COPPE
Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia
UFRJ

MODELLING THE INFLUENCE OF ENVIRONMENTAL PARAMETERS ON THE MARINE MICROBIAL ABUNDANCE IN ABROLHOS BANK USING RANDOM FOREST AND BOOST REGRESSION TREE

Reza Amir Ahmadi

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientadores: Laura Silvia Bahiense da Silva
Leite
Diogo Antonio Tschoeke

Rio de Janeiro
Novembro de 2021

MODELLING THE INFLUENCE OF ENVIRONMENTAL PARAMETERS ON THE MARINE MICROBIAL ABUNDANCE IN ABROLHOS BANK USING RANDOM FOREST AND BOOST REGRESSION TREE

Reza Amir Ahmadi

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Orientadores: Laura Silvia Bahiense da Silva Leite
Diogo Antonio Tschoeke

Aprovada por: Prof. Laura Silvia Bahiense da Silva Leite
Prof. Diogo Antonio Tschoeke
Prof. Fabiano Lopes Thompson
Prof. Priscila Machado Vieira Lima

RIO DE JANEIRO, RJ - BRASIL
NOVEMBRO DE 2021

Ahmadi, Reza Amir

Modelling the Influence of Environmental Parameters on the Marine Microbial Abundance in Abrolhos Bank using Random Forest and Boost Regression Tree / Reza Amir Ahmadi. – Rio de Janeiro: UFRJ/COPPE, 2021.

VIII, 149 p.: il.; 29,7 cm.

Orientadores: Laura Silvia Bahiense da Silva Leite
Diogo Antonio Tschoeke

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2021.

Referências Bibliográficas: p. 143-147.

1. Marine microbial abundance. 2. Abrolhos bank. 3. Influence of environmental parameters. I. Leite, Laura Silvia Bahiense da Silva *et al.* II. UniversidadeFederal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

MODELANDO A INFLUÊNCIA DE PARÂMETROS AMBIENTAIS NA ABUNDÂNCIA MICROBIANA MARÍTIMA DO BANCO DE ABROLHOS USANDO FLORESTA RANDÔMICA E ÁRVORES DE REGRESSÃO DE CRESCIMENTO PROGRESSIVO

Reza Amir Ahmadi

Novembro/2021

Orientadores: Laura Silvia Bahiense da Silva Leite
Diogo Antonio Tschoeke

Programa: Engenharia de Sistemas e Computação

O Banco de Abrolhos é uma extensão da plataforma continental do leste brasileiro localizada no sul do estado da Bahia, Brasil, formando os maiores e mais ricos recifes de coral do Atlântico Sul. O objetivo deste estudo consistiu em modelar a influência de parâmetros ambientais na abundância microbiana dos recifes de coral de Abrolhos usando Random Forest e Boost Regression Tree. Nossos resultados mostraram que carbono orgânico dissolvido, nitrogênio total e silicato são os fatores mais importantes que regulam a abundância microbiana. A hidrodinâmica e a temperatura também influenciam. Os arcos internos de Abrolhos, principalmente Pedra de Leste, apresentaram baixa hidrodinâmica, maior tempo de residência da água, maior concentração de nutrientes, maior proliferação de micróbios e possivelmente mais doenças de coral. Já os arcos externos de Abrolhos, principalmente Arquipélago, apresentaram alta hidrodinâmica - favorecendo a "lavagem" dos recifes, menor temperatura, menor concentração de nutrientes e menor proliferação de micróbios. Esses achados podem fornecer subsídios importantes para uma melhor gestão ambiental do Banco de Abrolhos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

MODELLING THE INFLUENCE OF ENVIRONMENTAL PARAMETERS ON THE MARINE MICROBIAL ABUNDANCE IN ABROLHOS BANK USING RANDOM FOREST AND BOOST REGRESSION TREE

Reza Amir Ahmadi

November/2021

Advisors: Laura Silvia Bahiense da Silva Leite
Diogo Antonio Tschoeke

Department: Systems Engineering and Computer Science

The Abrolhos Bank is an extension of the eastern Brazilian continental shelf located in the south of Bahia State, Brazil, comprising the largest and richest coral reefs of the South Atlantic. The aim of this study was to model the influence of environmental parameters on the marine microbial abundance of the Abrolhos coral reefs using Random Forest and Boost Regression Tree. Our findings showed that dissolved organic carbon, total nitrogen and silicate are the most important factors regulating the microbial abundance. The hydrodynamics and temperature also have influence. The internal arcs of Abrolhos, specially Pedra de Leste, presented low hydrodynamics, longer residence time of the water, higher concentration of nutrients, greater proliferation of microbes and possibly more coral disease. While the external arcs of Abrolhos, specially Arquipélago, presented high hydrodynamics - favoring the "washing" of the reefs, lower temperature, lower concentration of nutrients, and minor proliferation of microbes. These findings can provide important insights for a better environmental management of the Abrolhos Bank.

Acknowledgments

I would like to thank my supervisory team that has been there for me throughout all times. Thank you, Prof. Laura Silvia Bahiense da Silva Leite, Prof. Diogo Antonio Tschoeke and my wife, Dr. Tooba Varasteh, for your limitless support. I can’t fully express how much help you three have been to me and how much you have aided in my growth through these years.

Several people have given me intellectual support and have worked so hard with me. I can’t thank enough, Prof. Fabiano Lopes Thompson, Prof. Cristiane Thompson, Prof. Priscila Machado Vieira Lima, Gutierrez, Rosa koko Otsuki, and Angelica.

And finally, thank to my friends, family, Federal University of Rio de Janeiro (UFRJ), Post-Graduate Program in department of Systems Engineering and Computer Science, Brazil and Brazilian people that have been my emotional support and making my MSc a fun experience.

I dedicate this dissertation to my wife Tooba.

Contents

1. Introduction

Coral reefs are threatened worldwide, with both global changes and local impacts playing important roles in accelerated reef degradation (Bruce et al., 2012). The Abrolhos Bank on the eastern Brazilian continental shelf is recognized as the largest and richest coral reef system in the South Atlantic (Francini-Filho et al., 2013). Coral disease and massive declines in coral cover have recently occurred in the Abrolhos Bank (Francini-Filho et al., 2008). The Abrolhos Bank is an extension of the eastern Brazilian continental shelf (approximately 46,000 km2) located in the south of Bahia State, Brazil. The Abrolhos Bank comprises the largest and richest reefs of the South Atlantic, with at least 20 species of coral, including 6 that are endemic to Brazil (Leão et al., 2003). The coral reefs in Abrolhos are distributed into two arcs almost parallel to the mainland shore. The Internal Arc is located from about 10 to 20 km off the cost, and is formed by a complex of bank reefs and isolated coral pinnacles of varied dimensions. The Outer Arc, which borders the east side of the Abrolhos Islands, is formed by isolated giant coral pinnacles, located circa 70 km from the coast (Leão et al., 2003).

Microbial abundance is potentially important as food for filter-feeding fauna. There is evidence that physico-chemical variables influence the abundance of microbial (Goulder, 1980; Milner & Goulder, 1986; Morikawa, 1984). In response to different environments, microbes reflect differently in population size, distribution, and physiological state and cultivability. The use of statistical learning methods is necessary to analyze the new data collected from Abrolhos environment with physical-chemical parameters that modulate the abundance of microbial in the water of Abrolhos coral reefs. Harmful effects of eutrophication and fishing are interconnected and cause serious damage to reef biomes (Bell 2008). The absorption of organic matter by microbial is a major route of carbon flux, and its variability can change the overall patterns of carbon flow (Azam 1998). The dissolved organic carbon (DOC) released by algae beside other nutrients may promote the microbial growth that promote the death of the coral (Smith *et al.*, 2006). Therefore, it is necessary to develop a lighter prediction model, such as an empirical approach, for predicting microbial abundances in particular.

Regression analysis is a statistical technique for estimating the relationship among variables which have reason and result relation. Main focus of univariate regression is analyze the relationship between a dependent variable and one independent variable and formulates the linear relation equation between dependent and independent variable. Regression models with one dependent variable and more than one independent variables are called multilinear regression (Uyanık and Güler 2013).

