Estimating the number of LTE Cell IDs; a comparison of different species richness estimators

Milda Jokubauskaite

Abstract

Species richness estimation is an ongoing difficult statistical problem. Most of the research done in this subject focuses on biological data even though, a lot of the methods can be applied more widely. In this paper, we will focus on technological data and compare different known estimators (both, parametric and non-parametric) to see how well they work and compare in determining the total number of LTE Cell IDs (observed and unobserved) focusing on Northern Europe's region.

Contents

1. Data

The most widely used navigation system currently is GPS (Global Positioning System). It uses GPS satellites that collect and provide information to a GPS receiver on earth. One of the biggest drawbacks of this system is easy obstruction by mountains or buildings resulting in decreased accuracy. A company located in Lund, Sweden, Combain, tackled this problem and developed a global low-cost, high accuracy location system for M2M and IoT devices. The company is focusing on CELL ID and Wi-Fi technologies. They use Cell ID to locate any mobile device with a GSM/WCDMA/CDMA modem and have more than 106 million Cell IDs from more than 1000 operators in their database [8]. Combain themselves and their clients many times have asked about the total number of LTE cell IDs and WiFi APs and they approached the Lund University asking for help answering it. This project is looking at their LTE cell IDs database in order to help and answer this question. The data is frequency count data, which includes N_k = "number of cells that were observed exactly k times" and k = "frequency, i.e. number of times the cell was 'seen'". The data for the whole world is too big to handle with limited computation resources, therefore data was divided into individual countries databases and handled in smaller numbers. The main focus data was from Sweden.

1.1. Collection glitches and truncation

After plotting raw Swedish data (Figure 1), some obvious issues can be observed:

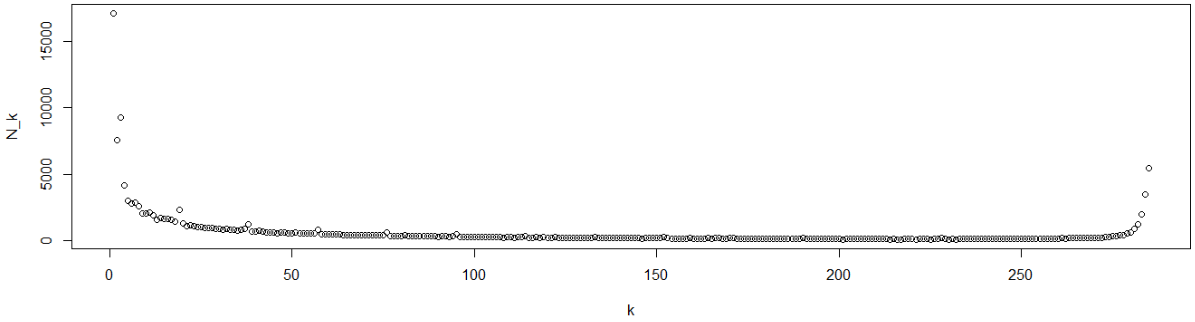


Figure 1: Raw Sweden's data.

- Due to the data collection system, observations above around $k = 280$ start to sum up and gives the 'tail' in the data plot, that can be corrected by truncating the data. In this project, the data was truncated from ($k_{max} = 285$ in Sweden's data) to $k = 280$ and therefore the total number of observed cells is not $N_{obs} = \sum_{k=1}^{k=k_{max}} N_k$ but instead $N_{obs} = \sum_{k=1}^{k=280} N_k$.
- Another issue with data collection can be noticed at $k = 3$. N_3 is suspected not to have its true value because of another collection issue. When device stays

for some time around a connection without WiFi, it registers a connection point multiple times and registration count of these points is set at most three. Because of multiple registrations from one connection, we assume that $N = 3$ is noted bigger. In order to negate this issue, we in this project we take $N_3 = \frac{N_2+N4}{2}$.

- Last artefact in the data can be seen at $k = n * 19 \quad n = 1, 2, 3, ...$ In the Figure 1 jumps in N_k can be seen in these points. It is once again due to the system's error when registering the collected data. Every time a cell is being registered $20th$ time, the system looks through the connections and throws away the one that had weakest signal. And the 20th count is then added to the $19th$ frequency, so therefore N_{20} should, in fact, be N_{21} and N_{19} is therefore equal to the sum of 'real' N_{19} and 'real' N_{20}. The same process is happening every $19th$ time, and jumps appear at $k = 19, 38, 57, ...$
  In order to correct this, my supervisor from Lund University wrote a smoothing R function. It was used on the data and can be found in Appendix A.

1.2. Data after corrections

After working out all the issues raw data set had, we can see the corrected data curve in Figure 2.

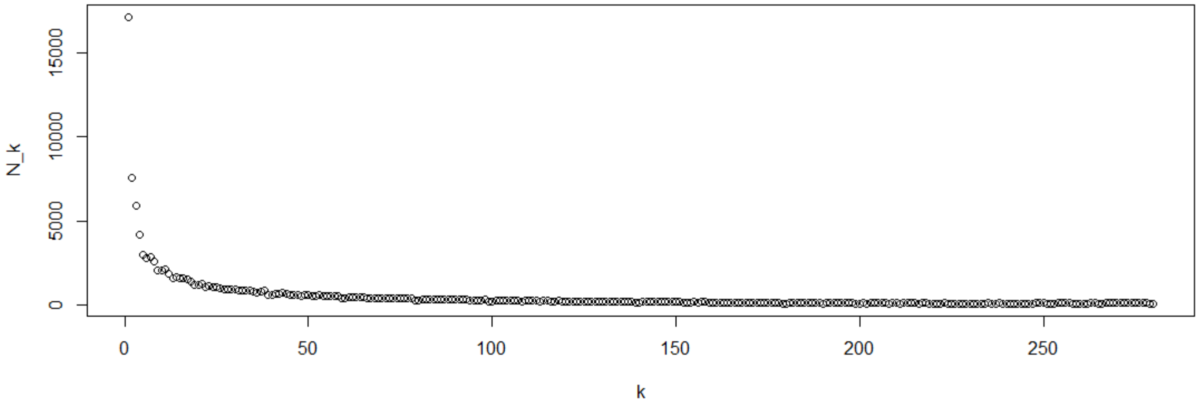


Figure 2: Corrected Sweden's data.

