Population Size Estimation Using a Few Individuals as Agents

Farid Movahedi Naini*, Olivier Dousse†, Patrick Thiran* and Martin Vetterli*
* EPFL, Lausanne, Switzerland
† Nokia Research Center, Lausanne, Switzerland

Abstract—**We conduct an experiment where ten attendees of an open-air music festival are acting as Bluetooth probes. We then construct a parametric statistical model to estimate the total number of visible Bluetooth devices in the festival area. By comparing our estimate with ground truth information provided by probes at the entrances of the festival, we show that the total population can be estimated with a surprisingly low error (1.26% in our experiment), given the small number of agents compared to the area of the festival and the fact that they are regular attendees who move randomly. Also, our statistical model can easily be adapted to obtain more detailed estimates, such as the evolution of the population size over time.**

I. Introduction

Nearly every current mobile phone is equipped with a Bluetooth radio interface, each having a unique MAC address. This technology was originally designed to replace wires between electronic devices. In order to ease the peering of devices, it includes a detection functionality, where enabled devices can detect each other within a small radius (typically 10-20m). It has also been observed [1] that a non-trivial fraction of mobile phone users leave the detection feature of their phone turned on constantly ("visible mode"), most probably because the energy autonomy of their phone is not much affected to attract their attention. A particularly interesting feature is that when they are in visible mode, phones broadcast their MAC address, which makes them *uniquely identifiable*. This possibility allows therefore to use mobile phones as sensing devices, and to evaluate different features of a population related to their mobility patterns. We focus here on a more specific problem, which is the population size estimation.

In this paper, we only consider the case where measurements are performed by mobile agents that move randomly, as every other user (the mobile phones carried by "standard" users), and not by agents who would carefully swipe the monitored area. Is it possible to estimate with a good accuracy the size of the population in a closed environment from such traces? To the best of our knowledge, this is the first effort to use such measurements for population size estimation.

In order to study the feasibility and accuracy of population size estimation, we conducted an experiment at Paléo Music Festival [2] that took place in Nyon, Switzerland in July 2010. As explained in the next section, this festival provides a good environment to perform experiments related to population sampling. We use the obtained data from this experiment as a basis to benchmark our method.

The problem and the solution exposed in this paper are closely related to problems addressed in some fields such as ecology, biostatistics and information theory. Ecologists and biostatisticians are interested in estimating population sizes of certain animals (refer to [3], [4], [5] for a review). One of their techniques is called "Capture-Recapture", where some of the animals in a population are first caught (by setting up traps), marked and released. In the recapture process, some of the animals are captured again and the number of previously marked animals will provide information that is used to infer about the population size [6], [7]. Thanks to the unique Bluetooth MAC address attached to every device, we can keep a similar record of the individuals who have already been seen and thus apply similar methods in our setting. In the field of information theory, alphabet size estimation [8], pattern likelihood maximization [9], and sequence probability estimation [10], [11] also address related problems.

In contrast to the above works, we do not place monitoring devices or traps at given places, and we cannot start and terminate the measurement campaign at given times. In our case, the "sensing devices" are carried by regular individuals from the population, with an uncontrolled, random mobility pattern, and who arrive and leave the monitored area at different, random times. Consequently, after describing the experiment we conducted at Paléo Music Festival and the obtained measurements in Section II, we develop a method that factors in these sources of uncertainty in Section III. We discuss the estimation results in Section IV.

