Dealing with Data Evolution and Data Integration: An approach using Rarefaction

Luca Del Core, Eugenio Montini, Clelia Di Serio, Andrea Calabria

Abstract Heterogeneity and unreliability of data negatively influence the effectiveness and reproducibility of the results in all fields involving sampling techniques. Heterogeneity is mainly due to technological advances which imply improvements in measurements resolution. Unreliability or under-representativeness in data may be due to machine/software or human variances/errors, or other unidentifiable external factors. In the era of big data, technological evolution, and continuous data integration, scientists are increasingly facing with the problems of how to (1) identify and filter-out unreliable data, and (2) harmonize samples gauged with different platforms improved over time. This work is aimed at developing a new statistical framework to address both issues, showing results in real case scenarios.

Luca Del Core
San Raffaele Telethon Institute for Gene Therapy (SR-Tiget),
IRCCS San Raffaele Scientific Institute,
Via Olgettina 58, 20132, Milano, Italy
e-mail: delcore.luca@hsr.it

Eugenio Montini
San Raffaele Telethon Institute for Gene Therapy (SR-Tiget),
IRCCS San Raffaele Scientific Institute,
Via Olgettina 58, 20132, Milano, Italy
e-mail: montini.eugenio@hsr.it

Clelia Di Serio
University Vita-Salute San Raffaele,
University Centre of Statistics in the Biomedical Sciences,
Via Olgettina 58, 20132, Milano, Italy
e-mail: diserio.clelia@unisr.it

Andrea Calabria
San Raffaele Telethon Institute for Gene Therapy (SR-Tiget),
IRCCS San Raffaele Scientific Institute,
Via Olgettina 58, 20132, Milano, Italy
e-mail: calabria.andrea@hsr.it

Key words: Heterogeneity, Rarefaction, Filtering, Abundance Models, Species Pooling, Generalized Nonlinear Models, Hurlbert-Heck Model, Entropy, Integration Sites, Gene Therapy.

Motivation and background

Nowadays, with the advent of high throughput technologies, new statistical challenges are aimed at combining large heterogeneous datasets produced under different platforms having different efficiency, reliability and resolution. This is the case of biomedical data, where the follow-ups of a clinical cohort of patients under treatment may last over several decades and the monitoring of patient health-care must benefit by the biotechnological improvements continuously consolidated. Therefore, new statistical methods are required to understand how to consider, within a unique framework, time series that are observed over a long period, and how to distinguish whether the increased number of events is attributable to the change in technology rather than to the disease change itself. Thus, it is crucial to obtain reliable and harmonized time-course data, addressing the problems of (1) the identification and filtering of unreliable data, and (2) how to scale heterogeneous integrated data to obtain consistent results in the application domain. This work is proposed as a first step towards a mathematical/statistical solution of both problems.

Material and Methods

The filtering problem is addressed through the expected richness estimation via the Hurlbert-Heck (HH) curve[1, 2] and the Species Pooling (SP) methods[3, 4, 5] for an estimation of the unseen species. The base statistical methods have been previously applied in ecological and population-based studies. We exploited the Generalized Nonlinear Model (GNM) as an estimator of the HH curve properly rescaled. Then, using an empirical approach, we identified a minimum threshold for the richness over the whole cohort of data to filter out under-representative observations. To address the problem of data integration for reproducible results over continuous technological improvements and the scaling problem, we used a Rarefaction Method[1, 2].Both methodologies have been applied in biomedical science using molecular data (retroviral vector integration sites, IS) of Gene Therapy (GT) clinical trials, a good case study for the presence of heterogeneous data. The filtering technique is used basing on the richness in distinct number of ISs. The rarefaction approach allowed improving data integration of IS by rescaling data in order to obtain rarefied population measures (such as entropy indexes[6, 7]) that are more robust and homogeneous than the un-scaled ones, thus potentially improving the assessment of safety and long term efficacy of the treatment. A discussion of results is finally presented.