In the further text, the corrected Sweden's data will be referred to as smooth Sweden's data. It means that first, the raw data was smoothed using R function and as a result of this modification now has $k = 300$ registered frequencies. It still has the 'tail' from summed observations from $k = 280$, so it then is truncated for the analysis. These observations, that we know exists, just not at the corrected frequencies, will be added to the total in the end. Finally, artefact at $k = 3$ is also corrected and the data is ready for analysis.

After a quick calculation we set a starting point with total observations of $N_{obs} = 154179$ for $k = 1, ..., 280$.

2. Estimators

In general, there are two types of data, incidence (collected using multiple sampling units) and abundance data (single sampling unit). Although most richness estimators can be easily adapted to fit both types of data, we will focus on the estimators that work for our abundance data. Richness estimators are also typically divided into two types:

- **parametric estimators** work by fitting a function or mixture of functions curve to the known abundance distribution and estimating the missing piece of the curve, for $k = 0$. The most widely used distributions in this type are log-normal, exponential, geometric series, Gamma, negative binomial and inverse Gaussian [6]. These estimators can typically be time consuming and mathematically challenging, especially with bigger, more heterogeneous data sets.
- **non-parametric estimators** do not require any assumptions about the data and its distribution. They are widely used due to their wide range of applicability even on heterogeneous data sets. These estimators use only rare abundances to estimate the missing part of the species.

2.1. Parametric richness estimators using CatchAll

One of the best suited parametric modelling in our case would be mixed parametric modelling, in which several models would be combined to best fit the data. It is quite intuitive decision to use combination, in which one model would fit the rare data and the other would be used for abundant part, possibly adding even more to fit the data frequency curve better. One of the advantages to this method is that it can be both, visually and quantitatively, assessed. The downside of this method are long and difficult, computationally expensive, numerical methods to obtain the best fitted parameters and their maximum likelihood estimates. Some algorithms were developed such as EM (expectation-maximization) algorithm and it was further researched and improved by researches after.

This problem was tackled by John Bunge [1]. He did extensive testing on the methods, mainly using microbial ecology data, and developed a program that runs the data and decides on the best fitted mixed parametric model just in a couple of minutes. He created a model selection algorithm that uses criteria based on Akaike Information Criterion (AICe) [5], p-values from two Pearson χ^2 statistics [1] and SE (standard error of the estimate). His method automatically goes through all the possible lengths (τ) of the frequencies and its counts to determine the most optimal data set length to be used for estimation. The mixed-exponential models fit the data curve in the most accurate way possible and connect different exponential functions automatically, where the data curve's inclination changes to fit the next exponential function.

For our data set this program will be used to find the most suitable parametric model and assess its prediction on population (cell IDs) richness.

CatchAll also includes two non-parametric estimators: Chao1 and ACE. They will be used to check that programs results coincide with other non-parametric calculations. The estimators will be described in the next subsection.

2.2. Non-parametric richness estimators

For the non-parametric richness estimators we will be looking at three groups: ACE (ACE and ACE-1), Jackknife (1st and 2nd orders), and Chao (Chao1 and iChao1), with the specific focus on Chao estimators. We will note $\hat{S}$ to be the estimation of the total richness.

2.2.1. ACE estimators

Abundance-based Coverage Estimator (ACE) was first introduced by Chao and Lee(1994) [6] and it uses first ten frequencies of the data ($\tau = 10$) to determine the estimated species richness:

- ACE:

$$\hat{S}_{ACE} = \frac{s_{-}(\tau)}{1 - \frac{N_1}{n_{-}(\tau)}} + s_{+}(\tau) + \frac{N_1}{1 - \frac{N_1}{n_{-}(\tau)}} \cdot \gamma_{rare}^2 \tag{1}$$

where,

$$\gamma_{rare}^2 = max\left(\frac{s_{-}(\tau)}{1 - \frac{N_1}{n_{-}(\tau)}} \cdot \frac{\sum_{i=1}^{\tau} i \cdot (i-1) \cdot N_i}{n_{-}(\tau) \cdot (n_{-}(\tau) - 1)} - 1, 0\right)$$

where in the equations, $s_{-}(\tau)$ is the number of detected cells for rare population part, $s_{+}(\tau)$ is the number of detected cells for abundant part and $n_{-}(\tau) = \sum_{i=1}^{\tau} i \cdot N_i$. In our case, we use $\tau = 10$.

- ACE-1 (highly heterogeneous cases):

$$\hat{S}_{ACE-1} = \frac{s_{-}(\tau)}{1 - \frac{N_1}{n_{-}(\tau)}} + s_{+}(\tau) + \frac{N_1}{1 - \frac{N_1}{n_{-}(\tau)}} \cdot \gamma_{rare}^{\prime 2} \tag{2}$$

where,

$$\gamma_{rare}^{\prime 2} = max\left(\gamma_{rare}^2 \cdot \left(1 + \frac{\frac{N_1}{n_{-}(\tau)}}{1 - \frac{N_1}{n_{-}(\tau)}} \cdot \frac{\sum_{i=1}^{\tau} i \cdot (i-1) \cdot N_i}{n_{-}(\tau) * (n_{-}(\tau) - 1)}\right), 0\right)$$

This estimator is a bias-corrected version of ACE, when data is highly-diversified. CatchAll uses CV (coefficient of variation) to determine which of the ACE estimators is best suited for use. If CV_{rare} is ≤ 0.8 ACE is better suited, and if CV_{rare} is > 0.8 ACE-1 is in the run. The squared CV_{rare} is γ^2_{rare} and γ'^2_{rare} for ACE-1 respectively.

2.2.2. Jackknife estimators

- First-order jackknife estimator was developed by Burnham and Overton (1978) in order to reduce the bias of biased estimators. It uses only singletons (data with only one observation) to estimate the total richness:

$$\hat{S}_{jackknife1} = S_{obs} + N_1 \tag{3}$$

where S_{obs} is the total number of detected cells.