II. Experiment

A. Experiment description

Paléo Music Festival is one the major music festivals in Europe, which attracts more than thirty thousand attendees per day. It is an open-air festival that allows GPS coverage, and takes place within a closed area with fixed entrance/exit points. The surface of the festival covers around 120000 m^2. These characteristics make this festival a good environment for performing experiments related to population sampling. In order to have a better understanding of the environment of the festival, a map is shown in Figure 1. Our idea is to sample the population by sending some attendees as "agents" inside the festival. Each agent is equipped with a mobile phone (Nokia N95) that is programmed to regularly scan for Bluetooth devices within its range (around 10-20 meters). The phone then collects Bluetooth MAC addresses of mobile devices that have

978-1-4577-0594-6/11/$26.00 ©2011 IEEE

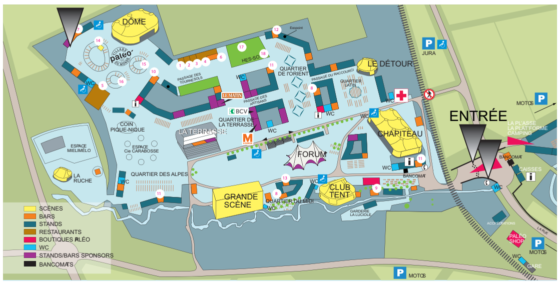

Fig. 1. Paléo Music Festival map. The surface covers around 120000 m^2. Position of the entrance phones is indicated by dark triangular markers.

their Bluetooth visibility turned on. Bluetooth MAC addresses are unique to each device and can be used as identifiers of attendees. The goal is to use this information to estimate the population size of attendees (or the subset of them that carry visible Bluetooth devices).

In order to have a ground truth of the number of visible Bluetooth devices at the festival, a regular Bluetooth scanning is done at the entrances of the festival as well. Two mobile phones are installed at the main entrance of the festival, and another phone is installed at the back entrance. The position of these three mobile phones is shown by markers in Figure 1. The same gates are used both for the entrance and for the exit of attendees. Some additional information, such as the estimated total number of attendees at the festival (obtained on the basis of the number of sold tickets and counted tickets at the entrance gates), is also provided by the organizers of the festival.

In our experiment, ten (unrelated) people were chosen to take part as agents. Agents' phones and entrance phones are programmed to perform Bluetooth scanning every 80 seconds. The experiment was performed during one day of the festival, and the duration of the festival on that day was 13 hours.

B. Measurements

In this section we discuss the measurements obtained in the experiment.

1) Preprocessing: The measurements are first preprocessed in order to discard irrelevant information. For the entrance phones, we consider only the Bluetooth traces that were collected during the opening hours of the festival. For the agents' phones, we consider only the Bluetooth traces that were collected during the period when the agents were on the festival grounds. Using the entrance phones traces, it is possible to determine the time period during which the agents were on the festival grounds.

2) Measurements at entrance: 3326 different Bluetooth devices were detected at the entrance. The estimated number of attendees given by the organizers of the festival is 40536. By dividing the number of detected Bluetooth devices at the entrance by the total number of attendees, we get the approximate percentage of attendees that have visible Bluetooth devices. This ratio is equal to 8.2%, which is close to the values reported in the literature (4.7% to 7% in [1])[1].

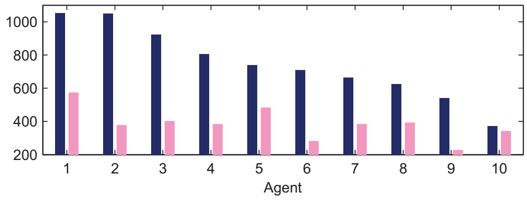

Fig. 2. Number of different Bluetooth devices detected by each agent (left bar), and the duration of stay (in minutes) for each agent (right bar).

3) Measurements by agents: The agents were able to detect 2637 out of 3326 Bluetooth devices detected at the festival, which corresponds to 79.3% of the Bluetooth devices. We expect this ratio to be less than 100%, because there were only a few agents present in the festival and the mobile phones have a short Bluetooth range. Nevertheless, this ratio is pretty large: 10 agents, spending a few hours at a large area, and with more than 3300 visible Bluetooth devices, have detected nearly 80% of them.

Figure 2 shows the number of different Bluetooth devices detected by each agent, and the time duration of stay (in minutes) for each agent. As mentioned before, our goal is to estimate the total number of visible Bluetooth devices at the festival (3326) based on agents' Bluetooth traces.