Results

Filtering Unreliable Data

Dealing with biological and molecular data, such as IS in GT studies, means dealing with high variability in data collection and sampling, due to the high variability of the available biological material (for example the different amount of DNA used in each test). Thus, the number of retrieved IS from each patient at different time points (IS_s Richness) may vary. Therefore, we have to evaluate the level of richness for each sample and filter-out those samples with an insufficient level of IS richness. To overcome this problem, the percentage $S_{\%}$ of richness in IS_s observed over the total can be defined as the ratio between the observed richness S and an estimator of the whole theoretical richness $\hat{S}_{tot}$, namely $S_{\%} = S/\hat{S}_{tot}$. The theoretical richness $\hat{S}_{tot}$ can be estimated in different ways, depending on the chosen Sampling Pooling (SP) technique[3, 4, 5]. In this work, the SP estimator is chosen based on the best Ranked Abundance Distribution (RAD)[8, 9] together with a p-leave out cross validation. Namely, if the general lognormal (gln) curve is chosen as the best AIC_c[10] RAD model among the candidates[1], or if, according to a χ^2 Goodness of Fit test, the gln distribution can be used in place of the optimal one, then the Preston estimator $\hat{S}^{tot}_{Preston}$[3] is used. Otherwise the performance of Chao and ACE estimators is compared[11] via a uniform leave-$p = .3$-out cross validation: the whole sample is considered as the sampling universe with a known total richness S_{obs} and the $[0,1]$-bounded quantity[2]

$$A^{est}_{absolute} = \min\{S_{obs}, \hat{S}_{est}\} / \max\{S_{obs}, \hat{S}_{est}\}$$

is used to compare the performance of the two non-parametric estimators. The more the accuracy of the estimator, the greater $A^{est}_{absolute}$ is: therefore, the estimator $best = argmin_{est}\{A^{est}_{absolute}\}$ is chosen if the gln distribution is rejected as RAD model. Also, in order to assess the accuracy of the species pooling estimator chosen among the candidates, during the $nFld = 100$ p-leave-out cross simulations, three additional accuracy indexes are calculated and compared with $A^{best}_{absolute}$. These are defined as

$$A_{effective} = 1 - |.7 - S^{0.7}_{obs}/\hat{S}^{0.7}_{tot}| \qquad A_{relative} = 1 - |S^{0.7}_{obs}/\hat{S}^{0.7}_{tot} - S^{1}_{obs}/\hat{S}^{1}_{tot}|$$

$$A_{cumulative} = \min\{\hat{S}^{0.7}_{tot}, \hat{S}^{1}_{tot}\} / \max\{\hat{S}^{0.7}_{tot}, \hat{S}^{1}_{tot}\}$$

where $S^{0.7}_{obs}$, $\hat{S}^{0.7}_{tot}$, S^{1}_{obs}, $\hat{S}^{1}_{tot}$ are the observed and estimated (using the chosen estimator) total richness in the 70% fold and in the whole sample respectively. By definition, they are $[0,1]$ bounded and the more the accuracy, the greater they are. As an

[1] Geometric Series, MacArthur's Broken Stick, Zipf-Mandelbrodt, Zipf, General lognormal are the candidate RAD models in the case study. All these distributions are fitted with the Maximum Likelihood Estimation (MLE) technique.

[2] $\hat{S}_{est}$ is the estimation of the total richness S_{obs} obtained using the the estimator $est \in \{Chao, ACE\}$ using the 70-random subsample.

example, these indexes are calculated in the case study, together with $A^{est}_{absolute}$, and the results are shown in Fig.1(a). Furthermore, in order to check robustness with respect to little variations, the model selection and inference are performed on a fixed number *nRnd* of binomial randomizations of the original data.

Therefore, a set of $S_{\%}$'s is collected among all the samples, and a minimum threshold for this quantity is identified to filter out under-representative observations. Without any ground-truth available, we used an empirical approach in which the shape of the $S_{\%}$'s empirical cumulative distribution (eCDF) is analyzed, and we selected as threshold (if existing) the main concave/convex inflection point over the eCDF curve[3], which could be interpreted as a signal of multimodal distribution. Then, a Generalized Pareto Distribution (GPD) is fitted on the excesses above that threshold (POT method[13]) and a QQ-plot between the empirical and GPD quantiles could provide feedbacks on the quality of the fitting such that an higher quality corresponds to a better overlap of the curve to the main diagonal. This method was applied in our case study and the results are shown in Fig.1(b,c).

