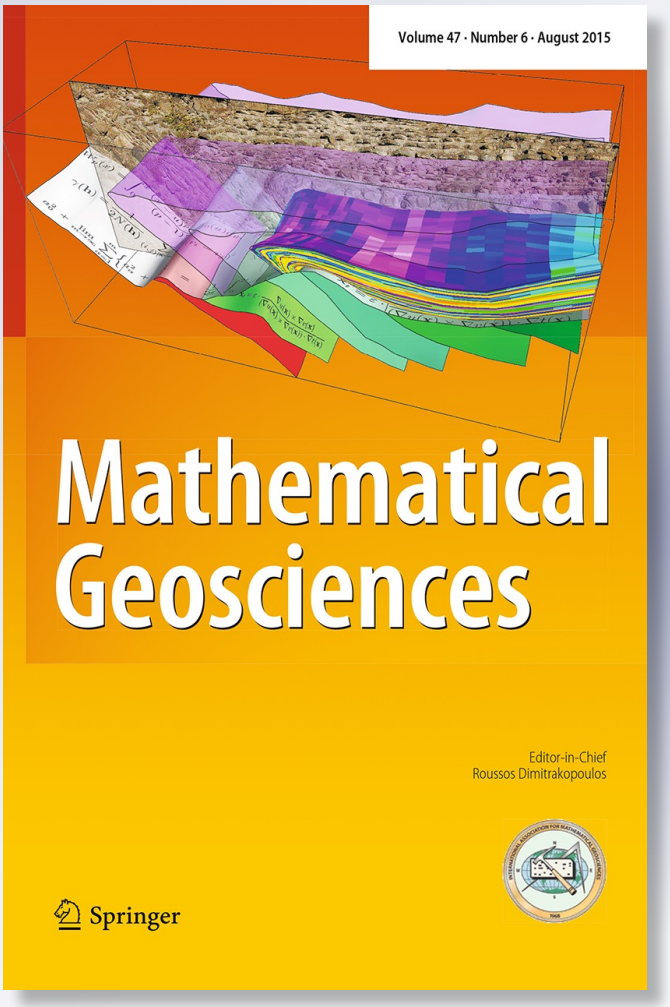
Mineral Species Frequency Distribution Conforms to a Large Number of Rare Events Model: Prediction of Earth's Missing Minerals

Grethe Hystad, Robert T. Downs & Robert M. Hazen

Mathematical Geosciences

ISSN 1874-8961
Volume 47
Number 6

Math Geosci (2015) 47:647-661
DOI 10.1007/s11004-015-9600-3



Springer

Math Geosci (2015) 47:647–661
DOI 10.1007/s11004-015-9600-3

Mineral Species Frequency Distribution Conforms to a Large Number of Rare Events Model: Prediction of Earth's Missing Minerals

Grethe Hystad[1] · **Robert T. Downs**[2] · **Robert M. Hazen**[3]

Received: 16 December 2014 / Accepted: 18 May 2015 / Published online: 11 June 2015

Abstract A population model is introduced to describe the mineral species frequency distribution. Mineral species coupled with their localities conform to a large number of rare events (LNRE) distribution: 100 common mineral species occur at more than 1,000 localities, whereas 34 % of the approved 4,831 mineral species are found at only one or two localities. LNRE models formulated in terms of a structural 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.

Keywords Statistical mineralogy · Mineral ecology · Mineral frequency distribution

1 Introduction

The search for predictive statistical models of natural systems represents an ongoing opportunity and challenge in applied mathematics. Here the extensive and growing data resources on mineral species and their localities are employed to identify and parameterize frequency distributions of Earth's mineral kingdom. These models, for the first time, facilitate prediction of Earth's total, but as yet undiscovered, mineralogical diversity.

✉ Grethe Hystad
ghystad@math.arizona.edu

1 Department of Mathematics, University of Arizona, 617 N. Santa Rita Ave., P.O. Box 210089, Tucson, AZ 85721-0089, USA

2 Department of Geosciences, University of Arizona, 1040 E 4th Street, Tucson, AZ 85721-0077, USA

3 Geophysical Laboratory, Carnegie Institution, 5251 Broad Branch Road NW, Washington, DC 20015, USA

The frequency distribution of mineral species in Earth's near-surface environment, as well as on other terrestrial planets and moons, arises from both deterministic factors and chance events (Hazen et al. 2015). Mineral diversity is defined as the number of different mineral species, each of which has a unique combination of crystal structure and chemical composition as approved by the International Mineralogical Association (http://rruff.info/ima). Deterministic factors in mineral diversity are illustrated by the positive correlation between the abundances of crustal elements and the numbers of mineral species containing those elements (Hazen et al. 2015). The most common and volumetrically significant minerals, which form from the most abundant crustal chemical elements (including oxygen, silicon, iron, magnesium, aluminum, and calcium), are known as the rock-forming minerals. Their role in the frequency distribution of terrestrial planets and moons is a necessity in mineral evolution, and thus potential Earth-like planets in other star systems are expected to show similar distributions of these common rock-forming minerals. However, chance also plays an important role in the diversity of mineral species, in particular for rare minerals found at only one or two localities.

As of February 2014, 4,831 mineral species have been identified and reported on Earth. The characteristic mineral species frequency distribution, which records the number of localities for each mineral species, is right skewed with a heavy tail, as illustrated in Fig. 1. The mineral species with the highest frequency (quartz with approximately 45,000 reported localities) is ranked number 1, while 100 mineral species (2 %) have been reported to occur at 1,000 or more localities. By contrast, 22 % of mineral species have been found at only one locality and 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. (2015) proposed that the frequency distribution of mineral species with rare occurrence indicates that many more mineral species are yet to be discovered, have occurred in the past but are lacking on Earth today, or never formed owing to chance events.