III. Model

A. Data structure and notation

1) Population: The population is comprised of attendees with visible Bluetooth devices. We denote its size by N. We call the population members *individuals* and use variable i for indexing them. Denote the festival duration by T_{fest}. For simplicity, we shift the time origin such that the festival opening time is at time 0 and its closing time is at time T_{fest}. Let st_i and dt_i denote, respectively, the entrance and departure times of individual i to the festival; these variables are not directly observable (at least not for all individuals), and will be treated as random variables, which are assumed to be i.i.d. across the population. We denote by $f(st, dt)$ their probability density distribution (pdf), on which we will elaborate later. Moreover, let t^i_{frst} and t^i_{last} denote the first and the last time, respectively, when individual i has been detected by any of the agents. This information indicates that individual i has been on the festival grounds between t^i_{frst} and t^i_{last}.

2) Agents: We denote the number of agents by M and use variable j for indexing them. Let st^A_j and dt^A_j denote the entrance and departure times of agent j to the festival. Note that, unlike individuals, agents' entrance and departure times are known to us. Let $t^j_{st_i,dt_i}$ denote the duration of time between the entrance and departure of individual i, which is overlapped with the entrance and departure of agent j[2]. We have $t^j_{st_i,dt_i} = \max\left(\min(dt^A_j, dt_i) - \max(st^A_i, st_i), 0\right)$.

3) Detection: The data that each agent provides consists of a list of MAC addresses detected by the agent together with the corresponding detection times. Denote the total number

[1] The ratio is a bit higher probably because the population structure (such as age) at Paléo is different than the population structure in [1].

[2] We assume that when an individual or an agent enters the festival, he stays on the festival grounds until he departs from the festival.

of detected MAC addresses by S and map the detected MAC addresses to the set $\{1, \ldots, S\}$. Note that this mapping is not unique. Denote by k_{ij} the number of times that individual i has been detected by agent j[3]. Let $n_i = \sum_{j=1}^{M} k_{ij}$ denote the total number of times that individual i has been detected. Note that individual i is *observed* if and only if $n_i > 0$ (if it has been detected by at least one of the agents).

B. Likelihood based estimation

Our model is mainly based on the following two assumptions.

- Poisson detection: We assume that the number of times an agent detects an individual is Poisson distributed.
- Independence: We assume that the detection of any individual by any agent is independent of all other individuals and agents.

More precisely, we assume that the number k_{ij} of times that agent j detects individual i is a Poisson random variable with parameter $\lambda_i t^j_{st_i,dt_i}$. In other words, we set the mean number of detections of individual i by agent j to be proportional to the amount of time during which both individual i and agent j are on the festival grounds ($t^j_{st_i,dt_i}$) and to a factor specific to i which we call *detection rate* (λ_i) of individual i.

Moreover, we treat λ_i as a random variable. We assume that for individual i, λ_i is drawn from a Gamma distribution with parameters α and β, independently from other individuals and from its arrival and departure time. We use the Gamma prior because it is a flexible distribution and it is the conjugate prior of the Poisson distribution. The probability density function of λ_i therefore reads: $f_{\lambda_i}(\lambda_i; \alpha, \beta) = \beta^{\alpha} e^{-\beta\lambda_i} \lambda_i^{\alpha-1}/\Gamma(\alpha)$.

The Poisson-mixed model has previously been used in the literature to address problems related to population size estimation [12], [13]. In these methods, all the population members (animals for example) are vulnerable to the sampling process (traps for example) for the entire duration of experiment. However, in our experiment, this assumption does not hold, and we account for this by using the pdf $f(dt, st)$. Some other methods [7], [3] could be applied to this problem, but they will only account for whether individual i has been detected by agent j or not. In other words, they only take into account $\mathbb{1}_{\{k_{ij}>0\}}$ and not k_{ij}. These methods attack the problem by modeling the detection probability of an individual. A limitation of this approach is that the detection probability of an individual does not linearly scale with time and hence the effect of time cannot be readily included. In contrast, in the Poisson model, the average number of times agent j detects individual i scales linearly with time, as one would expect. Moreover, parameters λ_i and $t^j_{st_i,dt_i}$ have meaningful interpretations.