Another interesting question to deal with is the saturation problem: finding the total abundance Ab_{tot} needed to reach a certain percentage level p of richness, say the 90%. For this reason, the Hurlbert-Heck (HH) rarefaction curve[1, 2] $E(S)$ is calculated, over a properly chosen grid of rarefaction levels of total abundance, as a pointwise estimation of the expected richness associated with each rarefaction level. Then a family of generalized nonlinear models defined by a log-linear mixture regression function $E(Y) = g^{-1}(\alpha log(X) + \beta X)$ and binomial distribution for the response variable with *probit*, *logit*, *cloglog* as candidate link functions g are applied on the percentage ratio $S^{\%} = E(S)/\hat{S}_{tot}$ over Ab_{tot}, and the one with the maximum R^2 index is chosen. Also, in order to calculate the total abundance $Ab^{p\%}_{tot}$ associated with a certain percentage level p of richness, the regression function was inverted via the numerical resolution of the Lambert function $W(z)e^{W(z)}, z \in \mathbb{C}$. In Fig.1(d) a graphical representation of the percentage HH curve and its regression estimator are shown for one sample of the case study.

Scaling of the Heterogeneous Data

Another problem regarding the IS_s data being analyzed in the case study concerns the reliability of data interpretation due to the different orders of magnitude in total abundance Ab_{tot}, and indeed in richness S_{obs}, of IS_s reached in each sample mainly due to the variation in resolution of the gauge instruments adopted during time.

Therefore, in order to compare any measure of safety obtained in each sample (e.g. an entropy index), the whole cohort of data should be first rescaled to the same magnitude level of total abundance. In this work, the minimum total abundance level $Ab_{raremax} = \min_{samples} Ab_{tot}$ among the samples is used as rarefaction level.

[3] The eCDF inflection points are found via the Extremum Distance Estimator (EDE) algorithm [12].

Then, a rarefaction technique[2], which essentially consists in random subsampling with proportional abundances $p_i = Ab_i/Ab_{tot}$ as probability weights with Ab_i the abundance of the i-th species in the original sample, is applied in order to generate a rarefied version of that sample. This results in a more homogeneous pool of samples which can be used for entropy measures comparison during time of therapy. As an example, the Renyi Entropy Spectre (RES)[7] is calculated on the IS_s data of the case study, before and after rarefaction. The results are shown in Fig.1(e,f).