- Second-order jackknife estimator was derived by Smith and van Belle (1984) and uses both, singletons and doubletons (data observed once and twice) in order to give the total richness estimation:

$$\hat{S}_{jackknife2} = S_{obs} + 2 \cdot N_1 - N_2 \tag{4}$$

Higher orders of the jackknife estimators were developed by Burnham and Overton too, mainly focusing on capture-recapture problems and incidence data. Even though expressions for abundance data of the higher-orders were derived later too, but we will not focus our discussion on them here.

2.2.3. Chao estimators

Chao non-parametric estimators were developed as the lower bound of species richness estimation. It was first introduced by by Anne Chao in 1984 [7]. Chao1 stands for abundance data and Chao2 stands for incidence data. Correspondingly, iChao1 uses abundance data type and iChao2- incidence. Since our data is abundance frequency count, we will focus and discuss only Chao1 and iChao1 estimators, though they are easily derived for incidence data too.

- **Chao** proposed the estimate of the total species population through the estimation of $E(N_0)$, the expectation of the unobserved, by a non-parametric approach. She used the assumption that detection probabilities $(p_1, p_2, ..., p_S)$, with S different species in a community are fixed unknown parameters and therefore sample frequencies $(X_1, X_2, ..., X_S)$, where we set $k_{max} = n$, $(S_{obs} = \sum_{i=1}^{s} I(X_i > 0) = \sum_{k \geq 1} N_k)$ follows a binomial distribution [2]:

$$E(N_k) = E\left[\sum_{i=1}^{S} I(X_i = k)\right] = \sum_{i=1}^{S} \binom{n}{k} p_i^k (1 - p_i)^{n-k}, \quad k = 0, 1, 2, ..., n. \tag{5}$$

Based on the (5), Cauchy-Schwarz inequality shows:

$$\left[\sum_{i=1}^{S}(1-p_i)^n\right]\left[\sum_{i=1}^{S}p_i^2(1-p_i)^{n-2}\right] \geq \left[\sum_{i=1}^{S}p_i(1-p_i)^{n-1}\right]^2$$

A theoretical lower bound $E(N_0)$ is then derived as:

$$E(N_0) \geq \frac{n-1}{n}\frac{[E(N_1)]^2}{2E(N_2)}.$$

And the lower bound for species richness S is then:

$$\hat{S}_{Chao1} = E(S_{obs}) + E(N_0) \geq E(S_{obs}) + \frac{n-1}{n}\frac{[E(N_1)]^2}{2E(N_2)}. \tag{6}$$

After replacing the expected values in the (6) above with observed data, **Chao1** species lower bound estimator is then obtained:

$$\hat{S}_{Chao1} = S_{obs} + \frac{n-1}{n}\frac{N_1^2}{2N_2}$$

The estimator becomes unbiased if the data is perfectly homogeneous, (i.e. $p_1 = p_2 = ... = p_S$), but if the data has any heterogeneity to it, the estimator becomes accordingly biased. And the bias of this estimator is what leads us to the next segment.

- **iChao** total richness lower bound estimator was developed as a result for correcting the bias of Chao estimator using modified Good-Turing frequency formula (by Chun-Huo Chiu, Yi-Ting Wang, Bruno A. Walther and Anne Chao in 2014) [2]. They derived the approximation of the bias magnitude of the Chao1 estimator to be:

$$\left|bias(\hat{S}_{Chao1})\right| \approx \frac{1-\alpha_3}{\alpha_3}\left[\frac{1-\alpha_1}{\alpha_1} - \frac{1-\alpha_3}{\alpha_3}\right]\frac{2E(N_2)}{n(n-1)} \tag{7}$$

In which, α's come from the modified Good-Turing formula. Originally, Good and Turing defines for some data $\alpha_k = \sum_{i=1}^{S} p_i I(X_i = k)/N_k, \quad k = 0, 1, ...$ as the true mean relative abundance of those species that appeared k times in a sample of size n $(= k_{max})$. They state then, that $\alpha_k \quad k = 0, 1, ...$ is estimated by:

$$\tilde{\alpha}_k = \frac{(k+1)}{n}\frac{N_{k+1}}{N_k}, \quad k = 1, 2, ... \tag{8}$$

Good used a Bayesian approach to obtain (8), whereas Chun-Huo Chiu, Yi-Ting Wang, Bruno A. Walther and Anne Chao [2] used a more direct approach to derive the modified estimator of α_k:

$$\hat{\alpha_k} = \frac{(k+1)N_{k+1}}{(n-k)N_k + (k+1)N_{k+1}}, \quad k = 1, 2, ... \tag{9}$$

From there, they evaluate the magnitude associated with the Chao1 starting with the first-order bias magnitude from (6):

$$\begin{aligned} \left|bias(\hat{S}_{Chao1})\right| &= E(N_0) - \frac{(n-1)}{n}\frac{[E(N_1)]^2}{2E(N_2)} \\ &= \frac{E(N_0)(2E(N_2)/[n(n-1)]) - [E(N_1)/n]^2}{2E(N_2)/[n(n-1)]} \end{aligned} \tag{10}$$

They use the definition of α_k in the Good-Turing frequency formula and Cauchy-Schwarz inequality to separately approximate each term in the numerator of (10) (for details see [2]) and as a result get:

$$\left|bias(\hat{S}_{Chao1})\right| \approx \frac{1-\alpha_3}{\alpha_3}\left[\frac{1-\alpha_1}{\alpha_1} - \frac{1-\alpha_3}{\alpha_3}\right]\frac{2E(N_2)}{n(n-1)}. \tag{11}$$

Using α_1 and α_3 in (9) for (11) the lower bound of species richness is obtained:

$$\hat{S}_{Chao1} + \frac{(n-3)}{4n}\frac{N_3}{N_4} \times max\left(N_1 - \frac{(n-3)}{2(n-1)}\frac{N_2N_3}{N_4}, 0\right). \tag{12}$$

For big n in (12) $(n-3)/n$ and $(n-3)/(n-1)$ can be omitted and following improved lower bound of species richness estimator **iChao1** is obtained:

$$\hat{S}_{iChao1} = \hat{S}_{Chao1} + \frac{N_3}{4N_4} \times max\left(N_1 - \frac{N_2N_3}{2N_4}, 0\right). \tag{13}$$

