Supplementary Material to "Earth's 'missing' minerals" by Hazen et al.

Earth's "missing" minerals

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

Joshua Golden[3], Alex J. Pires[3], and Edward S. Grew[4]

On the nature of LNRE functions

(after Hystad et al. 2015a, 2015b)

In our efforts to model the frequency distribution of minerals, we have modified and adapted techniques and models developed in lexical statistics to characterize word frequency distributions. In particular, the notation and techniques used in Baayen (2001) and Evert (2004) are as follows.

Let S denote the population size of mineral species and denote the ith mineral species by x_i for $i = 1, 2, \ldots, S$. Assume that each mineral species x_i has a population probability π_i (relative abundance) of being sampled at an arbitrary locality, where $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_S$ defines the ordering schemes and $\sum_{i=1}^{S} \pi_i = 1$. Let N, the sample size, be the number of distinct mineral species-locality pairs, which is the sum of the number of mineral species found in all localities. We assume that a sample of N mineral species-locality pairs is randomly and independently drawn from the population of distinct minerals species. Let $f_i(N)$ denote the frequency of the ith mineral species x_i in the sample of N mineral species-locality pairs; thus, $f_i(N)$ 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:

$$P[f_i(N) = m] = \binom{N}{m}\pi_i^{m}(1 - \pi_i)^{N-m} \approx [(N\pi_i)^{m}/m!]\exp(-N\pi_i). \qquad (1)$$

In Equation (1) the binomial probabilities are 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. A sample of N mineral species-locality pairs will not contain all the different mineral species in the population; therefore, 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 $I_{[fi(N)>0]}$ 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 $V(N) = \sum_{i=1}^{S} I_{[fi(N)>0]}$. Refer to the growth curve of $V(N)$, as V varies with N, as the mineral species accumulation curve. As of February 2014, the total number of distinct mineral species found is $V(N) = 4831$ for $N = 652{,}856$. Let $I_{[fi(N)=m]}$ 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 $V_m(N) = \sum_{i=1}^{V(N)} I_{[fi(N)=m]} = \sum_{i=1}^{S} I_{[fi(N)=m]}$. Notice that the sum was extended to include the entire population size S, because 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 ($V_1(N)$, $V_2(N)$, …, $V_{V(N)}(N)$) is called the observed frequency spectrum. For example, the number of distinct mineral species found at only one or two localities is $V_1(N) = 1062$ and $V_2(N) = 569$. Notice the following identities: $N = \sum_m mV_m(N)$ and $V(N) = \sum_m V_m(N)$. Using Equation (1) from Baayen (2001), the expected values of $V_m(N)$ and $V(N)$ are given, respectively, by:

$$E(V_m(N)) = \sum_{i=1}^{S}[(N\pi_i)^m/m!]\exp(-N\pi_i) \qquad (2)$$

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

Notice also that the growth rate of $E(V(N))$ is given by:

$$(\frac{d}{dN})E(V(N)) = [E(V_1(N))/N], \qquad (4)$$

which 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 large number of rare events (LNRE) distribution (Baayen 2001; Khmaladze 1987).

The Generalized Inverse Gauss-Poisson Model

We employ two different LNRE models: the generalized inverse Gauss-Poisson (GIGP) and finite Zipf-Mandelbrot (fZM) models. The structural type distribution is defined by $G(\bar{\pi}) = \sum_{i=1}^{S} I_{[\pi_i \geq \bar{\pi}]}$, which is the number of mineral species in the population that have probability greater than or equal to $\bar{\pi}$ (Baayen 2001). The structural type distribution $G(\bar{\pi})$ will be approximated by a continuous function $G(\bar{\pi}) = \int_{\bar{\pi}}^{\infty} g(\pi)d\pi$, where $g(\pi)$ is a type density function that satisfies $g \geq 0$ and $\int_0^{\infty} \pi g(\pi)d\pi = 1$ (Evert 2004). The population size is given by $S = \int_0^{\infty} g(\pi)d\pi$. Since G is of bounded variation, the expressions in Equations (2) and (3) can be written in terms of the Stieltjes integrals (Baayen 2001):

$$E(V_m(N)) = \int_0^{\infty}[(N\pi)^m/m!]\exp(-N\pi)g(\pi)d\pi, \qquad (5)$$

and

$$E(V(N)) = \int_0^{\infty}[1 - \exp(-N\pi)]g(\pi)d\pi, \qquad (6)$$

Observe that the model is a mixed-Poisson distribution, where the population abundances of the individual mineral species can be considered independent random variables.