The data of this study consist of the list of localities, as well as the list of mineral species found at those localities, as of February 2014 from the crowd-sourced web-

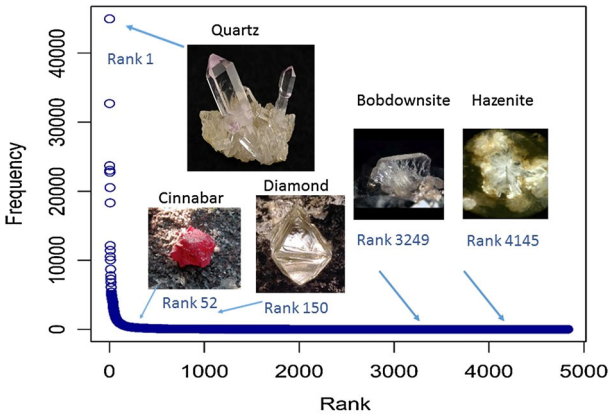


Fig. 1 Observed number of localities reported for each mineral species, called the frequency, versus the rank. The mineral species are ranked according to the frequencies

site http://Mindat.org. There are 135,415 distinct localities and, when counted over all mineral species, these data provide 652,856 observations (each observation being a unique mineral species-locality pair), which are referred to as the sample size. By comparison, in lexical statistics the sample size is the total number of words in a book (Baayen 1993), while in ecology the sample size could be the total number of individual plants (Shen et al. 2003). The objective is to identify a population model for the frequency distribution of mineral species on Earth. Discovering such models will permit the estimation of how many mineral species are yet to be discovered. Furthermore, the number of mineral species that would differ from those that have been discovered on Earth as of today can be predicted if a re-sampling of mineral species on Earth with a sample size of 652,856 observations could be performed, where approximately 4,831 minerals were discovered anew. Answers to these questions inform efforts to characterize Earth-like planets in other star systems.

The http://Mindat.org database is crowd-sourced, containing data both from the literature and from the mineral collecting community. As such, it contains inherent biases. The list of localities of the most common minerals likely has some problems. For instance, on the one hand, quartz localities will be dominated by those with aesthetic pieces, while on the other hand, the presence of minor or insignificant quartz might not have been noted by the collectors. In contrast, it is likely that the data on the extremely rare species will be quite accurate because the rare occurrences are usually noted in the literature and the low numbers make it easy to keep records. The data from http://Mindat.org represent a snapshot of their database recorded on February 2014. The snapshot was copied and the analysis was conducted on the copy.

Models for the frequency distribution used in the fields of ecology and linguistics can provide insight to mineral frequency distributions because they are also concerned with estimation of the sizes of type-rich populations. The challenge of estimating the number of biological species in an ecological population has been studied for decades (Fisher et al. 1943). For example, biologists and ecologists are concerned with assessing the diversity and richness of plant and animal species (Miller and Wiegert 1989), as well as with the effectiveness of predicting the number of new species, those presumed to exist but that have not yet been observed, in further taxonomic sampling (Shen et al. 2003). Microbial ecologists may be interested in estimating the number of taxa in a microbial population (Bunge et al. 2014). Alternatively, the field of linguistics uses models that estimate an author's vocabulary size and idiosyncrasies (Efron and Thisted 1976).

Bunge and Fitzpatrick (1993) provide an overview of research and methods to estimate the number of species. Several statistical sampling theoretical methods exist for estimating the population size (Burnham and Overton 1978, 1979; Chao 1984; Chao and Lee 1992; Chao et al. 2000; Chao and Bunge 2002). Methods based on non-parametric maximum likelihood estimation can be found in Norris and Pollock (1998) and Wang (2010). One method for estimating the number of new species in taxonomic sampling involves extrapolation of a fitted parametric model to a species accumulation curve representing the number of observed species as a function of sample size (Keating et al. 1998; Soberón and Llorente 1993). Bunge and Barger (2008) provide an overview of parametric models used in the literature that are based on mixed-Poisson distributions fitted to the frequency count data by maximum likelihood

methods. Several choices exist for mixture models, including lognormal, gamma, inverse Gaussian, the generalized inverse Gaussian, and a mixture of two or three exponentials (Bunge and Barger 2008). For example, the generalized inverse Gauss–Poisson distribution (GIGP) was introduced by Sichel (1975, 1986) and was applied in Heller (1997), who used a maximum likelihood approach for estimation of the parameters regarding counts of filarial worms on mites that live on rats (Baayen 2001). The GIGP distribution is used only rarely in ecology, probably because of numerical difficulties in fitting the model (Bunge and Barger 2008).

Baayen (2001) presents a family of models called the large number of rare events (LNRE) models for studying word frequency distributions. A LNRE distribution is characterized by numerous species/words that have extremely low relative abundance probabilities (Baayen 2001). A LNRE model is formulated in terms of a structural type distribution and takes into account the number of unobserved species/words in the population, thereby allowing the calculation of the total size of the population by extrapolation (Baroni and Evert 2007). Baayen (2001) presents two LNRE models: the lognormal and the GIGP structural type distribution. He also presents models that are not identified as proper LNRE models but recognized as generalized Zipf's law, for instance, the Yule–Simon model. Evert (2004) introduced two additional LNRE models that are based on the Zipf–Mandelbrot law: the Zipf–Mandelbrot (ZM) and the finite Zipf–Mandelbrot (fZM) LNRE models. Baroni and Evert (2005, 2007) found that the ZM, fZM, and GIGP LNRE models have better or at least the same extrapolation qualities as other models. In the following sections, these varied models will be applied to mineral species-locality data.