The Boost Regression Tree (BRT) and Random Forest (RF) models are two relatively new tree-based models that have been developed to optimize predictive performance by combining a large number of simple trees into a powerful model rather than using a single tree model based on traditional regression trees (Breiman, 2001; Skurichina and Duin, 2002; Friedman, 2001, 2002). In the BRT model, the fitted model is a simple linear combination of many trees that are fitted iteratively and boosted to reweight poorly modeled observations (Elith *et al.*, 2008). The RF model is constructed in a random vector of the data feature space sampled independently (Breiman, 2001). Being data mining methods, the BRT and RF models have several common advantages, including a limited number of user-defined parameters and the ability to model non-linear relationships, manage qualitative and quantitative variables, remain robust despite missing data and outliers, reduce overfitting, and evaluate, summarize and interpret final models (Breiman, 2001; Friedman and Meulman, 2003). Owing to these merits, BRT and RF models have been widely applied in various scientific fields, including ecological modeling (Peters *et al.*, 2008; T. Froeschke and F. Froeschke, 2011).

1.1 Objective

The aim of this study was to model the influence of environmental parameters on marine microbial abundance using the BRT and RF regression models, in Abrolhos bank, Brazil. Six sites were selected for this study. The three outer reefs (External Arc) within the no-take area of the National Marine Park of Abrolhos (NMPA) included in this study (Parcel dos Abrolhos, Mato Verde and Arquipelago, hereafter, PAB, MV, and AR, respectively) are protected. The three inner reefs (Internal Arc) (Timbebas, Pedra de leste, andSebastião Gomes, hereafter, TIM, PL and SG, respectively) are unprotected (Francini- Filho and de Moura 2008). Spatial management through implementations of the NMPA can be considered a large-scale ecological experiment that can provide important insightsinto ecosystem functioning and management success (Knowlton and Jackson, 2008).

We evaluated the performance of and differences between the BRT and RF models in mapping the variability of environmental parameters on marine microbial abundance of the Abrolhos Bank. The specific objectives of this research were to:

- Apply Data Exploratory Analysis to the data set.
- Develop BRT and RF models to predict the microbial abundance content based on 544 samples and environmental variables (Biochemical and Biophysical parameters).
- Quantify the effects of various environmental variables on the microbial abundance variation.
- Compare the predictive qualities of the BRT and RF models.
- Compare the difference of biophysical (Temperature and Hydrodynamic velocity), biochemical parameters and microbial abundance in Internal and External arc of Abrolhos.

We evaluated this hypothesis for influence of environmental parameters on marine microbial abundance of the Abrolhos Bank: less hydrodynamics in the reefs of the internal arc of Abrolhos (TIM, PL and SG) has a longer residence time of the water, a higher concentration of nutrients, greater proliferation of microbes and possibly more coral disease. On the other side external arc of Abrolhos (MV, PAB and AR) favors "washing" of the reefs by the greater hydrodynamics, lower temperature, a lower concentration of nutrients, and minor proliferation of microbes.

2. Materials and Methods

2.1. Study Area

Six sites between 13 and 90 km off the coast were selected for this study (Fig. 1). The seawater samples were obtained in the inner reefs of SG (17°54′42.49″ S 39°7′45.94″ W), PL (39º2’00”W 17º46’00”S) and TIM (17°27'57.1"S 39°01'00.5"W) and in the outer reefs, PAB (18°00'52.4"S 38°40'00.6"W), MV (18°01'59.8"S 38°39'60.0"W) and AR (17°57'50.5"S 38°42'03.7"W). PAB, MV and AR are completely within NMPA, and enforcement is performed by the Brazilian Environmental Agency (ICMBio). The water sample collection were performed in years 2011-2016 (Except 2015). PAB also has unique coral reef structures known as Chapeiro˜es (mushroom-like structures). The three inner reefs (TIM, PL and SG) are unprotected and heavily fished. Sampling in six years and in different locations allowed us to determine the temporal and spatial variations in water quality and microbial diversity. The seawater samples were collected close (<1 m) to the reef structures at a depth of between 10 and 15 m at the SG, PL and TIM, and at 20 m at PAB, MV, and AR Reefs.

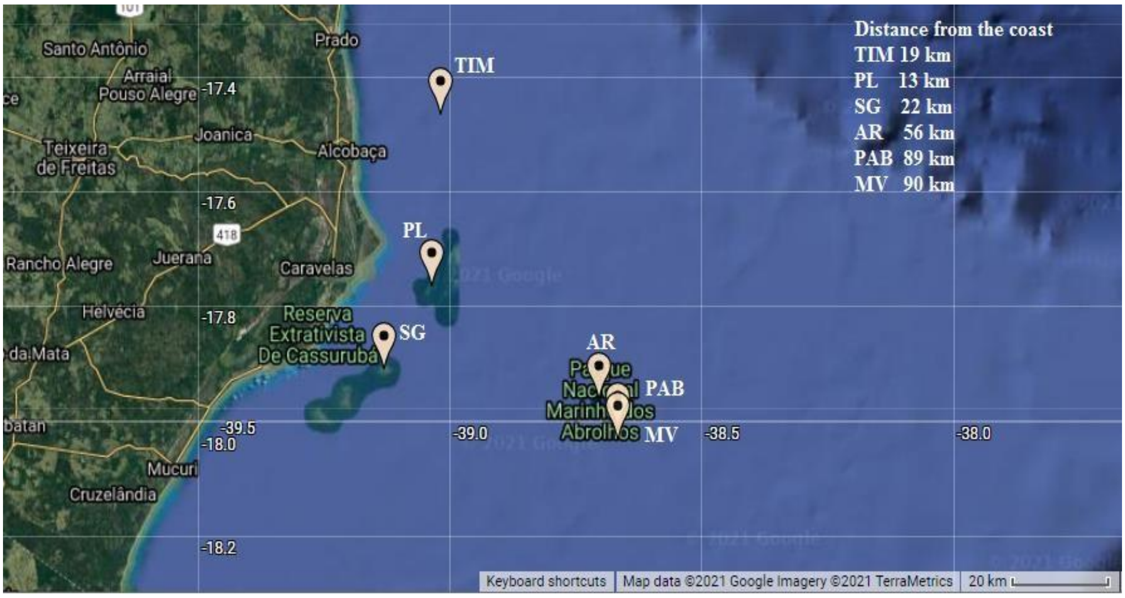


Fig. 1 Study area. SG, PL and TIM are located in internal arc of Abrolhos. PAB, MV and AR are located in external arc of Abrolhos. The distances from the coast are presented in the in the upper right frame, blank on the map (https://maps.co/).

2.2. Sampling and Data Collection

The variables studied in this work was presented in Fig. 2. 544 water samples were assessed for levels of biophysical and biochemical parameters by using oceanographic methods previously described by Grasshoff *et al.* (2009). There are three generic variables year, site, and season with 18 biological variables, Dissolved Organic Carbon (DOC), Orthophosphate, Total Phosphate (TP), NH3, Silicate, Nitrit, Nitrate, Total Nitrogen (TN) are independent variables and abundance of Chlorophyll a (Cl-a), Pheophytin, Bacterial, High nucleotide (HNA), Low nucleotide (LNA), Prochlorococcus, Synerochocos, Picoeukaryote, Nanoeukaryote and Virus are dependent variables (Fig. 2). The temperature data during years 2011 to 2016 and hydrodynamic velocity data only for 2010 among Internal and External arcs are available. We measured the water temperature (Temp) using a CTD device. Hydrodynamic data was determined with magnitude of depth averaged velocity (m/s). Concentrations of Total Phosphorus (TP) were determined by acid digestion to phosphate, and concentrations of Total Nitrogen (TN) were determined by digestion with potassium per sulfate following nitrate determination. Biological parameters were also measured for each sample. Microbial abundance was measured using flow cytometry (Flux) as described by Cabral *et al.* (2017). Chlorophyll-a analyses were performed as described by Coutinho *et al.*, 2019.