The generalized inverse Gauss-Poisson structural (GIGP) type distribution (Baayen 1993, 2001) can be used as a model for $G(\bar{\pi})$. This model was introduced by Sichel (1971, 1975, 1986), where the type density function is given by:

$$g(\pi) = [(2/bc)^{\gamma+1}/2K_{\gamma+1}(b)]\pi^{\gamma-1}\exp[-(\pi/c)-(b^2c/4\pi)], \qquad (7)$$

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 finite Zipf-Mandelbrot Model

An alternate LNRE model is provided by a structural type distribution reformulated for the finite Zipf-Mandelbrot (fZM) law as a model for $G(\bar{\pi})$. The Zipf-Mandelbrot law is reformulated as a LNRE model using the type density function of Evert (2004), where $g(\pi) = C\pi^{-\alpha-1}$ for $A \leq \pi \leq B$ and otherwise $g(\pi) = 0$, and $C = [(1-\alpha)/(B^{1-\alpha}-A^{1-\alpha})]$ is the normalization constant. The upper cutoff parameter is $B > 0$, the lower cutoff parameter A satisfies $0 < A < B \leq 1$, and $0 < \alpha < 1$. The population size is given by $S = \frac{1-\alpha}{\alpha}[(A^{-\alpha}-B^{-\alpha})/(B^{1-\alpha}-A^{1-\alpha})]$ and from Evert (2004) we have:

$$E(V_m(N)) = \frac{C}{m!}N^{\alpha}\Gamma(m-\alpha,NA), \qquad (8)$$

and

$$E(V(N)) = CN^{\alpha}\frac{\Gamma(m-\alpha,NA)}{\alpha} + (C/\alpha A^{\alpha})[1-\exp(-NA)], \qquad (9)$$

where

$$\Gamma(m-\alpha,NA) = \int_{NA}^{\infty} t^{m-\alpha-1}\exp(-t)dt. \qquad (10)$$

Additional information on LNRE models and their application to predicting mineral species accumulation curves is provided by Hystad et al. (2015a, 2015b).

Reference Cited in Supplementary Material

Baayen, R.H. (1993) Statistical models for word frequency distributions: a linguistic evaluation. Computational Humanities, 26, 347-363.

Baayen, R.H. (2001) Word Frequency Distributions, Text, Speech and Language Technology, volume 18. Dordrecht, The Netherlands: Kluwer.

Evert, S. (2004) Simple LNRE model for random character sequences. In Proceedings of the 7th Journées Internationale d'Analyse Statistique des Données Textuelles. Louvine-la-Neuve, 411-422.

Hystad, G., Downs, R.T., and Hazen, R.M. (2015a) Mineral frequency distribution conforms to a Large Number of Rare Events model: Prediction of Earth's missing minerals. Mathematical Geosciences, DOI 10.1007/s1 1004-015-9600-3.

Hystad, G., Downs, R.T., Grew, E.S., and Hazen, R.M. (2015b) Statistical analysis of mineral diversity and distribution: Earth's mineralogy is unique. Earth & Planetary Science Letters, in press.

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

Sichel, H.S. (1971) On a family of discrete distributions particularly suited to represent long-tailed frequency data. Proceedings of the Third Symposium on Mathematical Statistics, Pretoria, South Africa, 51-97.

Sichel, H.S. (1975) On a distribution law for word frequencies. Journal of the American Statistical Association, 70, 542-547.

Sichel, H.S. (1986) Word frequency distributions and type-token characteristics. Mathematical Sciences, 11, 45-72.