2 Introduction to LNRE Mineral Frequency Distribution

In statistical mineral ecology, the techniques and models developed in lexical statistics to model the word frequency distribution are adapted to model the frequency distribution of minerals. In particular, the notation and techniques used in Baayen (2001) and Evert (2004) are as follows.

Let S denote the population size of distinct mineral species on Earth and denote the ith mineral species by x_i for $i = 1, 2, \ldots, S$. Assume each mineral species x_i has a population probability π_i (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 $\sum_{i=1}^{S} \pi_i = 1$. Let N be the number of distinct mineral species-locality pairs, which is the sum of the number of mineral species found in all localities. Refer to N as the sample size. Assume that a sample of N mineral species-locality pairs is randomly and independently drawn from the population of distinct minerals species with outcomes in any of S mineral species. Let $f_i(N)$ denote the frequency of the ith mineral species x_i in the sample of N mineral species-locality pairs. Here $f_i(N)$ is the number of distinct localities for x_i as a function of the sample size N. 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_i. Let m denote the number of localities, also called frequency class. Thus, the probability that the ith mineral species x_i is found at exactly m localities is given by

$$P(f_i(N) = m) = \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m} \approx \frac{(N\pi_i)^m}{m!} \exp(-N\pi_i). \quad (1)$$

In the last line, the binomial probabilities were approximated with the Poisson probabilities with mean $N\pi_i$ for mineral species x_i, because N is large and π_i is small for all i. Since a sample of N mineral species-locality pairs will not contain all the different mineral species in the population, standard practice in this type of modeling is not to focus on the individual species, but instead to group the species within the same frequency class m.

Let $I_{[f_i(N)>0]}$ be the indicator function, which is 1 if the ith mineral species x_i is present in the sample of size N and 0 otherwise. Denote the number of distinct mineral species in a sample of N mineral species-locality pairs by $V(N) = \sum_{i=1}^{S} I_{[f_i(N)>0]}$. The graph of V as a function of N is called the mineral species accumulation curve. As of February 2014, the total number of distinct mineral species found is $V(N) = 4{,}831$ for $N = 652{,}856$. Let $I_{[f_i(N)=m]}$ be the indicator function, which is 1 if the ith mineral species x_i has frequency m and 0 otherwise. Denote the number of distinct mineral species with exactly m localities in a sample of N mineral species-locality pairs by $V_m(N) = \sum_{i=1}^{V(N)} I_{[f_i(N)=m]} = \sum_{i=1}^{S} I_{[f_i(N)=m]}$. Notice that the sum was extended to include the entire population size S, since the number of unobserved mineral species $V_0(N)$ has frequency zero in the sample. Thus, the population of distinct mineral species is split into the observed and unobserved mineral species, that is $S = V(N) + V_0(N)$. The sequence $(V_1(N), V_2(N), \ldots, V_{V(N)}(N))$ is called the observed frequency spectrum. For example, the number of distinct mineral species found at only one or two localities is $V_1(N) = 1{,}062$ and $V_2(N) = 569$. Notice the following identities $N = \sum_m m\, V_m(N)$ and $V(N) = \sum_m V_m(N)$. Using Eq. (1) (Baayen 2001), expected values of $V_m(N)$ and $V(N)$ are given, respectively, by

$$E(V_m(N)) = \sum_{i=1}^{S} \frac{(N\pi_i)^m}{m!} \exp(-N\pi_i), \quad (2)$$

and

$$E(V(N)) = \sum_{i=1}^{S} (1 - \exp(-N\pi_i)). \quad (3)$$

Notice also that the derivative of $E(V(N))$ with respect to N

$$\frac{\mathrm{d}}{\mathrm{d}N} E(V(N)) = \frac{E(V_1(N))}{N}, \quad (4)$$

is equal to the joint probability of the unobserved mineral species in the sample of size N (Baayen 2001). Given the large number of minerals with extremely low relative abundance probabilities, the mineral frequency distribution is a LNRE distribution (Baayen 2001; Khmaladze 1987). This distribution indicates that many of Earth's mineral species remain undiscovered.

For example, 22 % of all known mineral species are found at only one locality. In this distribution, the sample relative frequency of each species cannot be used to estimate

the corresponding population probability even for very large sample sizes. The sample relative frequencies tend to overestimate the population probabilities because there is no knowledge about the weight of the unobserved portion of the distribution. Furthermore, the sample relative frequencies will also change with the sample size N. Baayen (2001) asserted that the sample size has to be extremely large in order to observe the asymptotic behavior of the mineral species accumulation curve. A parametric model from the family of the LNRE models that takes into account the unobserved species can be used to handle this problem. Note that the mineral frequency distribution, like most word frequency distributions, does not rigorously satisfy the mathematical condition of the LNRE property as given in Khmaladze (1987) and Khmaladze and Chitashvili (1989). However, in practice, the mineral frequency distribution and many empirical word frequency distributions behave like a LNRE distribution for which the law of large numbers does not hold (Baayen 1993).