3. Sweden's Analysis & Results

Sweden's data was analysed after smoothing all the artefacts described in Section 1. The analysis was done using CatchALL [1] and R package's SpadeR function ChaoSpecies described in details in [3]. Some of the non-parametric models are calculated in both sources and were used to confirm the consistency of the calculations. Table 1 shows that the best estimation is the Two-Mixed-Exponential parametric model composed by CatchAll [1]. All of the non-parametric models perform less, but considering that Chao estimators are derived as lower bound for the total richness, they are

Table 1: Sweden's data results (smooth data)

Nr. of observations		154179
Estimator	Estimation	Confidence Interval
Non-parametric:		
Chao1	173493	(172735,174282)
iChao1	177640	(177054,178241)
ACE	168465	(168027,168918)
ACE-1	172975	(172341,173632)
First-order jackknife	171297	(170938,171664)
Second-order jackknife	180829	(180208,181465)
Best parametric:		
Two-mixed-exponential	177760	(177190,178345)

expected to be lower than the best estimated value. iChao has a notable increase for a lower bound from original Chao estimator, the increase is due to the bias of the Chao estimator (2.937 % of the total observations) given by (7).

Note that the parametric model's and iChao1 estimators confidence intervals overlap considerably. Given that iChao1 is the lower bound estimation of the total richness, we can conclude that this parametric model gives very conservative estimation.

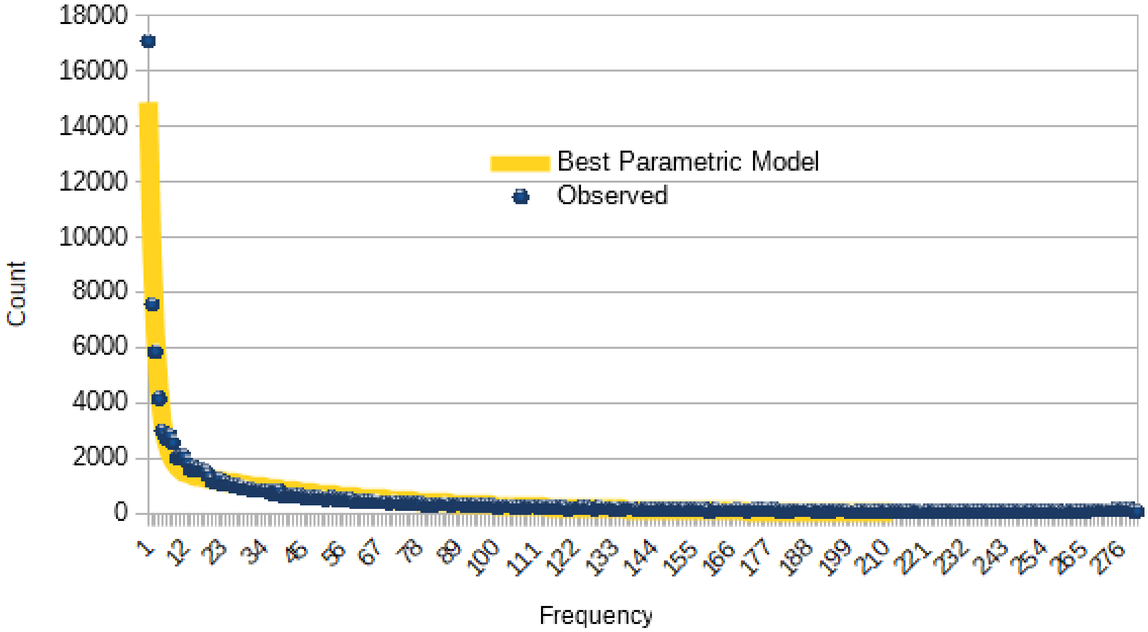


Figure 3: Parametric model's fitting to the smooth Sweden's data

In the Figure 3 is a plot of fitted parametric model that gave the best approximation to the observed data points. The two components, though not visible separately, are split to fit two parts of data with different declination angles. This particular parametric model uses 221 of the total 280 observations, which means that majority of all the available information had a role in total richness estimation.

I want to put a big emphasis in this paper about one of the main differences between the different estimator' groups (parametric and non-parametric). The parametric estimators, as mentioned in previous sections, are highly dependant on the underlying distribution of the data. As we can see from the Sweden's results above, parametric model is the one which gave the highest estimation with the highest confidence interval, therefore mostly over-performed in this particular case the non-parametric estimators, but the crucial condition for this performance was a smooth data curve, whose distribution can be traced. If the data quality or modifications to the original data were to be argued, the estimator's goodness could be argued as well, since it heavily relies on vast majority of data. Even through we trust our modifications and reasoning behind this smoothing of the data, I want to illustrate the sensitivity of the parametric model by simply recalculating the same estimators using raw Sweden's data with all its artefacts.

Table 2: Sweden's data results (raw data)

Nr. of observations		157544
Estimator	Estimation	Confidence Interval
	Non-parametric:	
Chao1	176858	(176100,177647)
iChao1	181684	(181164,182217)
ACE	171052	(170641,171476)
ACE-1	174796	(174224,175387)
First-order jackknife	174662	(174303,175029)
Second-order jackknife	184194	(183573,184830)
	Best parametric:	
Two-mixed-exponential	180146	(179636, 180669)

The changes from Table 1 to Table 2 are easily observed in all of the estimators and their performance, even though the same principal data was used. The overall increase in all of the non-parametric estimators is expected, because even after adding the truncated part's count, the total observations number is higher due to the remaining artefacts and all of those estimators consider that in their calculations. Due to the existing artefact at $k = 3$, the magnitude of Chao estimator's bias increases to 3.06 % and as a result improvement of the iChao estimator is even greater. The differences in the data cause some more changes to the non-parametric estimators, yet they are still consistent and intuitively predictable. The biggest change happens to the parametric model. CatchAll [1] model selection showed two-mixed-exponential as the best fitting model to our raw data curve. Figure 4 is visually eloquent of the changes in parametric modelling.

Even with the bigger number of the observations in raw data, parametric estimation of the total cell amount is has not increased as much (177760 with smooth data and 180146 with raw). It became even lower than the iChao1 lower bound estimation with the confidence intervals not even overlapping. It is due to the roughness of the data curve. Because of the artefacts, the curve has peaks at some points which makes it much harder to fit the parametric model's curve with high precision, and this loss in the precision causes drop in the estimation.