In order to derive the estimator for N, we compute the probability of observing the obtained measurements under the model described above with parameters N, α, β. This is usually called the *likelihood function*. We then pick the set of parameters, in particular N, that maximize this likelihood. The likelihood function has the following form:

$$L(N, \alpha, \beta) = \underbrace{\binom{N}{S} (1 - p_{det}(\alpha, \beta))^{N-S}}_{L_1(N,\alpha,\beta)} \cdot \underbrace{\prod_{i=1}^{S} \mathbb{P}_i}_{L_2(\alpha,\beta)}, \qquad (1)$$

where p_{det} and $\mathbb{P}_i$ are given below.

The first term (L_1) is related to the likelihood of the unobserved individuals, and the second term (L_2) is related to the likelihood of the pattern of the observed individuals. We discuss below each part of the likelihood function.

1) Likelihood of the unobserved: Let $p_{det}^{(st,dt,\lambda)}$ be the probability of observing an individual having detection rate λ, and entrance and departure times st, dt. Using the Poisson detection assumption, we have

$$p_{det}^{(st,dt,\lambda)} = 1 - \prod_{j=1}^{M} e^{-\lambda t^j_{st,dt}} = 1 - e^{-\lambda \sum_{j=1}^{M} t^j_{st,dt}}. \qquad (2)$$

Since λ, st and dt are random variables, we compute the expectation of this probability over (st, dt, λ):

$$p_{det}(\alpha, \beta) = 1 - \mathbb{E}_{st,dt}\left[\left(\frac{\beta}{\beta + \sum_j t^j_{st,dt}}\right)^{\alpha}\right]. \qquad (3)$$

The likelihood of the unobserved individuals is equal to the probability of not observing $N - S$ of the individuals:

$$\begin{aligned} L_1(N, \alpha, \beta) &= \binom{N}{N-S} (1 - p_{det}(\alpha, \beta))^{N-S} \\ &= \binom{N}{S} \left(\mathbb{E}_{st,dt}\left[\left(\frac{\beta}{\beta + \sum_{j=1}^{M} t^j_{st,dt}}\right)^{\alpha}\right]\right)^{N-S}. \end{aligned} \qquad (4)$$

2) Likelihood of the observed: We first compute the probability of the observed pattern of detection by each agent for one of the observed individuals. Given that individual i has detection rate λ and entrance and departure times st, dt, the probability for him to be detected k_{ij} times by agent j for $j = 1, \ldots, M$, with $t^i_{frst} > st$ and $t^i_{last} < dt$, is

$$\mathbb{P}_i^{(st,dt,\lambda)} = \prod_{j=1}^{M} e^{-\lambda t^j_{st,dt}} \frac{(\lambda t^j_{st,dt})^{k_{ij}}}{k_{ij}!} \mathbb{1}_{\{st < t^i_{frst}, dt > t^i_{last}\}}. \qquad (5)$$

Again taking expectations, we get

$$\mathbb{P}_i = \mathbb{E}_{st,dt}\left[\frac{\Gamma(\alpha + n_i)\beta^{\alpha} \mathbb{1}_{\{st < t^i_{frst}, dt > t^i_{last}\}}}{\Gamma(\alpha)(\beta + \sum_{j=1}^{M} t^j_{st,dt})^{\alpha+n_i}} \prod_{j=1}^{M} \frac{(t^j_{st,dt})^{k_{ij}}}{k_{ij}!}\right]. \qquad (6)$$

The second part of the likelihood is equal to the probability of the observed pattern for all the observed individuals. Using the independence assumption we have

$$L_2(\alpha, \beta) = \prod_{i=1}^{S} \mathbb{P}_i. \qquad (7)$$

[3] As Bluetooth scanning is performed every 80 seconds, if we observe a burst of repeated detections of individual i by agent j, we only consider the first detection of the burst.