The structural type distribution is defined by $G(\tilde{\pi}) = \sum_{i=1}^{S} I_{[\pi_i \geq \tilde{\pi}]}$, which is the number of mineral species in the population that have probability greater than or equal to $\tilde{\pi}$ (Baayen 2001). The structural type distribution $G(\tilde{\pi})$ will be approximated by a continuous function $G(\tilde{\pi}) = \int_{\tilde{\pi}}^{\infty} g(\pi)\mathrm{d}\pi$, where $g(\pi)$ is a type density function that satisfies $g \geq 0$ and $\int_0^{\infty} \pi g(\pi)\mathrm{d}\pi = 1$ (Evert 2004). 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)

$$E(V_m(N)) = \int_0^{\infty} \frac{(N\pi)^m}{m!} \exp(-N\pi) g(\pi)\mathrm{d}\pi, \tag{5}$$

and

$$E(V(N)) = \int_0^{\infty} (1 - \exp(-N\pi)) g(\pi)\mathrm{d}\pi. \tag{6}$$

Observe that the model is a mixed-Poisson distribution, where the population abundances of the individual mineral species can be considered independent random variables. In this paper, the GIGP structural type distribution (Baayen 1993, 2001) will be used as a model for $G(\tilde{\pi})$. This model was introduced by Sichel (1971, 1975, 1986), where the type density function is given by

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

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). It follows from integration of Eqs. (5) and (6)

$$E(V_m(N)) = \frac{2Z}{bK_{\gamma+1}(b)(1+N/Z)^{\gamma/2}} \frac{\left(\frac{bN}{2Z\sqrt{1+N/Z}}\right)^m}{m!} K_{m+\gamma}(b\sqrt{1+N/Z}), \tag{7}$$

and

$$E(V(N)) = S - E(V_0(N)) = \frac{2Z}{b}\frac{K_\gamma(b)}{K_{\gamma+1}(b)}\left[1 - \frac{K_\gamma(b\sqrt{1+N/Z})}{(1+N/Z)^{\gamma/2}K_\gamma(b)}\right], \quad (8)$$

where the population size is $S = \frac{2}{bc}\frac{K_\gamma(b)}{K_{\gamma+1}(b)}$ (Baayen 2001). Here $Z = \frac{1}{c}$ represents the unit of measurements of mineral species-locality pairs for which the number of occurrences for a particular mineral species is $Z\pi$.

2.1 Expected Differences in Numbers of Mineral Species from Two Earth-Like Planets

In this section, some of the non-parametric estimators that are used in statistical ecology to estimate the number of species in a population will be described. Subsequently, a description on how to estimate the expected number of new mineral species to be discovered in a second sample among the unobserved species in the initial sample from the population of mineral species will be given. The coverage of a sample is defined to be the total relative abundance of mineral species discovered in the sample. That is, the coverage of a sample of size N is given by $C = \sum_{i=1}^{S} \pi_i I_{[f_i(N)>0]}$ (Solow and Polasky 1999; Wang 2011). According to Baayen (2001), one estimator for the sample coverage is given by

$$\hat{C} = \frac{1}{N}\sum_{m=1} m^* E(V_m(N)) = 1 - \frac{E(V_1(N))}{N} \approx 1 - \frac{V_1(N)}{N},$$

where $m^* \approx (m+1)\frac{E(V_{m+1}(N))}{E(V_m(N))}$ are the Good–Turing estimates (Good 1953).

Let S_2 be the number of new mineral species that are to be discovered among the unobserved species $V_0(N)$ in a second sample of size M. The estimation of the expected number of new mineral species $E(S_2)$ (Shen et al. 2003; Solow and Polasky 1999) is described in this section. Denote its estimator by $\hat{E}(S_2)$. Solow and Polasky (1999) proposed a quick estimator for the expected number of new species, given the information in the initial sample and assuming equal abundance for the unobserved species. The estimator proposed by Solow and Polasky (1999) is given by

$$\hat{E}(S_2) = \hat{V}_0(N)\left(1 - \left(1 - \frac{1-\hat{C}}{\hat{V}_0(N)}\right)^M\right), \quad (9)$$

where the estimator of $V_0(N)$ is obtained by Chao (1984)

$$\hat{V}_0(N) = \frac{V_1^2(N)}{2V_2(N)}, \quad (10)$$

and $\hat{C} = 1 - \frac{V_1(N)}{N}$ is the Good–Turing estimate. The estimator $\hat{V}_0(N)$ could be replaced by other estimators. Shen et al. (2003) showed that the estimator in Eq. (9) is

also valid under the assumption that species with equal frequencies in the sample also have equal relative abundances in the population. They replaced the estimator of the unseen species $\hat{V}_0(N)$ in Eq. (9) with an estimator derived by Chao and Lee (1992) and Chao et al. (2000), where the estimator incorporates the first k rare species in the sample. Define the total number of rare species in the sample by $S_0 = \sum_{m=1}^{k} V_m(N)$ and the sample coverage by $\tilde{C} = 1 - \frac{V_1(N)}{\sum_{m=1}^{k} m V_m(N)}$. Then the estimator proposed by Chao and Lee (1992) and Chao et al. (2000) of the unobserved species is given by

$$\hat{V}_0(N) = \frac{S_0}{\tilde{C}} + \frac{V_1(N)}{\tilde{C}}\hat{\gamma}^2 - S_0, \tag{11}$$