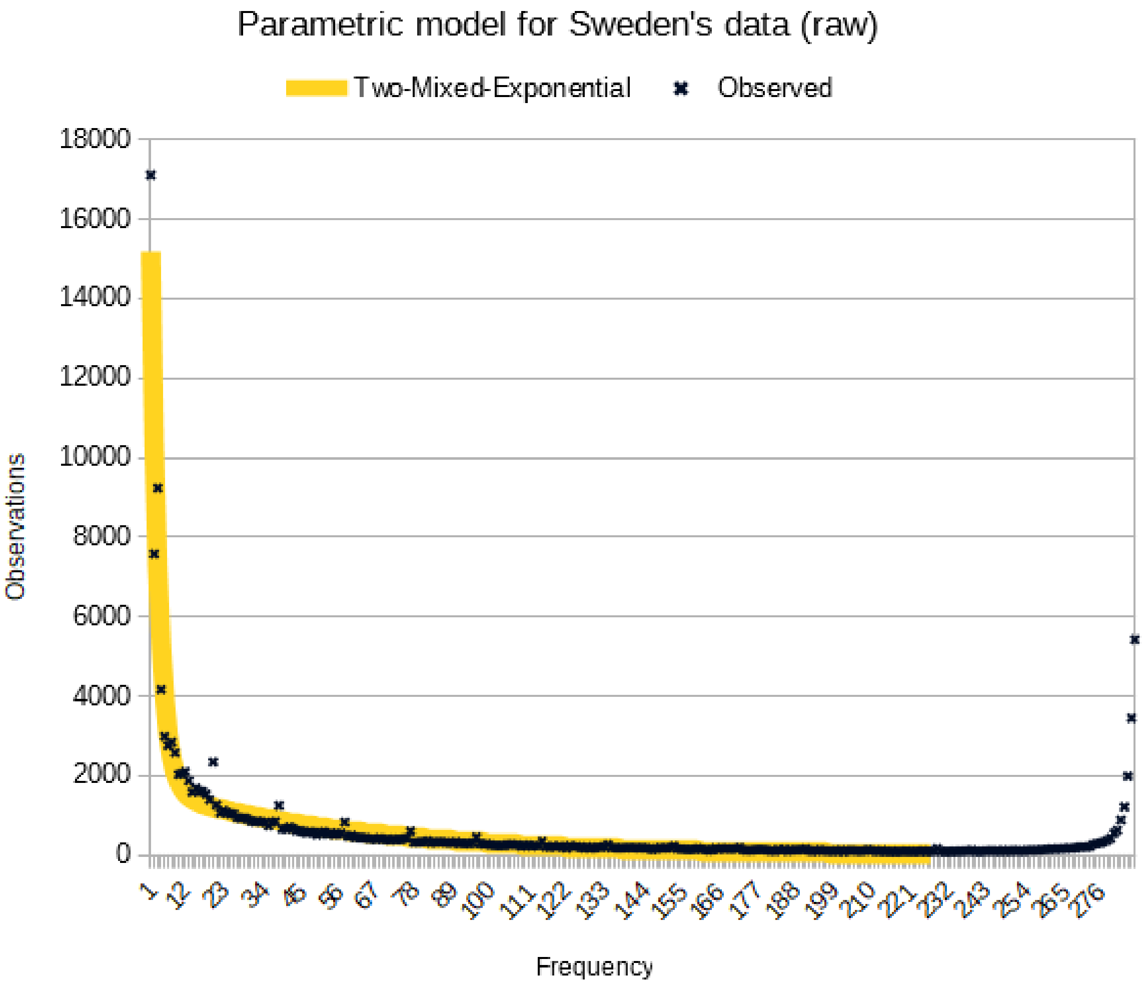


Figure 4: Parametric model's fitting to the data

This example illustrates that there is no uniformly best estimator, even if the data is from the same source. Given enough of information about collection errors that would cause the artefacts, we can apply corrections to the data, and get a set of information that is assumed to be close to the true values, hence get a better precision in estimation of the total amount. But we also need to keep in mind, that for all of those frequency points, where errors were found, there is no way of knowing the true value of the count, and in the smoothed data, those values are assumed. Before running any calculations, plotting the data curve and inspecting it visually can be best way to start, one can observe consistency and smoothness of the observation points and decide on the method of estimation knowing the qualities and information needed for them.

4. Nordics Analysis & Results

In addition to Sweden's, I have analysed and worked with other northern Europe's countries' data. It includes Lithuania's, Latvia's, Estonia's, Finland's, Denmark's, Norway's and Netherlands's data. To stay short, I will just call all of their data collectively, Nordic data. I have analysed it separately in both, before and after smoothing and removal of the artefacts. Each of these countries amount of total observations correlates well with the size and population of the countries. All of them, same as Sweden, are assumed to have very high collection rates, i.e. very close to all of the existing LTE cells are assumed to have been seen and therefore the data is of a high quality.

All the calculations done using the Nordic data were consistent with the discussion in Section 3. For the smoothed data, parametric modelling was producing the best estimator of the total richness, and was ill-fitting when the data had artefacts. Because of the reconcilable results, I will not be presenting and talking about each one separately, instead, I will look into all of the data in bulk, and also see if analysis of all the data at once is more beneficial than analysing one by one and summing the results.

4.1. Summed Data Analysis & Results

In the Table 3 can be found results from the summed Nordic data analysis. All of the countries observation count data was summed and the already described analysis was performed. For comparison reasons same analysis was done on both sets of data, for the raw untouched one and for the smoothed (cleaned from its artefacts) data. From the table several interesting things can be noted.

The most obvious happening in this table is the increase in every estimator for raw data due to the higher number of the total observations due to the described artefacts. The more interesting result in this table is in the parametric model's line. Its performance did not worsen even with the raw data set.