3) Maximum likelihood estimator: We define the maximum likelihood estimators for N, α, β as

$$(\hat{N}, \hat{\alpha}, \hat{\beta}) = \underset{N,\alpha,\beta}{\arg\max} \log L(N, \alpha, \beta). \qquad (8)$$

Where $L(N, \alpha, \beta)$ is the full likelihood given by (1), (4) and (7). $\hat{N}$ is the maximum likelihood estimator for the population size.

C. Estimating the total number of attendees

Remember that N is the number of attendees who carry visible Bluetooth devices. By applying the ratio of attendees that have visible Bluetooth devices to the estimated N, we can estimate the total number of attendees. Let N_{Tot} be the total number of attendees and let r be the ratio of attendees carrying visible Bluetooth devices: $r = N/N_{Tot}$. Let $\hat{N} = N(1+\Delta N)$ and $\hat{r} = r(1+\Delta r)$ be the estimates for N and r, respectively, with relative errors equal to ΔN and Δr. If $|\Delta N| \ll 1$ and $|\Delta r| \ll 1$ then,

$$\hat{N}_{Tot} = \frac{\hat{N}}{\hat{r}} = \frac{N(1+\Delta N)}{r(1+\Delta r)} \approx N_{Tot}(1 + \Delta N - \Delta r),$$

which means in the worst case, the relative error in estimating the total number of attendees is approximately equal to the sum of the relative errors in estimating N and r.

IV. Results

In this section we discuss some results from the application of our model to the data. We first elaborate on the choice of the model for arrival and departure times $f(st, dt)$.

A. Choice of $f(st, dt)$

We use three different entrance and departure times distributions which we discuss below.

1) Deterministic: One extreme choice for $f(st, dt)$ is a deterministic entrance time and departure time for all the individuals. We choose $f_1(st, dt) = \delta(st)\delta(dt - T_{fest})$, where $\delta(\cdot)$ is the Dirac function. This distribution assumes that all the individuals enter at the beginning of the festival (time 0) and leave at the end of the festival (T_{fest}), similarly to the studies in [7], [12], [13].

2) Estimated actual distribution: The opposite extreme choice for $f(st, dt)$ is to use the Bluetooth traces obtained from entrance phones to estimate the distribution of $f(st, dt)$. This information is in general not available, but is used in our experiment for benchmarking purposes. Recall that the entrance phones perform a Bluetooth scanning at the entrance gates; as a result, they measure entrance and departure times for all individuals. After observing entrance phones traces, we computed the empirical distribution of $f(st, dt)$.

3) Low informative: In practice, we do not have detailed enough information of entrance and departure times to estimate $f(st, dt)$. We assume that individuals enter uniformly at random between the start of the festival until the mid-time of the festival. In other words, $st \sim \mathcal{U}(0, T_{fest}/2)$. We also assume that the duration of stay for each individual is $\mathcal{N}(T_{fest}/2, 2^{\text{hours}})$ and is independent of the entrance time. For generating a valid (st, dt), we draw an entrance time and a positive duration of stay according to the described distributions; the departure time is accepted only if it is smaller than T_{fest}.

Parameter	Choice of $f(st, dt)$		
	$f_1(st, dt)$	$f_2(st, dt)$	$f_3(st, dt)$
$\hat{\alpha}$	1.588	1.994	1.935
$\hat{\beta}$	1669.4	1653.9	1624.2
$\hat{p}_{det}(\alpha, \beta)$	0.850	0.796	0.803
$\hat{N}$	3104	3314	3284
$(N - \hat{N})/N$	6.67%	0.36%	1.26%

TABLE I
Comparison of the estimated population size with the ground truth ($N = 3326$) for three different distributions of entrance and departure times.

Method	$\hat{N}$	$(N - \hat{N})/N$
M_{th} in [7]	3013	9.46%
[8]	2676	19.54%

TABLE II
Result of applying the estimators in [7], [8] to the measurements.

