On Estimating the Number of Worldwide LTE Cell-IDs and WiFi APs

Species Richness Estimation Applied on Huge Datasets

Anna Lindgren[1] Bengt Lindoff[2]

[1] *Centre for Mathematical Sciences, Lund University*,
anna.lindgren@matstat.lu.se

[2] *Bengt Lindoff Innovation AB, Bjärred, bengt@lindoffinnovation.com*

December 16, 2018

1 Background — The Worldwide Success of Wireless Communications

Cellular wireless communication has been a tremendous success all over the world. From the start in small scale in the 1980ies with the 1G analog system, to the first worldwide success with the 2G GSM system during the 1990ies focusing on voice services. In the beginning of the 21th century 3G, High Speed Packet Access (HSPA) with focus on wireless data services was launched and during the last 5–10 years, 4G Long Term Evolution (LTE) optimized for mobile broadband services has been deployed. The 4G LTE together with the development of smartphones has made the wireless data services exploding all over the world the last 5–10 years as well. For instance, the estimated number of smartphones sold 2018 is 1.54 billion units, [6], and total number of cellular subscriptions is, in 2018, above 8 billion, i.e., more than one subscription per human in the world, see [4].

Also, other radio access technologies used in unlicensed frequency bands, such as WiFi, have become extremely popular. For instance, the estimated number of devices in use having WiFi capability as of today is more than 10 billion, see [7]!

From all the above numbers one can realize that the number of radio base stations (Cells) for serving the number of cellular devices as well as WiFi Access Points (AP) around the world must be huge.

2 Position applications based on Cellular Radio Base Stations and WiFi Access Points

Positioning applications, such as positioning applications developed by Combain, [3], utilizing the huge amount of stationary cellular radio base stations (or cells) and WiFi APs, have also been developed. Such applications can support wireless devices, such as Internet-of-Things devices, without support for Global Navigation Satellite Systems (GNSS), such as Global Positioning System (GPS), to give a position estimate of the device.

2.1 Basic principle of positioning

Each cellular radio base station, Cell, has a unique Cell Identity, Cell ID. The same holds for WiFi APs and hence, based on the ID the LTE Cell or WiFi AP is uniquely determined. Measurement Devices, such as smartphones with specific installed apps, report to a server the detected Cell IDs and/or WiFi APs together with signal strength measurement reports for the Cell ID or WiFi AP. Attached to the measurement report, also GNSS information is sent to the server.

Based on such measurement reports the Cell IDs and WiFi APs can be triangulated to determine their positions. Once the Cell ID or WiFi AP position is known (the Cell ID/AP has been "seen"), it is stored in a database on the server and can later on be used to triangulate other devices' positions (without inbuilt GNSS), based on measurement reports for detected Cell IDs/APs received from that device. Based on advanced positioning algorithms the position can be determined with an accuracy of 10–200 m depending on the number of detected Cell IDs and WiFi APs in the measurement report transmitted from the device.

3 Current number of Cell IDs and WiFi APs detected by Combain

Currently, more than 100 million different 2G, 3G and 4G Cell IDs and approximately 1.5 billion WiFi APs have been detected and stored in Combain's database (October 2018). Based on these numbers one could ask the following question; How many Cell IDs and WiFi APs are there around the world? Is it possible to estimate the total number of Cell IDs and WiFi APs based on the currently detected Cell IDs and WiFi APs?

In this paper we will try to answer the question of how many LTE Cell IDs and WiFi APs there are worldwide. In Section 4 we discuss how to solve the estimation problem and show that the solution can be found based on estimation principles found in biodiversity research, namely species richness estimation. In Section 5 we then discuss various classical estimation methods in the area of species richness estimation and in Section 6 we apply these estimation methods to the LTE Cell ID and WiFi AP data and give estimates of the number of LTE Cell IDs and WiFi APs worldwide. Finally, in Section 7 we draw some conclusions and we also discuss the difference between traditional species richness estimation applications and our use case, where the main and *significant* difference is the huge amount of data available in our application compared to traditional species richness estimation applications. Based on this difference, directions for future research in the area of species richness estimation in the case where a huge amount of data is available is discussed.