Table 3: Summed Nordic data results

Nr. of observations				
Smooth data: 586258			Raw data: 597346	
Estimator	Estimation		Confidence Interval	
Non-parametric:				
	Smooth data:	Raw data:	Smooth data:	Raw data:
Chao1	632470	643558	(631383,633584)	(642471,644672)
iChao1	643094	654889	(642279,643921)	(654180,655606)
ACE	621220	630656	(620573,621879)	(630044,631279)
ACE-1	629358	638296		
First-order jackknife	633531	644619	(632932,634138)	(644020,645226)
Second-order jackknife	656625	667713	(655589,657676)	(666677,668765)
Best parametric:				
	Smooth data:	Raw data:	Smooth data:	Raw data:
Two-mixed-exponential	649188	660038	(648286, 650103)	(659196,660891)

In order to see better the behaviour of the parametric models, plots of the summed Nordic and Sweden's observations and their corresponding parametric curves can be seen in Figure 5 (for the untouched, raw, data) and Figure 6 (for the corrected, smooth, data). Looking at these plots, it is easy to see why the parametric models performed similarly for both data sets. The two-mixed-exponential curve is split to fit two parts of the data, one is fitted for the low frequency, and the other is fitted for the high frequency part. Both curves are very similar and therefore both gave appropriate estimations. Due to the weight of the non-corrupted observation points in the raw data, artefacts did not have such a high influence on the result. When analysis is performed on this big of a set, parametric model is still able to fit the curve with fair accuracy.

Another thing to note, is that coefficients of variation (CV) for ACE estimators are smaller than 0.8, for both sets, and therefore data is not classified as highly diversified, hence ACE and not ACE-1 estimator is more suited to be used, i.e. confidence interval only for ACE estimator is relevant.

We can also note, that the only overlapping confidence intervals are between Chao1 and First-order jackknife estimators. This could happen if singletons (observations seen once) in the data are approximately the same as squared singletons divided by twice the doubletons (observations seen twice), in which case, formulas for these two estimators become very similar.

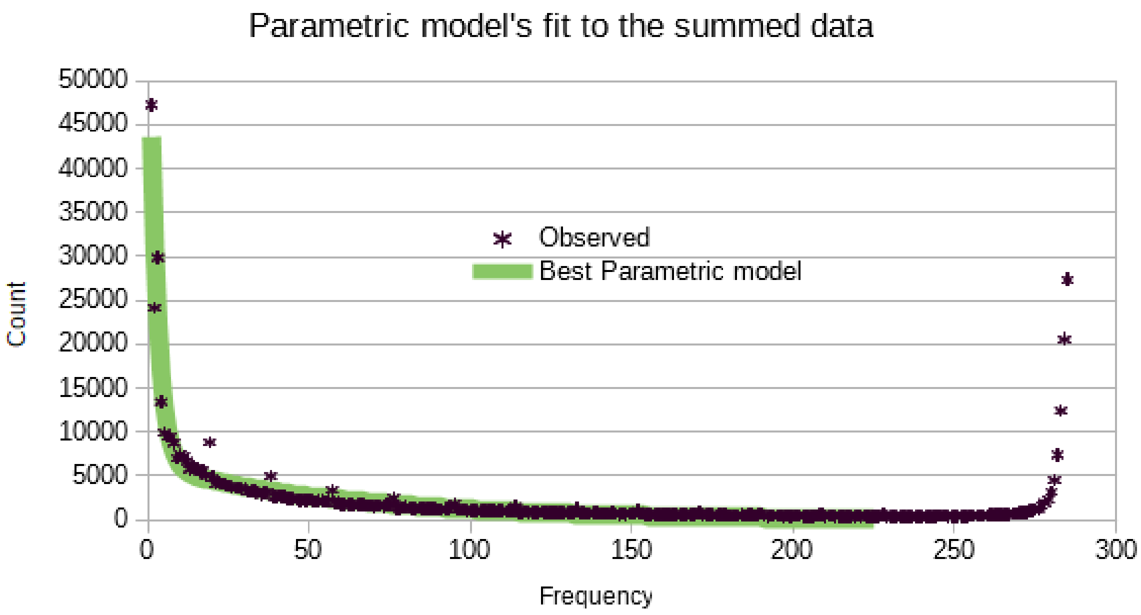


Figure 5: Parametric model's fitting to the raw data

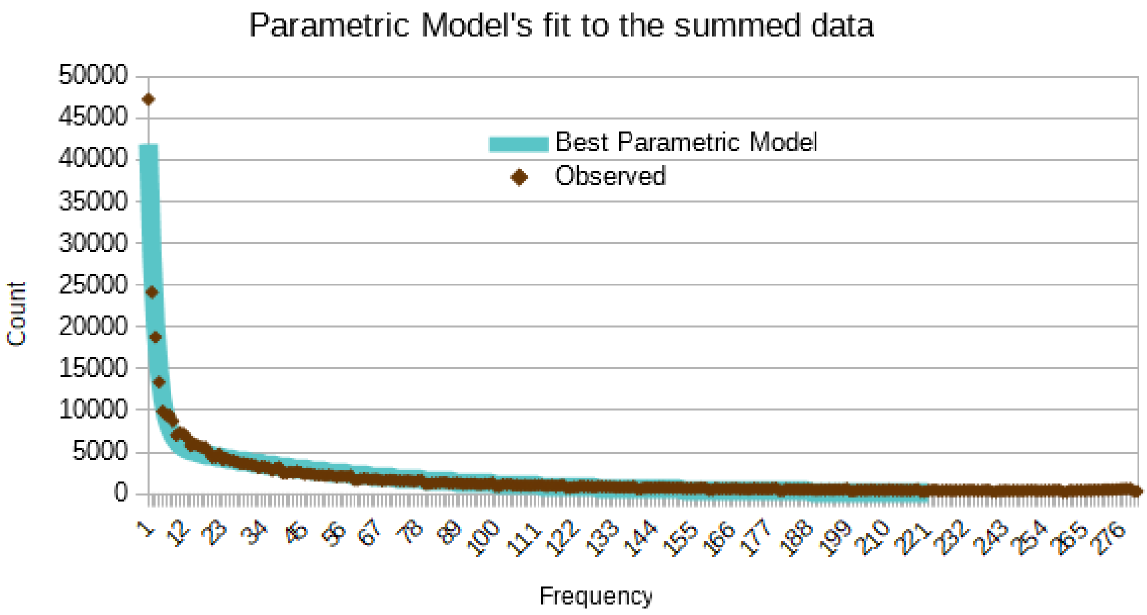


Figure 6: Parametric model's fitting to the smooth data

4.2. Separated Nordic Analysis

As mentioned before, the analysis was done on all of the Nordic countries separately and then also all together, with summed observations. And since the findings correlate well within all of the mentioned countries, I will not go into details of each analysis, but instead take a look at the final results and see which method gives a better (higher) estimation, the one where all the observations are summed and estimator is calculated for the whole data at once, or the one where estimators are calculated for each country separately and added in the end.