B. Estimating the population size

For each of the three pdf $f(st, dt)$ described above, we maximized the full likelihood give in (8) using numerical methods. The result is given in Table I.

We observe in the table that the naive choice of deterministic entrance and departure times gives a relatively large undershoot. An explanation for this undershoot is that based on $f_1(st, dt)$, all the individuals are in contact with all the agents, and hence the overlap time between agents and individuals is overestimated. The detection probability is overestimated, which results in an undershoot. By using a probabilistic $f(st, dt)$ instead, individuals are on average in contact with the agents for a smaller time duration, hence the detection probability decreases and we have an increase in the estimated population. We also observe that by estimating $f(st, dt)$ from the entrances traces, we get surprisingly close to the true value ($N = 3326$). Finally, the low informative $f_3(st, dt)$ gives a reasonably good result.

We compare our method with the capture-recapture method described in [7] and with the method in [8]. The results are shown in Table II. Both methods exhibit an undershoot. Remember that the time duration which each individual is vulnerable to the sampling process is random (according to its entrance and departure time), which is not taken into account in [7]. Therefore, the result has an undershoot similar to our method for the choice of $f_1(st, dt)$. The method in [7] assumes uniform sampling of the population, which is not valid in our experiment and is the reason for the undershoot. We remark that the approximation used in the estimator in [8] is not valid for our measurements, thus we have used the exact expression.

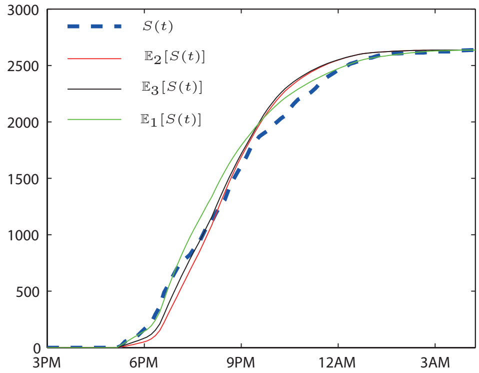

Fig. 3. The dashed line is the cumulative number of individuals detected over time ($S(t)$ defined in IV-C). The time axis is shifted to the opening/closing hours of the festival. The solid lines $\mathbb{E}_i[S(t)]$ are the average of $S(t)$, computed using the distribution $f_i(st, dt)$.

C. Average detected individual versus time

One way to compare the method against the actual traces is to look at the evolution of expected number of detected individuals versus time. Recall that the total number of detected individual is denoted by S. We denote by $S(\tau)$ the total number of individuals detected by agents up to time τ: $S(\tau) = \sum_{i=1}^{N} \mathbb{1}_{\{t^i_{frst} \leq \tau\}}$. In particular, $S = S(T_{fest})$. The obtained value of $S(\tau)$ based on agents' traces is the dashed line plotted in Figure 3. We observe that $S(\tau)$ is zero before any agent enters the festival, and then rapidly grows.

Based on the model, we can estimate $\mathbb{E}[S(\tau)]$ as follows. By linearity of expectation we have $\mathbb{E}[S(\tau)] = N\mathbb{P}[t_{frst} \leq \tau]$, where $\mathbb{P}[t_{frst} \leq \tau]$ is the probability for an individual to be detected by at least one agent before time τ. For any value of $0 \leq \tau \leq T_{fest}$, the agents can be categorized into two types. Type I agents are those who enter the festival after time τ. These agents cannot detect any individual before time τ. Type II agents are the remaining agents who enter the festival before time τ. Type II agents can detect an individual before time τ relative to the duration of time they stay on the festival up to time τ. In fact by setting $dt^A = \tau$ for type II agents, we can use (3) to estimate $\mathbb{P}[t_{frst} \leq \tau]$. We do this for the three choices of $f(st, dt)$ and using the estimated α and β that are given in Table I. $\mathbb{E}[S(\tau)]$ is then equal to $\hat{N}\mathbb{P}[t_{frst} \leq \tau]$, where $\hat{N}$ is the estimated population size. The results are plotted in Figure 3. Note that the dashed line in Figure 3 is one realization of $S(\tau)$. However, the solid lines are expectations of $S(\tau)$ based on the model for three different $f(st, dt)$. We observe that the solid lines follow the dashed line closely, and that the model can predict the time evolution of $S(\tau)$. Similarly, by restricting agents' entrance and departure times to a particular time interval, it is straightforward to use the method to estimate the size of the population present at the festival at that time interval.