where $\hat{\gamma}^2$ is an estimator of the squared coefficient of variation of species abundance. For the total number of mineral species $\hat{\gamma}^2 = 0.448$, computed with $k = 10$. According to Chao et al. (1993) and Shen et al. (2003), a value of $k = 10$ has shown to provide the best estimates in empirical studies. A third estimator to consider is to replace $\hat{V}_0(N)$ in Eq. (9) by the jackknife estimator (Burnham and Overton 1978, 1979). The kth order jackknife estimator of $\hat{V}_0(N)$ is given by

$$\sum_{m=1}^{k} (-1)^{m+1} \binom{k}{m} V_m(N). \tag{12}$$

The expected number of new mineral species can be estimated by extrapolating from the species accumulation curve fitted to the LNRE model from sample size N to sample size $2N$ and subtracting the expected number of mineral species at sample size N. That is

$$\hat{E}(S_2) = E(V(2N)) - E(V(N)), \tag{13}$$

where the two samples have equal sample sizes N. The resulting value is the estimate of the expected number of new species to be observed in a second sample of size N that were not expected to be observed in the initial sample of size N. The resulting value is multiplied by 2 in order to estimate the number of different mineral species distributed over the two samples. This value is an estimate of the expected number of mineral species that will be different in two random samples of the same size from two modeled Earth-like planets. The performance of the estimators described in this section will be examined with Monte Carlo simulations.

3 Results

In this section, a LNRE model is fit to the total mineral species frequency spectrum. The R-package, zipfR, (Evert and Baroni 2007, 2008) is used to fit the frequency spectrum of known mineral species to a LNRE model. Sichel's GIGP LNRE model fits well to the data. 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 are $\gamma = -0.419$, $b = 0.013$, and $Z = 70.462$, with $\chi^2 = 10.36$, $df = 13$, and

p-value $= 0.66$. Figure 2a, b illustrates the frequency spectrum for the observed values and expected values using Sichel's model at sample size $N = 652{,}856$. Notice that $E(V_1(N))$ underestimates $V_1(N)$ while $E(V_2(N))$ overestimates $V_2(N)$. The number of distinct mineral species on Earth is estimated to be $S = 6{,}394$. Thus, there are 1,563 mineral species yet to be discovered, assuming the application of current sampling and identification techniques and present-day mineral formation processes. The expected value and variance of the total number of distinct mineral species are 4,826 and 662, respectively.

An important aspect in LNRE modeling is the concept of the LNRE zone (Baayen 2001). The LNRE zone is defined by Baayen (2001) as the range 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. Accordingly, most books/texts are located in the central LNRE zone where $E(V_1(N))$ is still increasing. The mineral species frequency distribution appears to be in Baayen's (2001) definition of the late LNRE zone. The late LNRE zone is defined by the values of N in the LNRE zone for which $E(V_1(N))$ is decreasing. Outside the LNRE zone, the species accumula-

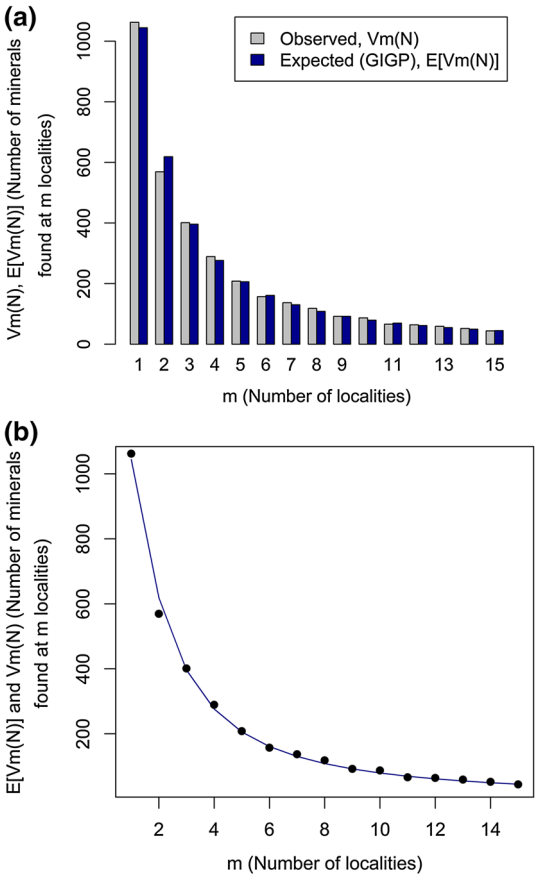


Fig. 2 In **a**, the *light barplots* represent the observed spectrum elements and the *dark barplots* represent the expected values of the spectrum elements. In **b**, the *filled circles* denote the observed spectrum elements. The *solid line* represents Sichel's fit

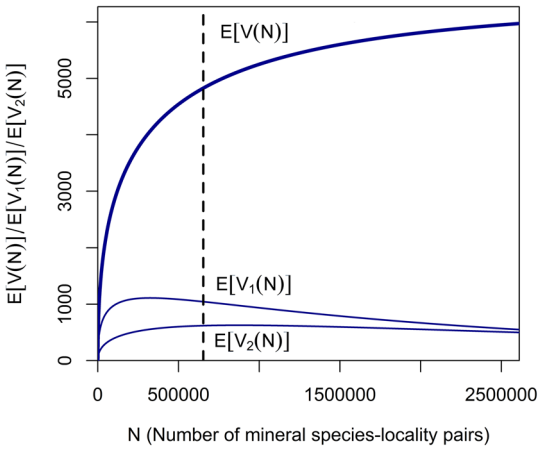


Fig. 3 Expected mineral species accumulation *curves*, extrapolated at four times the sample size ($N = 652{,}856$), using Sichel's model