4 Species Richness Estimation

From a data set of different units detected (detected LTE Cell IDs or WiFi APs in our case) we want to determine the total number of units, including also the non-seen, yet-to-be-discovered units. Is this at all possible? Well, the problem formulation is common in the area of biodiversity research where the aim is to estimate the total amount of species in, say, a certain geographical area, based on catched or seen species in that area, see [2]. The typical approach is to collect a sample of individuals, identify their species and then count the number of individuals of each species in the sample, and from this data estimate the total number of species, both seen and unseen, in the underlying population. This is called the "species richness" estimation.

Hence, one approach to solve our estimation problem is to try to apply species

richness estimation methods on Combain's data set in order to estimate the total number of LTE Cell IDs and WiFi APs worldwide.

5 Classical Estimation Methods

Assume that we have made n observations and classified each observation according to the observed species. For each observed species we then note how many times that species was observed and, finally, calculate the number of species that were observed only once, that were observed exactly twice, thrice, etc. We thus have the absolute frequencies, N_k, as the number of species that were observed exactly k times, for $k = 1, 2, \ldots, n$. If N_n is larger than zero, we have only see one single species in all our n observations. Sometimes data is truncated at $k_{\max} < n$. In that case we let $N_{k_{\max}}$ denote the number of species that were seen *at least* $k_{\max}$ times.

The problem is to estimate the total number of species, m, which is the sum of the number of species we have detected, $m_D = \sum_{k=1}^{n} N_k$, and the unknown number of species we have *not* detected. Many of the classical estimation methods are based on dividing the species into rare species, i.e. species only seen a few times, and abundant species, i.e., species seen several times, and below we describe some of the most commonly used estimators for species richness estimation.

Chao-estimator

One simple method, developed in order to give a lower bound for the total number of species, m, uses the total number of species already detected, m_D, adjusted by the relation between how many have been found once, N_1, and how many have been found twice, N_2. If N_1 dominates over N_2 we are still finding new species and there are likely many more left to find. If N_2 is larger, we have already begun to find the more elusive ones for the second time and there should be fewer left undetected.

The Chao-estimate of m is given by

$$\hat{m}_{\text{Chao1}} \quad = \quad m_D + \frac{N_1^2}{2N_2}.$$

If N_2 is small, or zero, we should use a bias correction. If $N_1 = 0$ the estimate becomes $\hat{m} = m_D$, and we have (probably) detected all the species. Refering

to the division of species into rare and abundant, the Chao-estimator defines rare species as species seen, at most, twice. For further details, see [2].

Abundance-based Coverage Estimator

The division of species into rare, i.e., seen only once or twice, and abundant, seen more than twice, as used for the Chao-estimator, can be generalized using a cut-off value, τ. We then define the rare species as those seen at most τ times and the abundant ones as those seen more than τ times. Hence, for the Chao-estimator we had $\tau = 2$.

Dividing the sample into rare and abundant species we get the number of observations of rare species as $n_{\text{rare}} = \sum_{k=1}^{\tau} kN_k$, the number of detected rare species as $m_{D,\text{rare}} = \sum_{k=1}^{\tau} N_k$ and the number of detected abundant species as $m_{D,\text{abund}} = m_D - m_{D,\text{rare}}$.

Given the cut-off value, $\tau = 2, 3, \ldots, k_{\max}$, we can then define the Abundance-based Coverage Estimator (ACE) as

$$\hat{m}_{\text{ACE}} = m_{D,\text{abund}} + \frac{m_{D,\text{rare}}}{\hat{C}_{\text{rare}}} + \frac{N_1}{\hat{C}_{\text{rare}}}\hat{\gamma}^2_{\text{rare}}$$

where the coverage estimate, $\hat{C}_{\text{rare}} = 1 - N_1/n_{\text{rare}}$, is the proportion of the rare species observations that are of species seen more than once, and $\hat{\gamma}_{\text{rare}}$ is the coefficient of variation of the rare species' relative abundances, N_k.