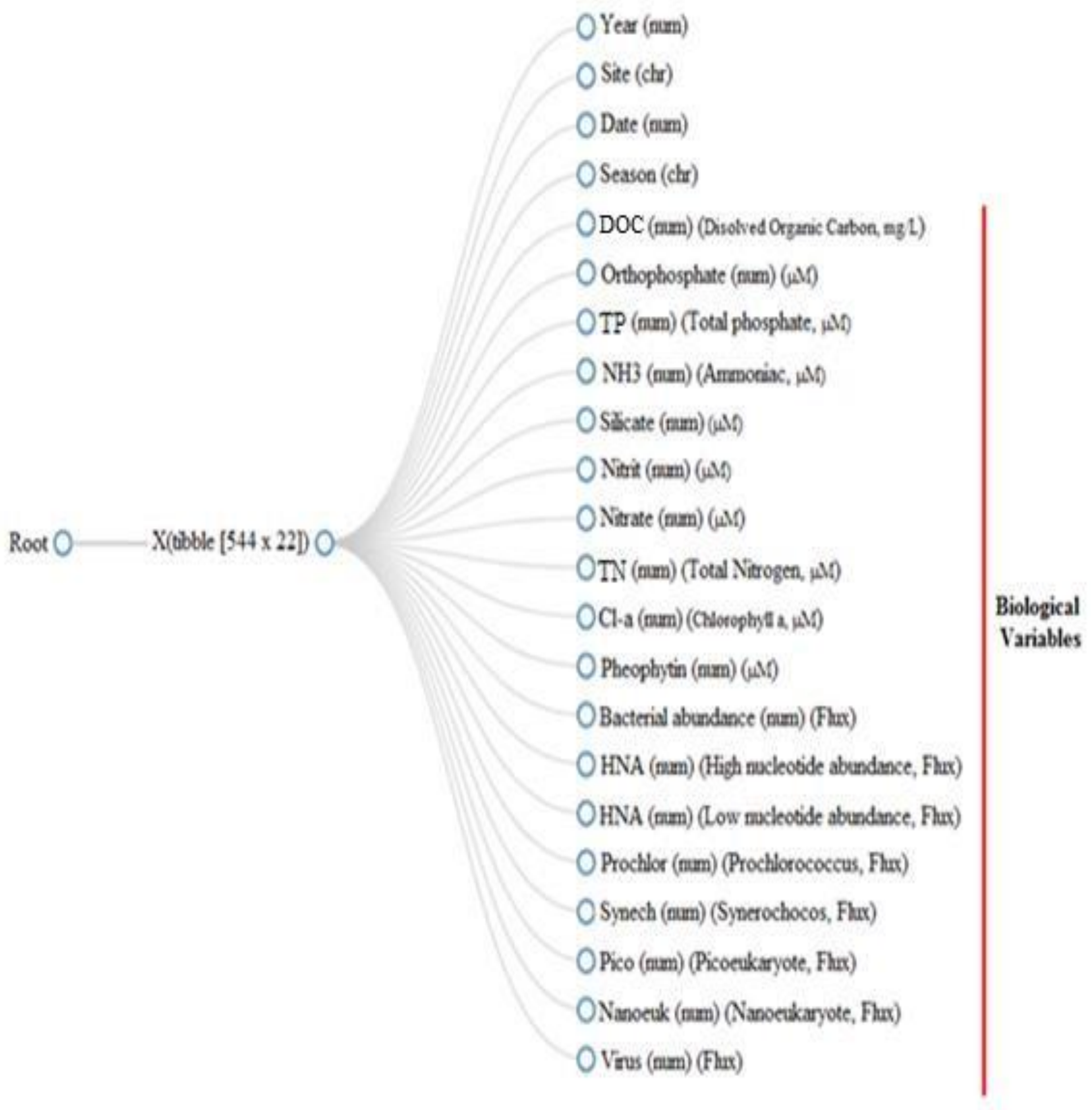

Root
X(tibble [544 x 22])
Year (num)
Site (chr)
Date (num)
Season (chr)
DOC (num) (Disolved Organic Carbon, mg/L)
Orthophosphate (num) (µM)
TP (num) (Total phosphate, µM)
NH3 (num) (Ammoniac, µM)
Silicate (num) (µM)
Nitrit (num) (µM)
Nitrate (num) (µM)
TN (num) (Total Nitrogen, µM)
Cl-a (num) (Chlorophyll a, µM)
Pheophytin (num) (µM)
Bacterial abundance (num) (Flux)
HNA (num) (High nucleotide abundance, Flux)
HNA (num) (Low nucleotide abundance, Flux)
Prochlor (num) (Prochlorococcus, Flux)
Synech (num) (Synerochocos, Flux)
Pico (num) (Picoeukaryote, Flux)
Nanoeuk (num) (Nanoeukaryote, Flux)
Virus (num) (Flux)
Biological Variables

Fig. 2 The variables studied in this work.

2.3. Data Exploratory Analysis

The number of samples per year and site is presented in Table 1, and Table 2 presents the number of observations per year and site.

Table 1. The number of samples per year and site.

	MV	PAB	AR	PL	TIM	SG	Total
2011	21	------	-----	-----	-----	------	21
2012	34	51	-----	10	27	23	145
2013	------	------	60	96	96	-----	250
2014	19	33	------	3	9	17	81
2016	------	12	13	11	-----	11	47
Total	74	96	73	120	130	51	

Table 2. The number of observations per years and site.

Year	Site	Silicate	TP	TN	Ortop.	Feof.	Nitrate	Nitrit	DOC	NH3	Cl_a	Bacterial	HNA	LNA	Prochlor	Synech	Picoeuk	Nanoeuk	Virus
2011	MV	21	17	16	19	21	18	20	21	15	21	21	21	21	12	21	21	21	21
2012	TIM	15	13	13	18	25	9	15	27	23	25	27	27	27	27	27	27	27	27
	PAB	30	15	18	41	42	16	26	51	42	42	36	36	36	30	36	36	36	36
	PL	3	4	3	8	8	3	5	4	8	8	9	9	9	9	9	9	9	9
	MV	30	13	13	12	18	12	22	34	31	18	30	30	30	30	30	30	30	30
	SG	14	6	6	6	16	6	7	23	14	16	18	18	18	18	18	18	18	18
2013	AR	24	24	24	24	0	24	23	60	18	17	54	54	54	54	54	54	54	53
	PL	36	32	32	36	27	36	36	96	23	27	85	85	85	83	83	83	83	83
	TIM	88	88	88	88	23	88	88	94	66	23	57	46	46	57	57	57	57	56
2014	MV	43	50	23	23	16	23	23	19	22	16	19	19	19	19	19	19	19	19
	PAB	33	23	23	23	5	23	23	33	12	5	33	33	33	30	33	33	33	33
	PL	23	33	23	12	3	23	23	3	22	3	3	3	3	3	3	3	3	3
	SG	33	34	23	22	14	54	54	17	22	14	17	17	17	17	17	17	17	17
	TIM	20	56	65	33	9	44	55	9	22	9	9	9	9	9	9	9	9	9
2016	PAB	10	10	10	10	23	10	10	12	10	23	11	11	11	11	11	11	11	11
	PL	11	11	11	11	23	11	11	11	11	23	10	10	10	10	10	10	10	10
	SG	10	10	10	10	21	10	10	11	10	23	10	10	10	10	10	10	10	10

Package ‘dlookr’, version 0.4.3, was used for Data Diagnosis, Exploration, and Transformation (Ryu 2019). Data diagnostics provides information and visualization of missing values and outliers and are necessary to understand the distribution and quality of data. Data exploration provides information and visualization of the descriptive statistics of univariate variables, normality tests and outliers, correlation of variables, and relationship between target variable and predictor. Data transformation supports binning for categorizing continuous variables, imputates missing values and outliers, resolving skewness. And it creates automated reports that support these three tasks (Ryu 2019). The dlookr package was used because it easily performs data diagnosis, automatically generates data diagnosis and exploratory data analysis reports, and treats skewed data.

2.3.1. Removing Data Outliers

Cleaning up data outliers is good method to see clear distribution of data. The outlier is a current problem faced by many data mining researches. Outliers are the patterns which are not in the range of normal behavior. Outliers in the dataset produce more false positive alarms. We have used Interquartile Range technique to identify the outliers (Vinutha *et al.*, 2018). In this, the continuous range of input is divided into quartiles and these quartiles are analyzed to target the range of outliers. Then the obtained outliers are removed (Fig. 3). In this figure, Median (Q2/50th Percentile) is the middle value of the dataset; First quartile (Q1/25th Percentile) is the middle number between the smallest number and the median of the dataset. Third quartile (Q3/75th Percentile) is the middle value between themedian and the highest value of the dataset. Interquartile range (IQR): 25th to the 75th percentile. Whiskers (shown in blue). Outliers (shown as green circles).