tion curve reveals its asymptotic limit for finite population sizes while $E(V_1(N))$ is negligible. Outside the LNRE zone, the sample relative frequencies can be used to estimate the population probabilities (Baayen 2001). Figure 3 shows the expected mineral species accumulation curves for $E(V(N))$ and the first two frequency spectrum elements $E(V_1(N))$ and $E(V_2(N))$. The point at which the vertical dashed line intersects the x-axis denotes the current value of the sample size $N = 652{,}856$. Figure 3 indicates that the distribution is in the late LNRE zone, because the maximum of $E(V_1(N))$ was expected to be achieved for a sample size smaller than $N = 652{,}856$. At that sample size, twice as many minerals species are expected to be found at only one locality as found at only two localities (Baayen 2001). Since $E(V(N))$ is still increasing, new minerals are predicted to be discovered but the growth rate of the discovery is expected to decrease since the inflection point was reached for N less than 652,856. Notice also that, as more minerals are continued to be sampled at new localities, the total number of minerals found at only one locality is predicted to decrease because mineralogists will continue to find previously discovered minerals at new localities. The current growth rate for the total number of new minerals species is 0.0016 using Eq. (4). Thus, if an additional mineral is sampled at a new locality at sample size $N = 652{,}856$, the probability that this mineral is a new species is 0.16 %.

3.1 Extrapolation Quality and Prediction Performance

The extrapolation quality of Sichel's GIGP LNRE model was tested using a related technique described in Baroni and Evert (2005). There were 100 random sub-samples selected without replacement of sample sizes $N/2$ from the original sample of $N = 652{,}856$ mineral species-locality pairs. The observed frequency spectra of the 100 sub-samples were then fit to Sichel's GIGP LNRE model and for each model extrapolated from sample size $N/2$ up to size N, which corresponds to two times the estimation size. The resulting averages of the expected values were compared to the expected values obtained by binomial interpolation using the original sample of size $N = 652{,}856$.

The binomial interpolation described in Baayen (2001) uses the frequency spectrum to compute expected values $E(V(N))$ for sample sizes up to the sample size N for which the frequency spectrum was obtained.

The experiment was repeated using 25 and 10 % of the data and extrapolated up to four and ten times the estimation sizes, respectively. The GIGP LNRE model extrapolates reasonably well when the parameters are estimated using 25 and 50 % of the data but not when using only 10 % of the data. These findings are consistent with those previously reported (Baroni and Evert 2005). The extrapolated values $E(V(N))$ tend to underestimate the values obtained by binomial interpolation. The average values of the 100 sub-samples at size N are 4,747 (80.6) and 4,796 (36.9) using 25 and 50 % of the data, respectively. The results from using 25 and 50 % of the data are illustrated in Fig. 4a, b. Solid curves represent the expected mineral species accumulation curves

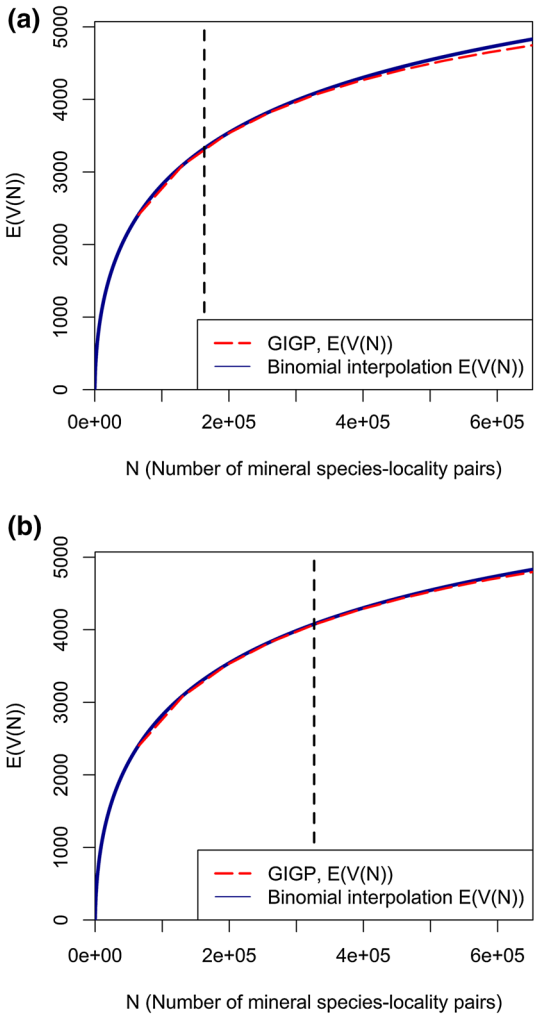


Fig. 4 *Solid curves* are the expected mineral species accumulation curves using binomial interpolation of the original sample. The *dashed curves* are the averages of the expected values computed using Sichel's model of sizes for **a** $N/4$ and for **b** $N/2$

using binomial interpolation of the original sample of size $N = 652{,}856$. The dashed curves are the averages of the expected values $E(V(N))$ computed using Sichel's model for sample sizes $N/4$ (Fig. 4a) and $N/2$ (Fig. 4b). The points at which the vertical dashed lines intersect the x-axis denote the values $N/4$ (Fig. 4a) and $N/2$ (Fig. 4b). In both figures the values of the GIGP curve to the left of these vertical lines are interpolated values and the values to the right are extrapolated values. It can be concluded that there exist more rare mineral species, each with a lower sample relative frequency, than expected. As a result, the number $S = \lim_{N\to\infty} E(V(N))$ of distinct mineral species in the population is likely underestimated (Baayen 2001). The reasons for this underestimation could be that the sample incorporates inter-species correlations, and thus is not random (Baayen 2001). For example, some groups of mineral species tend to be found together in the same locations. In addition, as a result of new and improved high-resolution sampling techniques, new mineral species are discovered more frequently than in the past. The numbers of new minerals being discovered are influenced by the mineralogical techniques employed. For example, an ongoing program at Caltech conducted by Ma et al. (2014) has identified a number of nanometer-scale minerals that could not have been identified by the light microscope techniques of earlier decades.