The cut-off value τ has to be chosen. The default is to use $\tau = 10$, but for very heterogenous data the value

$$\tau = \max(10, \frac{n}{m_D})$$

has been suggested. In the case where data has been truncated at $k_{\max}$ and n is unknown, this is a lower limit of the suitable τ-value.

Replacing $\hat{\gamma}^2_{\text{rare}}$ by a bias-corrected version, $\bar{\gamma}^2_{\text{rare}}$, used when the coefficient of variation is large, gives the alternative estimator $\hat{m}_{\text{ACE-1}}$. For further details, see [2].

If $\hat{C}_{\text{rare}}$ is small, i.e., many of the observations of the rare species have been of new species, $\hat{m}_{\text{ACE}}$ and $\hat{m}_{\text{ACE-1}}$ will be larger, reflecting that there still seems to be many rare and undetected species left. If γ^2_{rare} is large, the N_k for the rare species are very different. Typically this means that N_1 is much larger than the others and, again, there are still rare and undetected species left to discover.

Jacknife estimator

The Jackknife technique is a way to reduce the bias in an estimate. A first-order Jackknife estimate is the average of the n different estimates we get when we delete each of the n observations in turn. In a second-order Jackknife estimate we delete all possible pairs of observations, etc.

When estimating the number of species, m, and the number of observations is large we get the first and second order jackknife estimates as

$$\begin{aligned} \hat{m}_{\text{J1}} &\approx m_D + N_1, \\ \hat{m}_{\text{J2}} &\approx m_D + 2N_1 - N_2. \end{aligned}$$

Higher-order jackknife estimates will include observations of more and more abundant species, but with lower weights, compared to the rarer species. Higher-order jackknife estimates give lower bias at the price of a higher variablility. Properties of the jackknife estimates, as well as a procedure for selecting the best jackknife order, can be found in [1].

6 Estimation of the total number of LTE Cell IDs and WiFi APs

In Combain's database as of October 2018, close to 13.5 million different LTE Cell IDs have been detected from at least 1 billion LTE Cell ID observations. Since the data has been truncated at $k_{\max} = 285$, the exact number of observations, n, is unknown and hence 1 billion is only a lower bound. Furthermore, almost 1.5 billion different WiFi APs have been collected from more than 16 billion WiFi AP observations. In this case, tha data has been truncated at $k_{\max} = 200$. The values for n, m_D, τ, N_1 and N_2 can be found in Table 1. The entire data set for all N_k for both WiFi and LTE can be downloaded from [5].

Suppose that there are m unique species, or to be more correct in this case, units, i.e., LTE Cell ID or WiFi AP, where m is unknown. Using the methods described in Chapter 5, assuming optimal choice of τ for ACE estimates, and first, second, and optimal order for the Jackknife estimator, we get the estimates of m presented in Table 2.

Since the number of observations is extremely large the confidence intervalls will be very narrow (the width is around 0.04 % of the estimated number of LTE Cell IDs and 0.006 % of the WiFi APs) and are therefore not presented.

Observations	LTE Cell ID	WiFi AP
Total number of observations, $n \geq$	1 081 744 199	16 076 788 205
Total number of units detected, m_D	13 489 923	1 469 912 771
Lower bound $\tau = \max(10, n/m_D)$	80	10
Number of units detected once, N_1	1 589 326	592 998 180
Number of units detected twice, N_2	732 622	178 443 239

Table 1: Summary of the number of observations.

Estimated number of units:	LTE Cell ID	WiFi AP
$\hat{m}_{\text{Chao1}}$	15 213 839	2 455 231 249
$\hat{m}_{\text{ACE}}$	15 655 112	2 298 673 354
$\hat{m}_{\text{ACE-1}}$	16 469 045	2 785 117 081
$\hat{m}_{\text{J1}}$	15 079 249	2 062 910 951
$\hat{m}_{\text{J2}}$	15 935 953	2 477 465 892
$\hat{m}_{\text{J9}}$ (best)	—	3 950 841 895

Table 2: Estimated number of units for the different estimators.