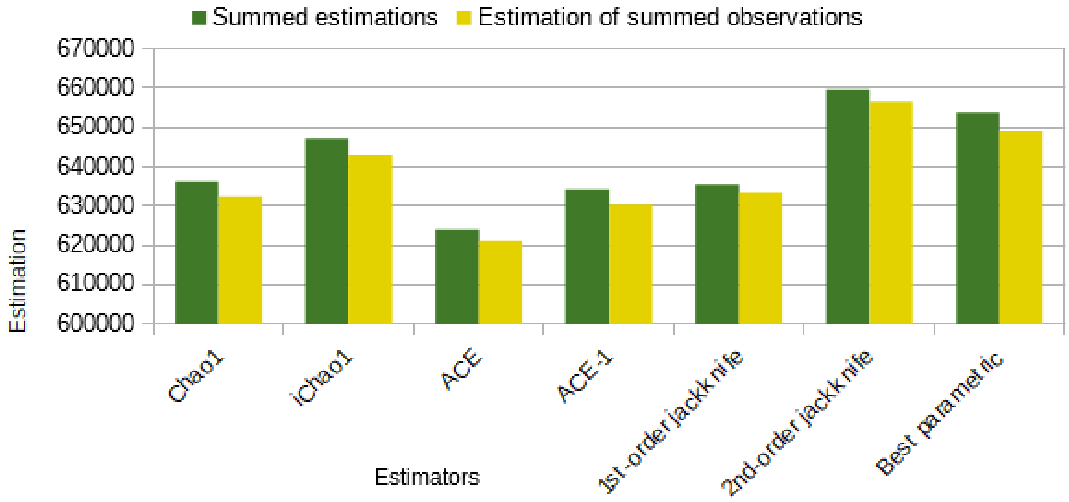


Figure 7: Differences in calculation order for smooth data

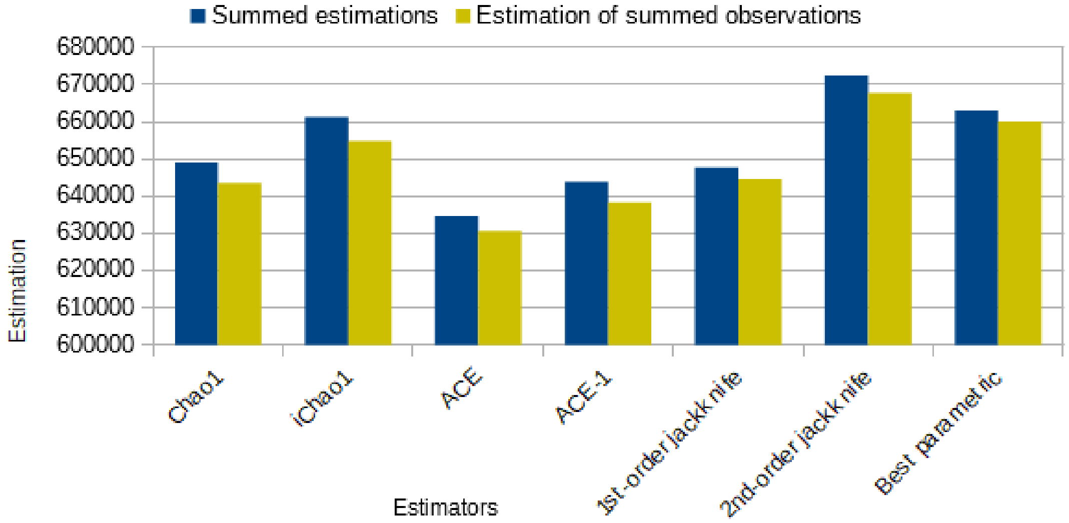


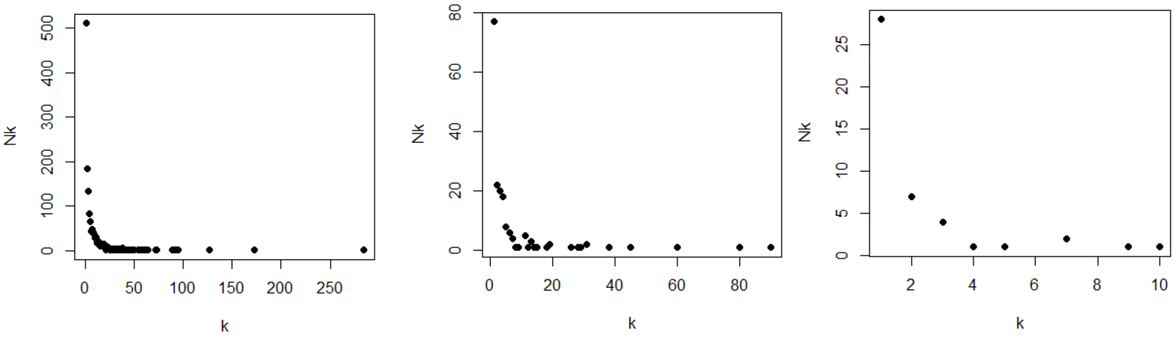
Figure 8: Differences in calculation order for raw data

Figure 7 (smooth data was used) and Figure 8 (raw data) shows the differences in the estimation for each estimator and the gap between methods. Darker columns indicate that the estimation was obtained by calculating estimators for each country individually and summing the estimations, the lighter columns indicate estimations calculated using the summed observations from the Nordics and Sweden. The result is uniform throughout all of the estimation methods and data types, summed estimations give a higher estimation for total population richness than estimation of the summed data.

Important to note here, is that the summed bias-corrected estimators iChao1 [2] give a lower bound for the countries higher (661352) than the best performed parametric model for the summed data (660038) for raw data. Even the smooth data's iChao1 estimator (647195) came very close to the best performed parametric model (649188) for summed observations analysis.