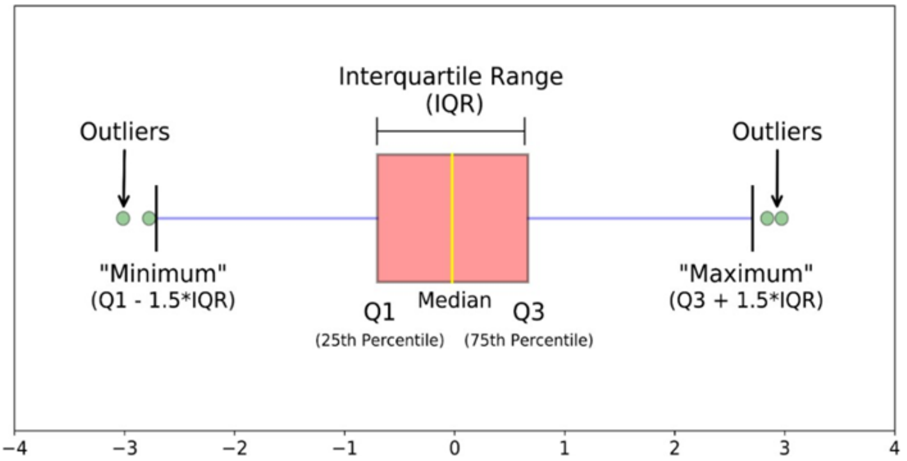


Fig. 3. Interquartile method to remove outliers.

2.3.2. Missing value Imputation by Principal Component Analysis (PCA)

Missing data are a common problem in most scientific research (Schmitt *et al.*, 2015). In statistics, imputation is the process of replacing missing data with substituted values. When substituting for a data point, it is known as "unit imputation"; when substituting for a component of a data point, it is known as "item imputation". There are three main problems that missing data causes: missing data can introduce a substantial amount of bias, make the handling and analysis of the data more arduous, and create reductions in efficiency. Because missing data can create problems for analyzing data, imputation is seen as a way to avoid pitfalls that have missing values.

We used Principal Component Analysis (PCA) to fill missing values. PCA was choosed because it simultaneously reduces the dimensionality of the dataset, increases interpretability and minimizes information loss. It does so by creating new uncorrelated variables that successively maximize variance. Finding such new variables, the principal components, reduces to solving an eigenvalue/eigenvector problem, and the new variables are defined by the dataset at hand, not a priori, hence making PCA an adaptive data analysis technique. It is adaptive in another sense too, since variants of the technique have been developed that are tailored to various different data types and structures (Davò et al., 2016).

Analog Ensemble (AnEn) post-processing was applied on the PCA output to obtain the final forecasts (Davò *et al.*, 2016). The data points were scored by how well they fit into a principal component (PC) based upon a measure of variance within the dataset. In this way, PCA result can be seen as a kind of clustering analysis (Bailey 2018). Finally, PCA is an imputation method of interest which deserves further consideration in practice and demonstrated higher capacity to impute threetypes of missing values: missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR) with lower Mean Squared Error (MSE) than others (Madley-Dowd *et al.*, 2019; Hegde *et al.*, 2019; Schmitt *et al.*, 2015) (Fig. 4).

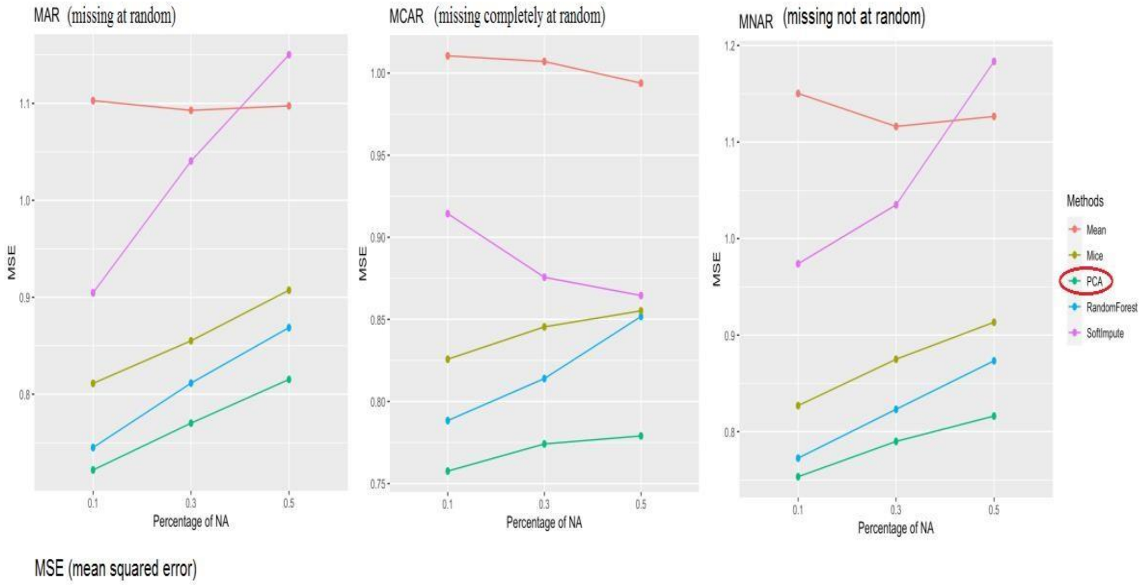


Fig. 4. Mean Squared Error (MSE) of PCA compared to other methods (Mean, Mice, Random Forest and Soft impute) in three types of missing values.

Absolute frequency histograms of each variable (Abs. Freq.), Relative frequency histograms of each variable (Rel. Freq.), Boxplots of each variable and Boxplots of each variable in sites/per years to Data Exploratory analysis were performed. Graphical abstract of histograms and boxplots of each variable is presented in Fig. 5.

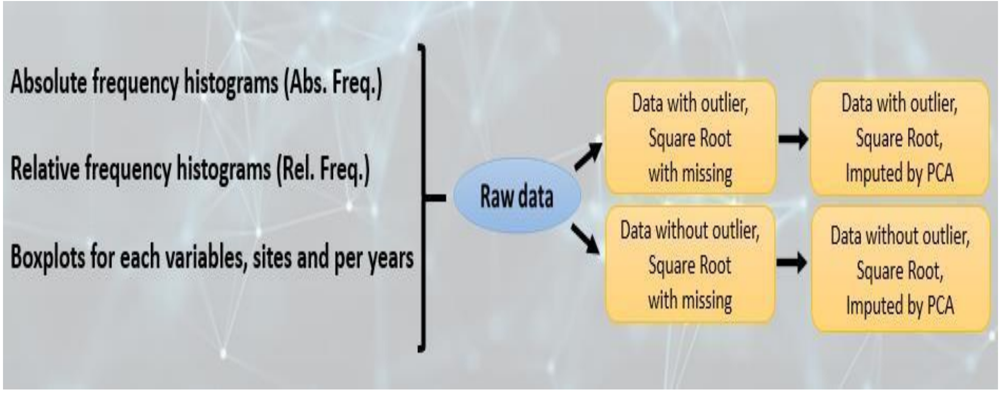

Absolute frequency histograms (Abs. Freq.)
Relative frequency histograms (Rel. Freq.)
Boxplots for each variables, sites and per years
Raw data
Data with outlier, Square Root with missing
Data with outlier, Square Root, Imputed by PCA
Data without outlier, Square Root with missing
Data without outlier, Square Root, Imputed by PCA

Fig. 5. Graphical Abstract of Histograms and Boxplots pattern of each variables.

2.4. Modelling

The Regression Tree Algorithms can be used to find one model that results in good predictions for our dataset, or, saying in other way, to perform causal inference. We can analize the statistics and confusion matrices of the current predictors to see if our model is a good fit to the data. Among the regression algorithms, Boosted Regression Trees (BRT) and Random Forest (RF) have been shown to be the strongest in the literature on Ecology. Ecologists use BRT and RF statistical models for both explanation and prediction, and need techniques that are flexible enough to express typical features of their data, such as nonlinearities and interactions.

The RF algorithm operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees, while the BRT algorithm combines the strengths of two algorithms: decision tree algorithms and boosting methods. As RF models, BRT repeatedly fits many decision trees to improve the accuracy of the model, but while RF models use the bagging method (each occurrence has an equal probability of being selected in subsequent samples), BRT uses the boosting method, in which the input data are weighted in subsequent trees and the weights are applied in such a way that data that was poorly modelled by previous trees has a higher probability of being selected in the new tree.

