthermore, 32 of Mars' mineral species are among the first 1713 ranked species observed on Earth, while one of the species is ranked number 3643 on Earth. Furthermore, almost half of the species found on Mars are among the 100 most common observed species on Earth. Recall that there are 1962 mineral species on Earth that have greater than 99.9 % chance of being observed in a sample of $N = 652{,}856$ observations. To determine if the mineral species found on Mars are in coincidence with the ranking of species from an Earth-like planet, the following simulation was performed. A random sample of 33 mineral species with no replacement is generated from the population of mineral species on Earth. The experiment is repeated 10^5 times and the number of species that are ranked 1963 or higher are counted. The result is provided in Fig. 6. The histogram shows that 30 % of the samples from Earth that contain 33 mineral species have exactly one species with rank greater than 1962. In comparison, Mars has one species, sanderite, which has rank greater than 1962 on Earth. Even though one cannot say for sure which species among the 4831 observed species on Earth have a population ranking from 1 to 1962, it is highly likely that the observed ranking is approximately equal to the population ranking for these common mineral species. Thus, 32 of the 33 mineral species found on Mars are among the species that are likely to be present on an Earth-like planet. It can be concluded that, on the continuous scale of what can be termed Earth-likeness, Mars is mineralogically Earth-like.

Figure 7 illustrates the 2.5th and 97.5th percentiles for the number of species that have rank greater than 1962 in a sample of v mineral species from Earth. The simulation was run with 10^5 samples of v species, each sampled with no replacement from the population of mineral species on Earth. For example, if a sample from an extraterrestrial planet has 300 mineral species, the 97.5th percentile for the number of species with rank greater than 1962 is 16, that is, if the number of mineral species with rank greater than 1962 is larger than 16 in a sample of 300 species, the planet is not likely Earth-like.

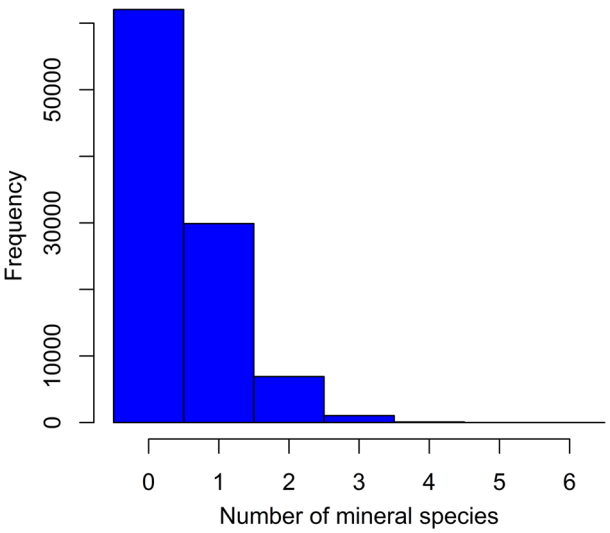

Fig. 6 The distribution of species with ranking greater than 1962 in a sample of 33 mineral species

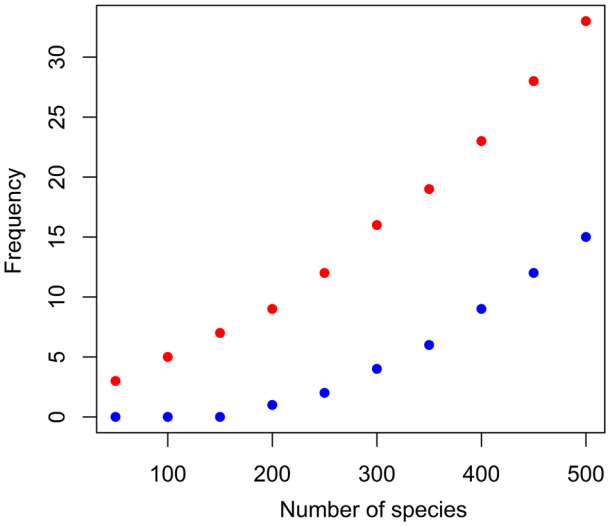

Fig. 7 Upper curve is the 97.5th percentile and the lower curve is the 2.5th percentile for the number of species with rank greater than 1962 in a sample of v mineral species selected with no replacement

6 Conclusions

In this paper, calculations of the relative abundances for all of Earth's crustal minerals, including the undiscovered species, were provided. These results led to a statistical measure on how to characterize Earth-like planets based on the mineralogy of today's Earth. In addition, it was demonstrated that, while Earth is mineralogically unique in

the known cosmos, mineralogical criteria can be used to quantify the extent to which a planet is Earth-like.

In the near future, the data for time of formation, diversity, and distribution of mineral species in Earth's history will be gathered such that characterization of an Earth-like planet in terms of mineralogy can be extended to include any time in its evolution.