The prediction performance of the model for predicting the number of distinct mineral species was examined using a cross-validation technique as described in Baroni and Evert (2007) with 20 non-overlapping sub-samples. The sample size of the test set was approximately three times the sample size of the training set. The root mean squared relative error goodness-of-fit measure is obtained to be 4.0 % using the GIGP LNRE model.

3.2 Expected Number of Different Mineral Species in Two Random Samples

Using the estimators given in Eqs. (9) to (13), the population size of all mineral species on Earth is estimated, as well as the expected number of new mineral species $\hat{E}(S_2)$ that would be discovered among the previously unobserved mineral species on Earth, in a second sample of the same sample size $N = 652{,}856$, which is the observed sample size. The results appear in Table 1. Method 1 refers to the estimator proposed by Solow and Polasky (1999) using the estimator by Chao (1984) for S. Method 2 refers to the estimator proposed by Shen et al. (2003) with the estimator by Chao and Lee (1992) and Chao et al. (2000) of order $k = 10$ for S. Method 3 refers to the estimator proposed by Solow and Polasky (1999) using the jackknife estimator of order 4 for S. Method 4 refers to extrapolation from the GIGP LNRE model. Since there is no known analytic form reported for the probability distribution of the GIGP model (Evert and Baroni 2008), the performance of the estimators in Table 1 was examined using 200 pairs of non-parametric Bootstraps samples (Efron and Tibshirani 1993; Shen et al. 2003). Random samples of size $N = 652{,}856$ were generated from the mineral frequency distribution and for each sample an additional sample of the same sample size was generated. The number of new mineral species observed in the second sample (the true estimated values) and the estimated values were calculated using the estimators given in Table 1. The average value of the 200 true values and the average

Table 1 Estimated values of S and $\hat{E}(S_2)$ for all the mineral species using sample size $N = 652{,}856$

Method	1	2	3	4
S	5,822 (80.1)	5,726 (56.7)	6,980 (215.2)	6,394
$\hat{E}(S_2)$	652	622	838	662

The standard errors of S are included in parenthesis for the first three estimators. The four methods are described in the text

Table 2 Using 200 Monte Carlo simulations to compare the values of $\hat{E}(S_2)$ for all the mineral species

Estimators by	$N = 10{,}000$	$N = 100{,}000$	$N = 652{,}856$
S. P. (1999)/Chao (1984)	379.9 (22.6)	588.7 (26.0)	318.3 (20.7)
Shen et al. (2003)	384.5 (20.8)	595.1 (20.5)	318.3 (15.3)
S. P. (1999)/jackknife	434.7 (22.8)	718.0 (35.9)	433.0 (23.3)
GIGP	396.9 (22.1)	621.3 (20.8)	357.3 (13.7)
True value	402.6 (18.0)	638.8 (22.8)	339.8 (17.1)

The standard errors are included in parenthesis. S. P. refers to the estimator proposed by Solow and Polasky (1999)

value of the 200 estimated values were obtained along with the sample standard errors of the estimators. The experiment was repeated for sample sizes $N = 10{,}000$ and $N = 100{,}000$. The results appear in Table 2.

For $N = 10{,}000$, the value of $\hat{E}(S_2)$ obtained by extrapolation of the GIGP model is close to the true value. It is well known that the estimator proposed by Chao (1984) underestimates the population size S (Shen et al. 2003; Wang 2011) and the same tendency is seen by the estimators proposed by Chao and Lee (1992) and Chao et al. (2000). Thus, these two estimators provide lower bounds for the population size of distinct mineral species and, hence, for $\hat{E}(S_2)$. Non-parametric bootstrap techniques are unreliable for estimation of uncertainty for heavy-tailed distributions (Kyselý 2010). An alternative technique should be employed in testing the estimators and to obtain standard errors of the estimators. For the total mineral dataset, the GIGP model performs better in computing $\hat{E}(S_2)$ compared to the other estimators. Therefore, the value of $\hat{E}(S_2)$ obtained from extrapolation of the species accumulation curve fit to the GIGP LNRE model was employed. It follows from Table 1 that $\hat{E}(S_2) = 662$. Thus, in two random samples from the population of mineral species on Earth, it is expected that $662 \times 2 = 1{,}324$ mineral species will be different. Therefore, $(\hat{E}(S_2)/E(V(N)) \times 100 = (662/4{,}826) \times 100 = 13.7\,\%$ of the species are expected to be different over two samples. An interpretation of this value indicates that two identical Earth-like planets are expected to have $4{,}826 - 662 = 4{,}164$ minerals species in common. Since extrapolation of the GIGP model tends to underestimate the number of distinct mineral species, the number of different mineral species in two random samples from two Earth-like planets is proposed to be larger than this value. Thus, the number 1,324 or approximately 13.7 % of species is a lower bound for the number of different mineral species on two Earth-like planets.