BRT and RF trees incorporate important advantages of tree-based methods, handling different types of predictor variables and accommodating missing data. Fitting multiple trees in BRT and RF overcomes the biggest drawback of single tree models: their relatively poor predictive performance.

2.4.1. Boosted Regression Trees (BRT)

The BRT method combines regression trees and a boosting technique to improve the predictive performance of multiple single models, where Boosting is a forward and stage-wise procedure in which a subset of the data is randomly selected to iteratively fit new tree models to minimize the loss function (Elith et al., 2008). This process introduces a stochastic gradient boosting procedure that can improve model performance and reduce the risk of overfitting (Friedman, 2002).

The BRT algorithm is an iterative process in which tree-based models were fitted iteratively using recursive binary splits to identify poorly modeled observations in existing trees until a minimum model deviance was reached. The final fitted model is a linear function of the sum of all trees multiplied by the learning rate (LR) based on all data (Elith et al., 2008).

In BRT modeling, four parameters are user defined: the learning rate (LR), tree complexity (TC), number of trees (NT) and bag fraction (BF). LR represents the contribution of each tree to the final fitted model, and TC controls the size of trees and whether inter-actions between variables should be considered. When TC = 1, each tree contains a single decision stump and models the effectof one variable; and when TC > 1, each tree fits a model that predicts the interactions of variables. NT is determined from the combination of LR and TC.

In practice, the BRT model needs to be regularized by setting up the parameters prior to making a prediction (Elith *et al.*, 2008). To perform this regularization, a few combinations of parameter values (LR, TC and BF) were tested. The optimal parameter combination was that which providedthe minimum predictive deviance. The final optimal values of LR, TC and BF were set to 0.0025, 9, and 0.75, respectively. This combination can generate an optimal NT of at least 1000 trees using a 10-foldcross-validation method. Elith *et al.* (2008) recommended the use of no fewer than 1000 trees when fitting such models. The relative importance of variables can be measured based on the number of times a variable is selected for modelingand weighted by the square improvement to each split and averaged across all trees (Friedman, 2001).

2.4.2. Random Forest (RF)

The RF algorithm generates multiple trees without pruning. In the training procedure, each tree is built based on a random subset of the original data (with replacement). In addition, a randomly selected subset of predictors is chosen for each built tree (Breiman, 2001).

Theuse of bootstrap sampling in RF modeling allows the remaining un-used subset (i.e., the out-of-bag data (OOB)) to be used for the estimation of general errors. RF predictions are the averaged output of all aggregations.

RF modeling requires three user-defined parameters: the number of variables used to grow each tree (mtry), the number of trees inthe forest (ntree) and the minimum number of terminal nodes (node size). The mtry parameter determines the strength of each individual tree and correlations between trees, and increasing mtry also increases the strength of each individual tree and correlations between trees (Peters *et al.*, 2008). However, the predictive performance of the RF modelis improved by increasing the tree strength and decreasing the correlations among trees (Ließ *et al.*, 2012).

To fit an RF model, default values of mtry (one third of the total numberof predictors) and node size (5) were used. The default value of ntree (500) has been provento be insufficient to yield stable results (Grimm *et al.*, 2008). Thus, we applied the RF model with ntree = 1000. The relative importance of variables can be estimated from the mean decrease in predictive accuracy when the variable is permuted (Prasad *et al.*, 2006).

2.4.3. Statistical Analyses and Model Validation

The Pearson correlation was used to relate the dependent variable of microbial abundance to independent quantitative variables. Statistical analysis and modeling in this study were performed using the R software (R Development Core Team, 2009). The BRT and RF models were developed using a BRT script provided by Elith *et al.* (2008) and the R Random Forest package (Liaw and Wiener, 2002).

The performance of the BRT and RF models was evaluated using a 10-fold cross-validation procedure that involved comparisons between the predicted and observed microbial abundance values. Three validation measurements were calculated: mean absolute prediction error (MAE), root mean square error (RMSE) and coefficient of determination (R^2) (Lin, 1989). MAE measures the average prediction bias, and RMSE represents the overall quality of the prediction. Predictions become increasingly optimal as MAE and RMSE approach zero.

3. Results

In this Section we present the results of the Data Exploratory Analysis, the results of the application of RF and BRT to perform causal inference over the dataset, the difference of biophysical and biochemical parameters and microbial abundance in the internal and external arcs of Abrolhos, and, finally, a comprehensive discussion over the achieved results.

3.1. Data Exploratory Analysis

The percent of missing for each variable was verified by ‘dlookr’ package (Fig. 6). Total missing values in data set verified as 37% (Fig. 6a) and then this missing value was imputed by PCA (Fig. 6b). The frequency of data per study sites and season are presentedin Fig. 7a and b respectively. The season parameter has high variety between Summer (523 samples) and Winter (21 samples) and has been neglected here (Fig. 7b). We verified the amount of missing data totally, per year, and site (Table 3), and we also analyzed the data per year, site and year/site separately. The missing number of observations in years/sites that were filled by PCA is presented in Table 3. A part of data set in Excel that shows the format of missing number of observations in years/sites, which were filledby PCA is showed in Fig. 8. Finally, the Pearson correlation plot of imputed data by PCA was showed in Fig. 9. Correlation amongvariables with R > 0.75 were showed in Table 4.

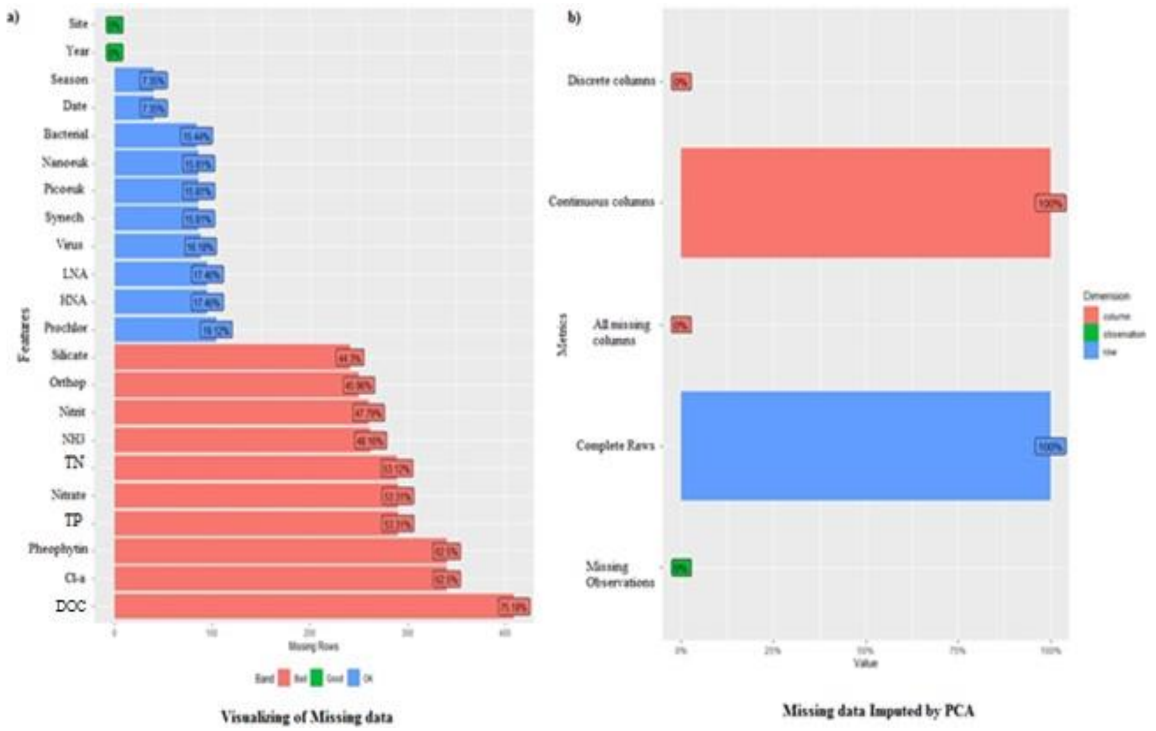


Fig. 6 a) Visualizing of missing data. Total missing in data set 37%. b) Missing Imputed by PCA.