5. Poor data quality areas Analysis & Results

Even though analysis is more reliable using observations from well covered areas (such as Nordics and Sweden), I decided to take a look at the poor quality data regions to see how a pattern of estimators' performance, that we have seen in all of the analysis above, will change. Three countries with poor coverage and data quality we will be discussing here are Laos, Papua New Guinea and Syria. No corrections for the data were made, because artefacts are not easily deduced, when the quality is this poor. Overview of observations can be seen in data plots in Figure 9 and results of the data analysis are presented in Table 4.



(a) Laos **(b)** Papua New Guinea **(c)** Syria

Figure 9: Data plots

Table 4: Poor data quality analysis results

	Laos	Papua New Guinea	Syria
Observations	1405	181	45
Chao1	2108	316	101
iChao1	2255	334	114
ACE	1173	264	111
ACE-1	2264	295	176
1st-order jackknife	1915	258	73
2nd-order jackknife	2240	313	93
Best Parametric Model	3160	319	121

Some of the differences in analysis results from the well documented countries should be discussed. In Syria's data's analysis, ACE-1 estimation is notably the highest, and this estimator was never the highest for any other instance. This is a good example of why this corrected version of ACE was introduced. As described in Section 2.2.1, ACE-1 is advised for highly diversified populations, and due to the lack of measurements, the smallest differences in smoothness of the data curve are significant, making coefficient of variation to grow and ACE-1 estimator to estimate higher than any other estimator.

Laos data has the most observations, and the data curve has more smoothness to it, so it is not surprising, that parametric model could be fitted well and give the highest estimation.

As it can be spotted in Figure 9 (b), Papua New Guinea's data is the most scattered, therefore making it hard for parametric models to fit it accurately, hence bias-corrected Chao1 estimator iChao1 using only singletons and doubletons gave the highest estimation. Keeping in mind, that it was developed as the lower bound estimator, it certainly over-performed other estimators.

From this small testing on these varying data sets, the need for consistent, sufficiently large data sets should be noted. It is in order to draw evident conclusions about the underlying distribution for parametric modelling. Also to have good area coverage to draw conclusions about non-parametric estimation methods, that uses rare parts of the observation. This section shows that when the data becomes small and poor enough, one looses the consistency and therefore reliability of the results.

6. Summary and Conclusion

The three estimation methods that consistently provided highest estimation for our both, corrected (smooth) and raw data are two non-parametric: iChao1 (13), Second-order jackknife (4) and the parametric estimator. In most cases, parametric estimator was two-mixed exponential, fitting the data in two parts, one for low frequency, another for high frequency parts.

I want to begin by talking about the jackknife estimators. They over-perform any other estimates in most cases, but they should be treated with high caution because Chiu [2] (2014) conducted intensive testing and concluded that jackknife estimators underestimate in small data sets and overestimate true species richness in big data sets. They state that the threshold, where these estimators are accurate varies with the model and can not be predicted. They claim that this feature is what led to many researches to find jackknife estimators best-fitting. Our data sets, even when taken one country at the time can be considered big and therefore, the best performing estimator, second-order jackknife, should not be taken as the best true estimator for the total richness of LTE Cell IDs.

The final conclusion lies between the bias corrected iChao1 estimator [2] and parametric model [1]. The main advantage of the parametric modelling is its result, in most cases it gives higher estimation than the iChao1 and uses majority of the data to find it. But it is also a very sensible method to be used, because for this method to perform well, the data curve has to be smooth enough in order to fit the parametric distribution to it. In order to achieve that, one should have sufficient information about the data and its collection methods and be able to use that information to correct the artefacts. In our case, all this was known and we were able to assume a smooth enough data curve for the parametric model. However, iChao1 estimator was developed as the lower bound estimator of the total richness, hence it serves that purpose impeccably. It puts the lower bound estimation above many that estimate the total richness. In addition, it is not computationally expensive and works well even with the artefacts that do not affect rare group or total count of the observations.

7. Further research

This project has a massive potential to be taken over and further researched. Due to the computational power limits, the whole world's data was not investigated. Statistical methods are also not widely explored for this type of abundance data without underlying distribution and further research on the estimators for such big abundance frequency count data could be carried out.

8. Appendix A

R function written by the supervisor, that was used to smooth the the artefacts at every 19th frequency point due to registration of the collected data errors:

```
avskrynkla <- function(skrynk, p = 0.5) {
        flat <- skrynk
        flat <- data.frame(k = seq(1, 300))
        flat$Nk <- NA
        for (k in seq(1, 15)) {
                flat$Nk[19*(k - 1) + (k - 1) + seq(1, 19)] <-
                        skrynk$Nk[19*(k - 1) + seq(1, 19)]
                if (k < 15) {
                        flat$Nk[19*(k - 1) + (k - 1) + 19] <-
                                ceiling(p*skrynk$Nk[19*(k - 1) + 19])
                        flat$Nk[19*(k - 1) + (k - 1) + 20] <-
                                floor((1 - p)*skrynk$Nk[19*(k - 1) + 19])
                }
        }
        return(flat)
}
```

References

[1] John Bunge. *Estimating the number of species with CatchAll.* 2012 Apr 1; 28(7): 10451047.

[2] Chun-Huo Chiu, Yi-Ting Wang, Bruno A. Walther and Anne Chao. (2014) *An Improved Nonparametric Lower Bound of Species Richness via a Modified Good-Turing Frequency Formula.* Biom, 70: 671-682. doi:10.1111/biom.12200

[3] Chao, A. and Chiu, C. (2016) *Species Richness: Estimation and Comparison.* In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J. L. Teugels).

[4] Nicholas J. Gotelli and Robert K. Colwell *Estimating species richness* Chapter4

[5] Trevor Hastie, Robert Tibshirani, Jerome Friedman *The Elements of Statistical Learning; Data mining, Inference and Prediction*

[6] Chao, A. and Lee, Sh. *Estimating Population Size Via Sample Coverage for Closed Capture-Recapture Models* Biometrics, Volume 50, Issue 1 (Mar., 1994), 88-97

[7] Anne Chao *Nonparametric Estimation of the Number of Classes in a Population* Scand J Statist 11:265-270, 1984

[8] https://combain.com/