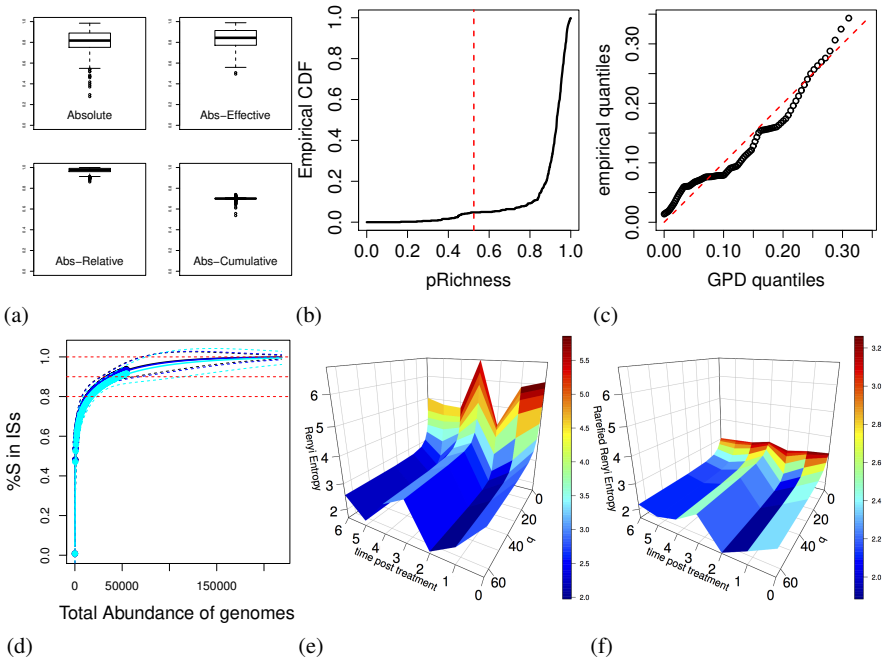

Fig. 1: (**a**) From top-left, the boxplot of the Absolute, Effective, Relative and Cumulative accuracy indexes are shown respectively. (**b**) The empirical CDF associated with the collected sample of $S^{\%}$. The vertical red highlighted line represents the threshold $S^{\%}_{thr}$ estimated via the EDE algorithm. (**c**) The scatterplot between the empirical quantiles associated with the excesses below that threshold and the GPD quantiles fitted on the same quantities. (**d**) The Hurlbert-Heck curve rescaled by the sampling pooling species estimators $S^{tot}_{Preston}$, S^{tot}_{Chao}, S^{tot}_{ACE} are drawn in black, blue and cyan. The observed and predicted ratio $S^{\%} = S_{obs}/S^{tot}$ are respectively represented by the thick and thin lines. The 80%, 90% and 100% richness thresholds are also shown as red dotted lines. This figure is related to a single sample. (**e**) The Renyi Entropy Spectre is shown during time of therapy on the heterogeneous (un-rarefied) and (**f**) homogeneous (rarefied) samples respectively.

Discussion and Conclusion

As a result of the filtering method, some of the correlations expected to be biologically relevant (e.g. between DNA nanograms DNA_{ng} and total abundance Ab_{tot} in IS_s) slightly increased, suggesting to further characterize the discarded samples.

Moreover, the comparison between the non-rarefied and rarefied entropy curves shows the positive effect in data harmonization by reducing the fluctuations in results due to change in sampling technology. These findings shed lights on reliability and reproducibility in continuous data integration over improvements in technological changes, critical challenges in the era of big data and improvements in high-throughput technologies. Species diversity could also be better addressed using the Renyi Entropy Spectre, where the effect of most abundant clones is visible at higher levels of q.

In conclusion, this work provided new methods to address the data integration and rescaling from technological sources continuously evolving and the problem of filtering unreliable data. Both problems approach the reproducibility of results in science even over time, and data accuracy and reliability.

References

1. S. H. Hurlbert, "The Nonconcept of Species Diversity: A Critique and Alternative Parameters," *Ecology*, vol. 52, no. 4, pp. 577–586, 1971.
2. K. L. Heck, G. van Belle, and D. Simberloff, "Explicit Calculation of the Rarefaction Diversity Measurement and the Determination of Sufficient Sample Size," *Ecology*, vol. 56, no. 6, pp. 1459–1461, 1975.
3. F. W. Preston, "The Commonness, And Rarity, of Species," *Ecology*, vol. 29, no. 3, pp. 254–283, 1948.
4. A. Chao, "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability," *Biometrics*, vol. 43, no. 4, p. 783, 1987.
5. R. B. O'Hara, "Species richness estimators: How many species can dance on the head of a pin?," *Journal of Animal Ecology*, vol. 74, no. 2, pp. 375–386, 2005.
6. T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition*. 2006.
7. J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. No. XIV, 2010.
8. R. H. Whittaker, "Dominance and diversity in land plant communities," *Science*, vol. 147, no. 3655, pp. 250–260, 1965.
9. J. B. Wilson, "Methods for fitting dominance / diversity curves," *Journal of Vegetation Science*, vol. 2, no. 1, pp. 35–46, 1991.
10. K. P. Burnham and D. R. Anderson, "Multimodel inference: Understanding AIC and BIC in model selection," 2004.
11. A. E. Magurran and B. J. McGill, "Biological diversity: frontiers in measurement and assessment," *Challenges*, p. 368, 2011.
12. D. T. Christopoulos, "Developing methods for identifying the inflection point of a convex/concave curve," pp. 1–29, 2012.
13. G. Salvadori, C. De Michele, N. T. Kottegoda, and R. Rosso, *Extremes in Nature: An approach using Copulas*. 2005.