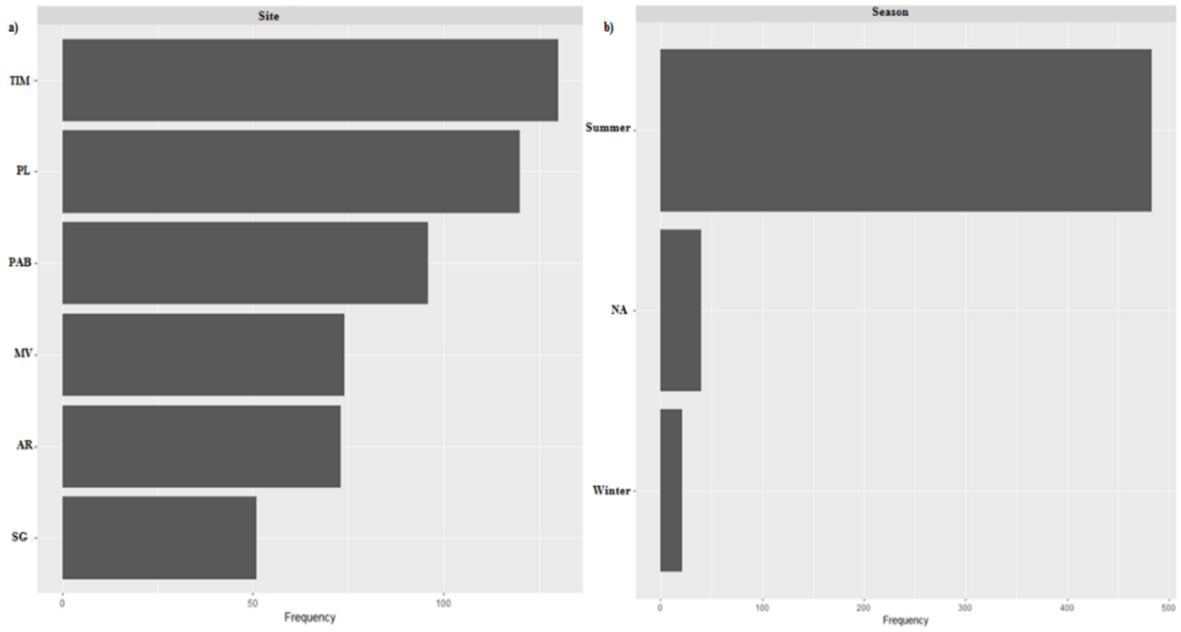
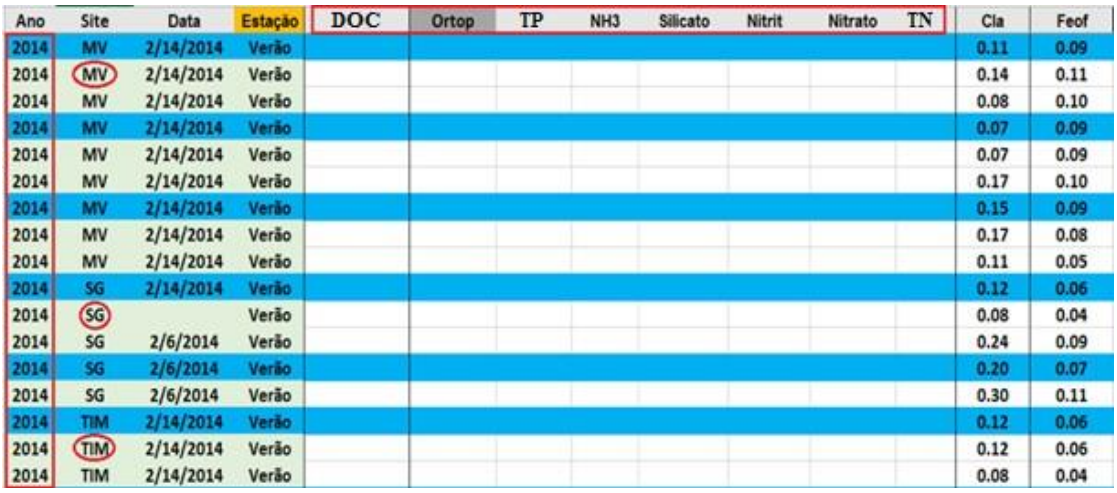


Fig. 7. Data frequency for study sites and seasons.

Table 3. The missing number of observations in years/sites.

Year	Site	Silicate	TP	TN	Ortoph.	Feof.	Nitrate	Nitrit	DOC	NH3	Cl_a	Bacterial	HNA	LNA	Prochlor	Synech	Picoeuk	Nanoeuk	Virus
2011	MV	8	4	5	2	8	3	1	6	6	8	8	8	8	9	7	6	6	6
2012	TIM	12	14	14	9	2	18	12	9	4	2	8	8	8	5	6	6	6	6
	PAB	21	36	33	10	9	35	25	31	9	9	15	15	15	21	15	15	15	15
	PL	7	6	7	2	2	7	5	4	2	2	1	1	1	1	1	1	1	1
	MV	4	21	21	22	16	22	12	22	3	16	4	4	4	4	4	4	4	4
	SG	9	17	17	17	7	17	16	11	9	7	5	5	5	5	5	5	5	5
2013	AR	36	36	36	36	60	36	37	60	42	60	6	6	6	6	6	6	6	7
	PL	60	64	64	60	69	60	60	69	73	69	11	11	11	13	13	13	13	13
	TIM	6	6	6	6	94	6	6	94	28	94	37	48	48	37	37	37	37	38
2014	MV	19	19	19	19	3	19	19	19	19	3	6	6	6	8	8	9	7	7
	PAB	33	33	33	33	28	33	33	33	33	28	6	6	6	7	7	7	8	9
	PL	3	3	3	3	12	3	3	3	3	11	6	6	6	6	7	7	9	8
2016	SG	17	17	17	17	3	17	17	17	17	3	7	7	7	8	7	7	7	7
	TIM	9	9	9	9	8	9	9	9	9	9	7	7	7	7	7	7	7	7
	PAB	2	2	2	2	12	2	2	1	2	12	1	1	1	1	1	1	1	1
	PL	8	7	7	7	11	7	7	7	7	11	1	1	1	1	1	1	1	1
	SG	1	1	1	1	11	1	1	8	1	11	1	1	1	1	1	1	1	1



Ano	Site	Data	Estação	DOC	Ortop	TP	NH3	Silicato	Nitrit	Nitrato	TN	Cla	Feof
2014	MV	2/14/2014	Verão									0.11	0.09
2014	MV	2/14/2014	Verão									0.14	0.11
2014	MV	2/14/2014	Verão									0.08	0.10
2014	MV	2/14/2014	Verão									0.07	0.09
2014	MV	2/14/2014	Verão									0.07	0.09
2014	MV	2/14/2014	Verão									0.17	0.10
2014	MV	2/14/2014	Verão									0.15	0.09
2014	MV	2/14/2014	Verão									0.17	0.08
2014	MV	2/14/2014	Verão									0.11	0.05
2014	SG	2/14/2014	Verão									0.12	0.06
2014	SG		Verão									0.08	0.04
2014	SG	2/6/2014	Verão									0.24	0.09
2014	SG	2/6/2014	Verão									0.20	0.07
2014	SG	2/6/2014	Verão									0.30	0.11
2014	TIM	2/14/2014	Verão									0.12	0.06
2014	TIM	2/14/2014	Verão									0.12	0.06
2014	TIM	2/14/2014	Verão									0.08	0.04

Fig. 8. A part of data set that shows the missing number of observations in years/sites.