4 Conclusion

This contribution is the first in a series of studies that attempts to apply large mineralogical data resources and statistical methods to understand the diversity and distribution of mineral species on Earth, as well as other terrestrial planets and moons. It is shown that the distribution of Earth's mineral kingdom, which is dominated by rare mineral species found at few localities, is effectively modeled using LNRE models. Two significant conclusions are: (i) that the present inventory of 4,831 described and approved mineral species is significantly incomplete. At least 1,500 mineral species remain to be discovered employing current techniques; and (ii) a replaying of mineral evolution on Earth, repeating the same deterministic factors (for example, the same ratios of chemical elements and biological processes), would result in more than 13.7 % of mineral species different from those discovered thus far.

Many avenues await further exploration. What model works best for other elements, or for subsets of mineral species-locality data that reflect geographic, spatial correlation effects, tectonic, age, or other restrictive factors? Are any aspects of mineral species frequency distributions indicative of life-biosignatures that might apply to other worlds? These questions will provide a dynamic focus for future studies.

Acknowledgments Joshua Golden, Edward Grew, and Dimitri Sverjensky provided valuable advice and discussions. We thank the Deep Carbon Observatory, the Keck Foundation, and a private foundation for support.

References

Baayen RH (1993) Statistical models for word frequency distributions: a linguistic evaluation. Comput Humanit 26:347–363

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

Baroni M, Evert S (2007) Words and echoes: assessing and mitigating the non-randomness problem in word frequency distribution modeling. In: Proceedings of the 45th annual meeting of the association for computational linguistics, Prague, pp 904–911

Baroni M, Evert S (2005) Testing the extrapolation quality of word frequency models. In: Danielsson P, Wagenmakers M (eds) Proceedings of corpus linguistics 2005, Birmingham, UK. The corpus linguistics conference series, vol 1

Bunge J, Barger K (2008) Parametric models for estimating the number of classes. Biom J 50(6):971–982

Bunge J, Fitzpatrick M (1993) Estimating the number of species: a review. J Am Stat Assoc 88(421):364–373

Bunge J, Willis A, Walsh F (2014) Estimating the number of species in microbial diversity studies. Annu Rev Stat Appl 1:427–445

Burnham KP, Overton WS (1978) Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika 65(3):625–633

Burnham KP, Overton WS (1979) Robust estimation of population size when capture probabilities vary among animals. Ecology 60(5):927–936

Chao A (1984) Nonparametric estimation of the number of classes in a population. Scand J Stat 11(4):265–270

Chao A, Bunge J (2002) Estimating the number of species in a stochastic abundance model. Biometrics 58(3):531–539

Chao A, Lee SM (1992) Estimating the number of classes via sample coverage. J Am Stat Assoc 87(417):210–217

Chao A, Ma MC, Yang MCK (1993) Stopping rules and estimation for recapture debugging with unequal failure rates. Biometrika 80:193–201

Chao A, Hwang WH, Chen YC, Kuo CY (2000) Estimating the number of shared species in two communities. Stat Sin 10:227–246

Efron B, Thisted R (1976) Estimating the number of unseen species: how many words did Shakespeare know? Biometrica 63(3):435–447

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap, monographs on statistics and applied probability, vol 57. Chapman & Hall/CRC, London

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, 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, posters and demonstrations session, Prague, 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 10 Nov 2008

Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. J Anim Ecol 12(1):42–58

Good IJ (1953) The population frequencies of species and the estimation of population parameters. Biometrika 40:237–264

Hazen RM, Grew ES, Downs RT, Golden J, Hystad G (2015) Mineral ecology: chance and necessity in the mineral diversity of terrestrial planets. Can Mineral 53(2). doi:10.3749/canmin.1400086

Heller G (1997) Estimation of the number of classes. S Afr Stat J 31:65–90

Keating KA, Quinn JF, Ivie MA, Ivie LL (1998) Estimating the effectiveness of further sampling in species inventories. Ecol Appl 8(4):1239–1249

Khmaladze EV (1987) The statistical analysis of large number of rare events. Tech. Rep. MS-R8804, Department of Mathematical Statistics, Center for Mathematics and Computer Science, CWI, Amsterdam, Netherlands

Khmaladze EV, Chitashvili RJ (1989) Statistical analysis of large number of rare events and related problems. Trans Tbilisi Math Inst 91:196–245

Kyselý J (2010) Coverage probability of bootstrap confidence intervals in heavy-tailed frequency models, with application to precipitation data. Theor Appl Climatol 101:345–361

Ma C, Beckett JR, Rossman GR (2014) Monipite, MoNiP, a new phosphide mineral in a Ca-Al-rich inclusion from the Allende meteorite. Am Mineral 99(1):198–205

Miller RI, Wiegert RG (1989) Documenting completeness, species-area relations, and the species-abundance distribution of a regional flora. Ecology 70(1):16–22

Norris JL, Pollock KH (1998) Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. Environ Ecol Stat 5(4):391–402

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, 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

Soberón J, Llorente J (1993) The use of species accumulation functions for the prediction of species richness. Conserv Biol 7(3):480–488

Solow AR, Polasky S (1999) A quick estimator for taxonomic surveys. Ecology 80(8):2799–2803

Wang JP (2010) Estimating species richness by a Poisson-compound Gamma model. Biometrika 97(3):727–740

Wang JP (2011) SPECIES: an R package for species richness estimation. J Stat Softw 40(9):1–15