From Table 2 one can see that there should be at least 15.2 million LTE Cell IDs in the world. More accurate estimators indicate 15.6–16.5 million LTE Cell IDs. For the WiFi APs, there should be at least 2.5 billion APs and the more accurate estimators give estimates in the range of 2.8–4 billion.

Figure 1 shows the abundance data, N_k, for LTE Cell IDs and WiFi APs, as well as estimates, for different τ (i.e., rare/abundant cut-off, for the ACE estimators), as well as different Jackknife order estimates. It can be noted that for LTE, the Jacknife estimator diverges with increasing order, giving an unrealistic estimate of 60 million LTE Cell IDs for the 10:th order Jacknife estimate, while for WiFi APs the 9:th order Jackknife was determined to be the best Jacknife-estimate (i.e., the estimate does not diverge with increasing order). It can also be seen that the ACE estimates diverge for large τ in the WiFi case. However, the optimal cut-off for WiFi APs is at $\tau = 10$ so the estimates for larger τ are irrelevant.

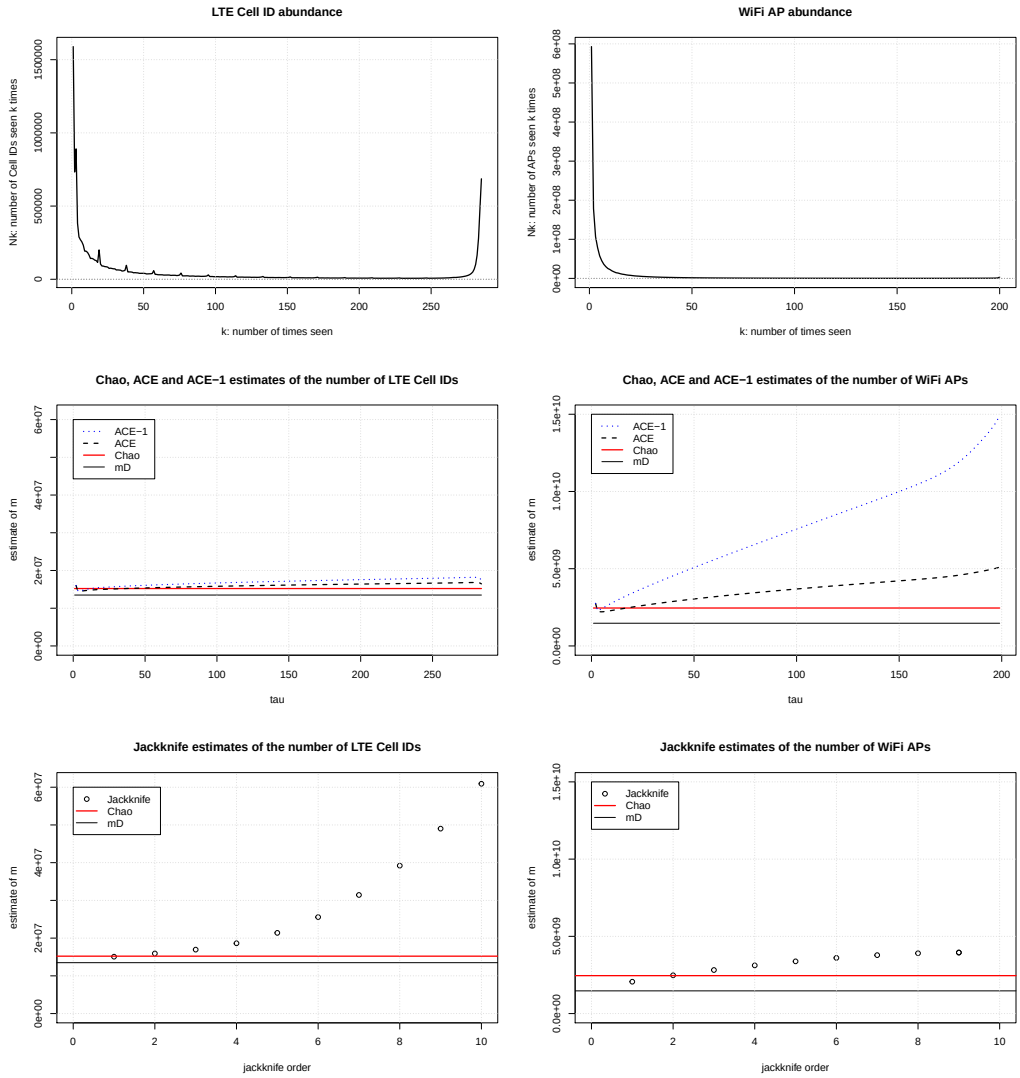

Figure 1: Plots over LTE and WiFi abundance data (top row), Chao and ACE estimates (middle row), and Jackknife estimates (bottom row).

7 Conclusions and Directions for Further Research

From the above results we can conclude that there are at least 15.2 million and, possible, as many as 16.5 million LTE Cell IDs, giving some 15–25 % more cells that remain to be found, and hence Combain's data set for LTE

Cell IDs is fairly complete. For the WiFi APs there are at least 2.5 billion, and, possibly, as many as 4 billion WiFi APs. Thus, Combain's current collection of WiFi APs is far from complete, which also explains the large differences between the different WiFi estimators.

As has been discussed in the paper, species richness estimation methods from the biodiversity research area can be applied to the problem of estimating the number of world-wide LTE Cell IDs and WiFi APs, based on Combain's data set. However, there are some significant differences between the data sets we analyzed here compared to classical species richness data sets, where the huge amount of observations in our data set is one of the differences. The huge amount of observations available opens up for other types of estimation approaches than the classical ones we used in this paper, mainly developed on small or moderate data sets.