V. Conclusion & future work

In this paper we introduced a novel application that exploits the opportunistic contacts between mobile devices, namely, population size estimation by using mobile devices to sample a population. In order to test the feasibility of this method, we conducted an experiment at Paléo Music Festival. We derived a model to estimate the population of people that carry visible Bluetooth devices. We observed that the resulting estimate is surprisingly close to the ground truth, even with a small number of agents.

Furthermore, the model that we presented can be easily applied to specific parts of the collected data in order to obtain more specific estimates. For example, a simple extension allows to estimate the population size at different time intervals. We believe that similar extensions can be made to estimate the population size in different areas of the festival, provided that we include some information about the agent's location in the dataset. Although having an estimate for the number of attendees requires the knowledge of the ratio of visible Bluetooth devices, some population characteristics such as the relative density of attendees in different time periods or in different areas of the festival scale linearly with the size of the subset of visible Bluetooth devices. Therefore, the method can be used to study such population characteristics. Our future work will focus on the inclusion of location information and local estimates.

Acknowledgment

We thank Stéphane Szilagyi for his help in organizing the experiment at Paléo, Pascal Viot for providing us with an estimate of the total attendees of the festival, Julien Eberle for his help in programming the mobile phones, and the agents for participating in the experiment. This work was supported by the Nokia Research Center grant 'Accidental sampling'.

References

[1] R. Jose, N. Otero, S. Izadi, and R. Harper, "Instant places: Using bluetooth for situated interaction in public displays," *Pervasive Computing, IEEE*, vol. 7, no. 4, pp. 52–57, 2008.

[2] http://www.paleo.ch.

[3] J. Bunge and M. Fitzpatrick, "Estimating the number of species: A review," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. pp. 364–373, 1993.

[4] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3/4, pp. pp. 237–264, 1953.

[5] C. J. Schwarz and G. A. F. Seber, "Estimating animal abundance: Review iii," *Statistical Science*, vol. 14, no. 4, pp. 427–456, 1999.

[6] A. Chao, "An overview of closed capture-recapture models," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 6, no. 2, pp. pp. 158–175, 2001.

[7] S. Lee and A. Chao, "Estimating population size via sample coverage for closed capture-recapture models," *Biometrics*, vol. 50, no. 1, pp. pp. 88–97, 1994.

[8] A. Orlitsky, N. Santhanam, and K. Viswanathan, "Population estimation with performance guarantees," in *ISIT'07*, 2007, pp. 2026–2030.

[9] J. Acharya, A. Orlitsky, and S. Pan, "The maximum likelihood probability of unique-singleton, ternary, and length-7 patterns," in *ISIT'09*, pp. 1135–1139.

[10] G. M. Gemelos and T. Weissman, "On the entropy rate of pattern processes," *IEEE Trans. on Info. Theory*, vol. 52, no. 9, pp. 3994–4007, 2006.

[11] A. B. Wagner, P. Viswanath, and S. R. Kulkami, "A better good-turing estimator for sequence probabilities," in *ISIT'07*, 2007.

[12] A. Chao and J. Bunge, "Estimating the number of species in a stochastic abundance model," *Biometrics*, vol. 58, no. 3, pp. 531–539, 2002.

[13] J. Wang, "Estimating species richness by a poisson-compound gamma model," *Biometrika*, vol. 97, no. 3, pp. 727–740, 2010.