Acknowledgements Two anonymous reviewers provided detailed and valuable suggestions for improving the manuscript. We received critical advice and data from Edward Grew. We gratefully acknowledge support from NASA Mars Science Laboratory Mission NNX11AP82A, as well as support from the Alfred P. Sloan Foundation (Grant Number 2013-10-01), the W. M. Keck Foundation (Grant Number 140002372), the Deep Carbon Observatory, the Carnegie Institution for Science, and an anonymous private foundation.

References

Baayen RH (2001) Word frequency distributions, text, speech and language technology, vol 18. Kluwer Academic Publishers, Dordrecht

Borucki W, Koch D, Batalha N, Caldwell D, Christensen-Dalsgaard J, Cochran WD, Dunham E, Gautier TN, Geary J, Gilliland R, Jenkins J, Kjeldsen H, Lissauer JJ, Rowe J (2008) Kepler: search for Earth-size planets in the habitable zone. Proc Int Astron Union 4:289–299

Borucki WJ, Koch DG, Basri GB, Caldwell DA, Caldwell JF, Cochran WD, DeVore E, Dunham EW, Geary JC, Gilliland RL, Gould A, Jenkins JM, Kondo Y, Latham DW, Lissauer JJ (2003) The Kepler mission: finding the sizes, orbits and frequencies of Earth-size and larger extrasolar planets. In: Deming D, Seager S (eds) Scientific frontiers in research on extrasolar planets, ASP conference series, vol 294, pp 427–440

Donahoe FJ (1966) On the abundance of Earth-like planets. Icarus 5:303–304

Downs RT (2015) Determining mineralogy on Mars with the CheMin X-ray diffractometer. Elements 11:45–50

Evert S (2004) A simple LNRE model for random character sequences. In: Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles, Louvain-la-Neuve, Belgium, pp 411–422

Evert S, Baroni M (2007) zipfR: word frequency distributions in R. In: Proceedings of the 45th annual meeting of the association for computational linguistics, Prague, Czech Republic, Posters and Demonstrations Session, pp 29–32

Evert S, Baroni M (2008) Statistical models for word frequency distributions, package zipfR. http://zipfr.r-forge.r-project.org/materials/zipfR_0.6-5.pdf. Accessed 1 June 2015

Grew ES, Hazen RM (2014) Beryllium mineral evolution. Am Miner 99:999–1021

Hazen RM, Papineau D, Bleeker W, Downs RT, Ferry JM, McCoy TJ, Sverjensky DA, Yang H (2008) Mineral evolution. Am Miner 93:1693–1720

Hazen RM, Bekker A, Bish DL, Bleeker W, Downs RT, Farquhar J, Ferry JM, Grew ES, Knoll AH, Papineau D, Ralph JP, Sverjensky DA, Valley JW (2011) Needs and opportunities in mineral evolution research. Am Miner 96:953–963

Hazen RM, Grew ES, Downs RT, Golden JJ, Hystad G (2015a) Mineral ecology: chance and necessity in the mineral diversity of terrestrial planets. Can Miner 00:1–29

Hazen RM, Hystad G, Downs RT, Golden JJ, Pires AJ, Grew ES (2015b) Earth's 'missing' minerals. Am Miner 100:2344–2347

Hystad G, Downs RT, Grew ES, Hazen RM (2015a) Statistical analysis of mineral diversity and distribution: Earth's mineralogy is unique. Earth Planet Sci Lett 426:154–157

Hystad G, Downs RT, Hazen RM (2015b) Mineral species fequency distribution conforms to a large number of rare events model: prediction of Earth's missing minerals. Math Geosci 47:647–661

Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis, 6th edn. Pearson, Upper Saddle River

Ryaben'kii VS, Tsynkov SV (2006) A theoretical introduction to numerical analysis. Chapman & Hall/CRC, Boca Raton

Seager S (2003) The search for extrasolar Earth-like planets. Earth Planet Sci Lett 208:113–124

Shen TJ, Chao A, Lin CF (2003) Predicting the number of new species in further taxonomic sampling. Ecology 84(3):798–804

Sichel HS (1971) On a family of discrete distributions particularly suited to represent long-tailed frequency data. In: Proceedings of the third symposium on mathematical statistics, Pretoria, South Africa, pp 51–97

Sichel HS (1975) On a distribution law for word frequencies. J Am Stat Assoc 70:542–547

Sichel HS (1986) Word frequency distributions and type-token characteristics. Math Sci 11:45–72

Ward PD, Brownlee D (2003) Rare Earth: why complex life is uncommon in the universe. Copernicus, New York

Williams RE, Blacker B, Dickinson M, Dixon WVD, Ferguson HC, Fruchter AS, Giavalisco M, Gilliland RL, Heyer I, Katsanis R, Levay Z, Lucas RA, McElroy DB, Petro L, Postman M, Adorf HM, Hook R (1996) The Hubble deep field: observations, data reduction, and galaxy photometry. Astron J 112:1335