# List of figures

[**Figure 1** Freshwater lakes from southern Norway and Sweden sampled for the COMSAT project. Secchi depth is displayed to provide a general impression of the longitudinal gradient in the dataset. 2](#_Toc27444547)

[**Table 1** Subsets of the ASV and metadata tables. The columns to the right show the first 5 ASVs in decreasing order of abundance while the columns to the left show linearly independent metadata variables that can be treated as explanatory to bacterial community composition. 2](#_Toc27444548)

# Introduction

## Chosen dataset

This study was performed on freshwater microbial ecology data generated by sampling 72 lakes from southern Norway and Sweden (fig. 1). It was part of a project designated COMSAT. The dataset comprises two tables. One consists of counts of amplicon sequence variants (ASV) for bacteria, while the other contains environmental metadata (table 1). Each observation in either table corresponds to a lake. Both tables are matched by observation and can thus be used as input and output for each other. Bacterial ASVs can be treated as a proxy for the abundance of bacterial species. These ASV counts can also be converted to binary with 0 equal to 0 and values above 0 set to 1 in order to study presence/absence patterns.



**Figure 1** Freshwater lakes from southern Norway and Sweden sampled for the COMSAT project. Secchi depth is displayed to provide a general impression of the longitudinal gradient in the dataset.

**Table 1** Subsets of the ASV and metadata tables. The columns to the right show the first 5 ASVs in decreasing order of abundance while the columns to the left show linearly independent metadata variables that can be treated as explanatory to bacterial community composition.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Site | ASV1 | ASV2 | ASV3 | ASV4 | ASV5 | … | Latitude | Longitude | Altitude | Area | Depth | Temperature | Secchi | O2 | CH4 | pH | TIC | SiO2 | KdPAR |
| 10000\_Hurdalsjøen | 18464 | 5231 | 6963 | 7563 | 9516 | … | 60.37648 | 11.04077 | 176 | 32.81 | 20.0 | 17.03 | 6.50 | 0.9044194 | 11.797343 | 6.870 | 0.82230 | 3312 | 0.62 |
| 10001\_Harestuvatnet | 15296 | 58728 | 30659 | 1614 | 17059 | … | 60.19323 | 10.71212 | 234 | 1.98 | 13.0 | 15.85 | 4.50 | 0.8468347 | 72.674567 | 7.365 | 4.05800 | 3783 | 0.89 |
| 170B\_Gjersjøen | 13356 | 52215 | 25810 | 1367 | 14586 | … | 59.78970 | 10.77485 | 40 | 2.64 | 22.0 | 19.65 | 3.30 | 0.8131012 | 52.953904 | 7.685 | 8.08500 | 3563 | 0.95 |
| 170\_Gjersjøen | 16227 | 53747 | 26456 | 2823 | 3119 | … | 59.78970 | 10.77485 | 40 | 2.64 | 22.0 | 19.65 | 3.30 | 0.8131012 | 52.953904 | 7.685 | 8.08500 | 3563 | 0.95 |
| 180\_Øgderen | 52862 | 4887 | 1361 | 14854 | 25616 | … | 59.71388 | 11.41303 | 133 | 12.66 | 9.5 | 18.61 | 1.10 | 0.8406025 | 85.639780 | 7.225 | 2.66800 | 1125 | 1.60 |
| 189\_Krøderen | 18830 | 53461 | 50015 | 12664 | 13253 | … | 60.13485 | 9.75860 | 133 | 43.91 | 14.0 | 15.44 | 2.80 | 0.8582522 | 29.100059 | 6.695 | 0.81360 | 2499 | 0.82 |
| 191\_Rødbyvatnet | 43828 | 7657 | 1836 | 34800 | 20517 | … | 59.58175 | 10.48715 | 118 | 1.16 | 10.0 | 18.55 | 2.10 | 0.8527711 | 260.596931 | 7.535 | 3.09200 | 2063 | 1.32 |
| 214\_Gjesåssjøen | 10532 | 588 | 9275 | 19181 | 5315 | … | 60.68167 | 11.99235 | 176 | 3.98 | 3.5 | 19.63 | 1.15 | 0.8360833 | 97.561306 | 7.070 | 1.73200 | 2924 | 2.27 |
| 2252\_Rotnessjøen | 14088 | 39265 | 35086 | 11061 | 7228 | … | 60.49690 | 12.34120 | 260 | 1.09 | 26.0 | 16.55 | 1.95 | 0.7350632 | 41.956068 | 6.635 | 0.77310 | 5559 | 1.08 |

# Methods

## Data filtering, formatting and transformation

The credit card dataset contained only continuous variables. There were missing values in the environmental metadata, which were replaced by intrapolation using Multivariate Imputation by Chained Equations (Buuren & Groothuis-Oudshoorn 2011). This decision to keep observations with missing values is motivated by the very small number of observations in the full dataset compared to the number of descriptors which would make it even harder to train models successfully, were the number of observations to be reduced any further. The ASV table was scaled using ranging (set values to interval of 0-1) due to this transformation yielding somewhat better clustering in terms of ecological meaningfulness compared to subtraction of mean and division by standard deviation (Legendre & Legendre 2012). Environmental metadata were used with or without scaling, the former being performed by subtracting the mean and dividing by standard deviation.

# References

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations inR. Journal of Statistical Software, 45(3). https://doi.org/10.18637/jss.v045.i03

Legendre, P., & Legendre, L. F. (2012). Numerical ecology(Vol. 24). Elsevier