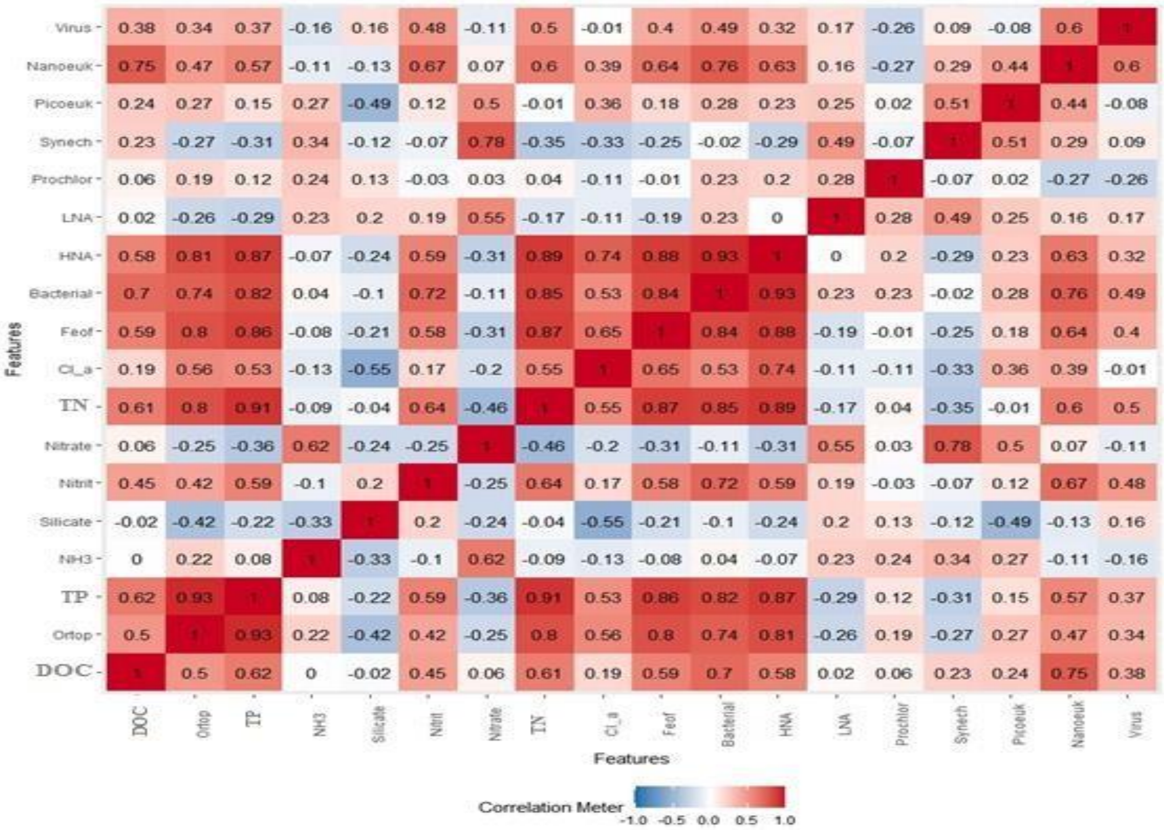


Fig. 9. Correlation among variables imputed by PCA.

Table 4. Correlation among variables with R > 0.75 in Imputed by PCA.

Variable 1	Variable 2	Correlation
DOC	Nanoeuk	0.75
Ortoph.	TP	0.93
	TN	0.81
	Feof	0.76
	HNA	0.84
TP	TN	0.92
	Feof	0.86
	Bacterial	0.82
	HNA	0.93
Nitrate	Synech	0.78
TN	Bacterial	0.87
	HNA	0.94
	Feof	0.9
Feof	Bacterial	0.83
	HNA	0.89
Bacterial	HNA	0.95
	Nanoeuk	0.76

3.2. Histograms and Box-Plots of Numerical Variables

For each variable, we produced graphs to help the data visualization. Collecting and summarizing data (numerically and graphically) help us understanding what is going on in the sample. The goal is to understand what is happening in the population from that sample. If there are many data points and we would like to see the distribution of the data,we can represent the data by a frequency histogram or a relative frequency histogram.

Looking at the histograms diagram, we can quickly understand the 4 factors: Pick (most frequency), Gap (no information), Concentration (Two or three bars of similar size) and Outlier. The only difference between a frequency histogram and a relative frequency histogram is that the vertical axis uses relative or proportional frequency instead of simple frequency (counts of data in percentage form).

In the present study, the data range and form vary a lot. For example, the range for bacteria is between 0 and 1.651.018, and for the DOC is between 1 and 6.310. In a normal distribution like Bacterial, Virus, LNA and TN points on one side of the average are as likely to occur as on the other side of the average. In a right-skewed distribution (also called a positively skewed distribution), a large number of data values occur on the left side with a fewer number of data values on the right side. A right-skewed distribution occurs in Nanoeuk, Picoeuk, Synech, HNA, Prochlor, Silicate, Ortop., Nitrite, NH3, Nitrate, TP, Feof, Cl-a and DOC.

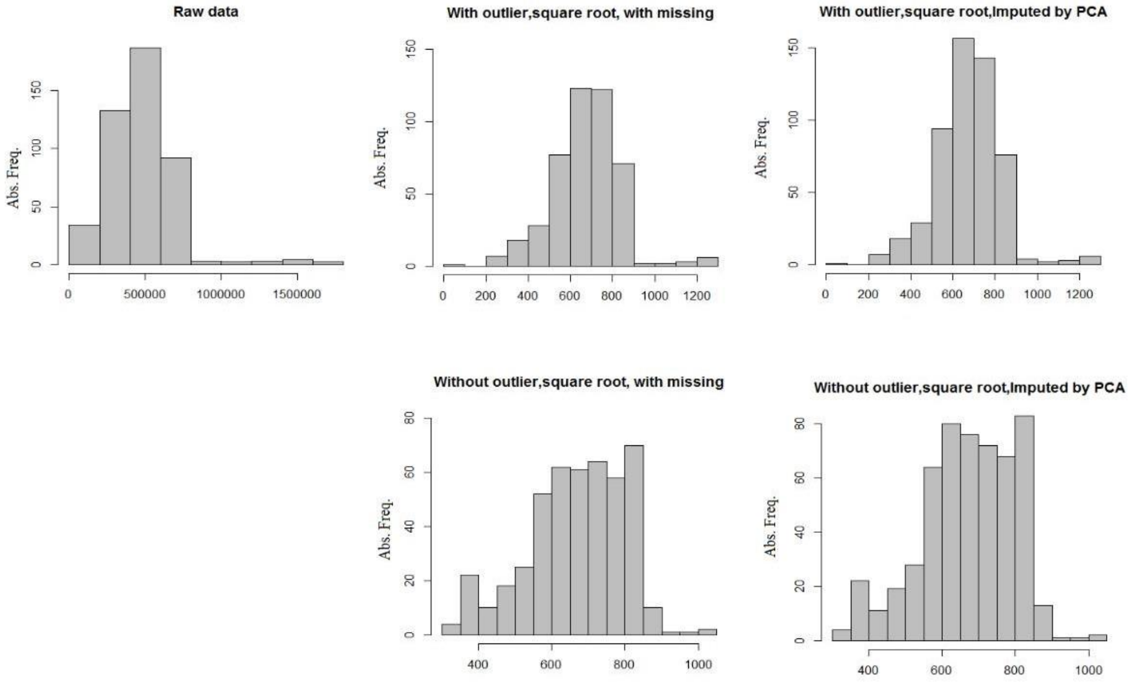
To better understand the behavior of the data we used the histograms and boxplot graphs. The Absolute Histogram presents the number of times the event occurred in the data set, while the Relative Frequency Histogram shows the fraction or proportion of times that a value occurs. The Box plot is a convenient way of graphically depicting groups of numerical data through their five-number summaries: the smallest observation, lower quartile (Q1), median (Q2), upper quartile (Q3), and the largest observation. A boxplot may also indicate which observations, if any, might be considered outliers.

Five data sets were designed to perform the Exploratory Data Analysis of each outcome variable: 1. Raw data (Original data), 2. Data with outlier, normalized by square root and with missing values, 3. Data with outlier, normalized by square root and without missing values (Imputed by PCA), 4. Data without outlier, normalized by square root and with missing values, 5. Data without outlier, normalized by square root and without missing values (Imputed by PCA).

Outliers can be important in biological data and they can indicate something scientifically interesting and increase the variability in data. On the other hand,we can see in data with outliers the clear normality shape both in Absolute/Relative frequency histograms of variables. We followed the pattern of Histograms and Boxplots that is presented in Fig. 5. Absolute/Relative frequency histograms and Boxplots of each variable were presented in Figs. 10-27. Data set with outliers, normalized by square root and imputed by PCA (with clear normality) was used for the rest of analysis.