For instance, we believe one can apply parametric models for the probabilities of observing individual units (LTE Cell ID/WiFi AP) based on the observed abundances, $N_1, N_2, \ldots, N_k$, and that way get an even better estimate of the total number of units, compared to the classical estimation approaches.

Furthermore, when the number of observations reaches 10–1000 millions, theoretical asymptotic results might be applied to determine, for instance, the remaining estimated number of observations needed before all units are observed.

Another difference between our data set and classical species richness data sets is the way the data is collected. In biodiversity research, the collection of observations are made on a dedicated basis (planned experiment), while in our case, the data are collected via crowd-sourcing over the entire world, from installed applications in smartphones. Hence, the data collection is not planned, since the data collection is done "in the background" and the observations happens to be where the smartphone happens to be, not where "unknown" LTE Cell IDs or WiFi APs are expected to be. This opens up for interesting data analyses on a per-regional basis, where the remaining number of yet-to-be-seen units may differs from, say country to country.

In short, the data set obtained by Combain opens up for a new field of research in the species richness area.

References

[1] K. P. Burnham and W. S. Overton. "Robust Estimation of Population Size When Capture Probabilities Vary Among Animals". In: *Ecology* 60.5 (Oct. 1979), pp. 927–936. DOI: 10.2307/1936861.

[2] Anne Chao and Chun-Huo Chiu. "Species Richness: Estimation and Comparison". In: Aug. 2016, pp. 1–26. ISBN: 9781118445112. DOI: 10.1002/9781118445112.stat03432.pub2.

[3] *Combain positioning solutions*. URL: https://combain.com.

[4] *Ericsson: mobility report*. 2017. URL: https://www.ericsson.com/en/mobility-report/reports/november-2017/mobile-subscriptions-worldwide-outlook.

[5] Anna Lindgren. *Downloadble data*. Centre for Mathematical Sciences, Lund University. 2018. URL: http://www.maths.lu.se/staff/anna-lindgren/downloads.

[6] *Smartphones industry: Statistics & Facts*. 2017. URL: https://www.statista.com/topics/840/smartphones.

[7] WiFi Alliance. *2018 WiFi predictions*. 2018. URL: https://www.wi-fi.org/news-events/newsroom/wi-fi-alliance-publishes-2018-wi-fi-predictions.