Fig. 10 shows the analysis of bacterial abundance. We can see the highest frequency of bacterial abundance is between the ranges 600 and 800 of observations (Fig 10. a and b). Boxplot of raw data shows outliers at the upper range of the data (above the box), the mean value (477712) is above the median (465039), the median line does not evenly divide the box, and the upper tail of the boxplot is longer than the lower tail, then the distribution of which the data were sampled may be skewed to the right (Fig. 1c).

a



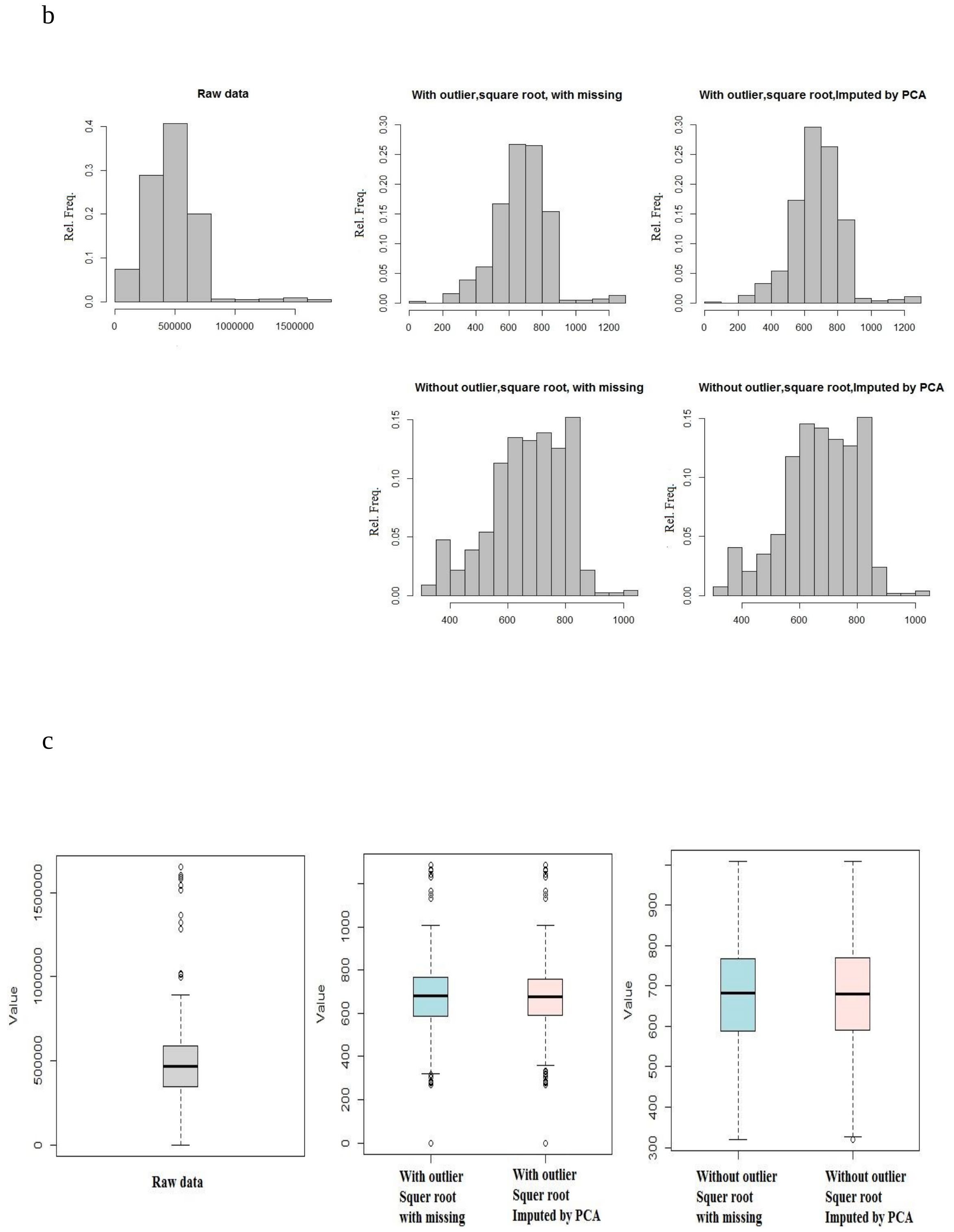
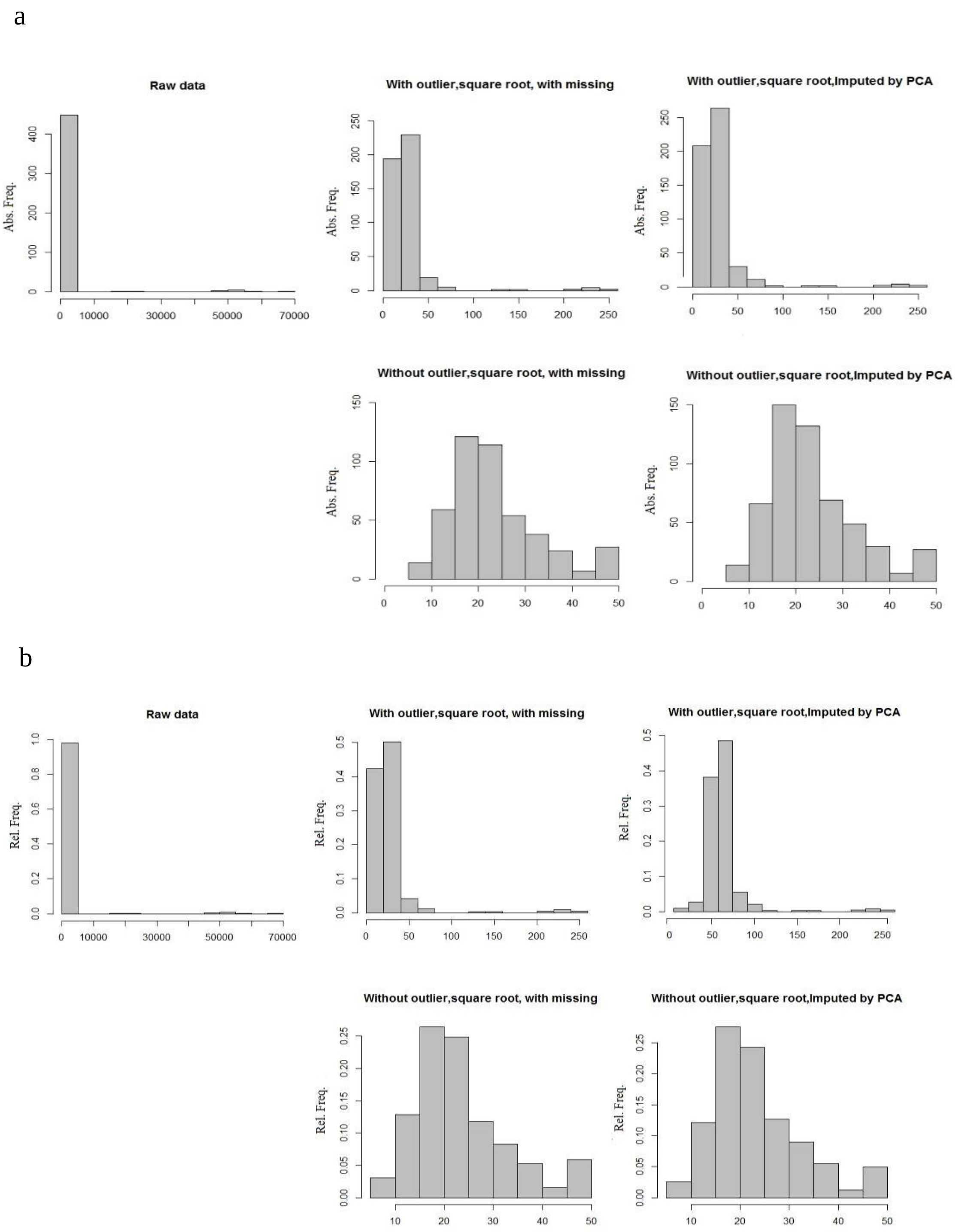


Fig. 10. a) Absolute count, b) Relative frequency histograms and c) boxplot of the bacterial abundance with 15% missing data that is shown for five data sets.

Fig. 11 shows the analysis of Nanoeuk abundance. We can see the highest frequency of Nanoeuk abundance is between the ranges 0 and 257.536 of observations (Fig. 11 a and b). Boxplot of raw data shows outliers at the upper range of the data (above the box), the mean value (27.423) is above the median (22.551), the median line does not evenly divide the box, and the upper tail of the boxplot is longer than the lower tail, then the distribution of which the data were sampled may be skewed to the right (Fig. 11 c).



c

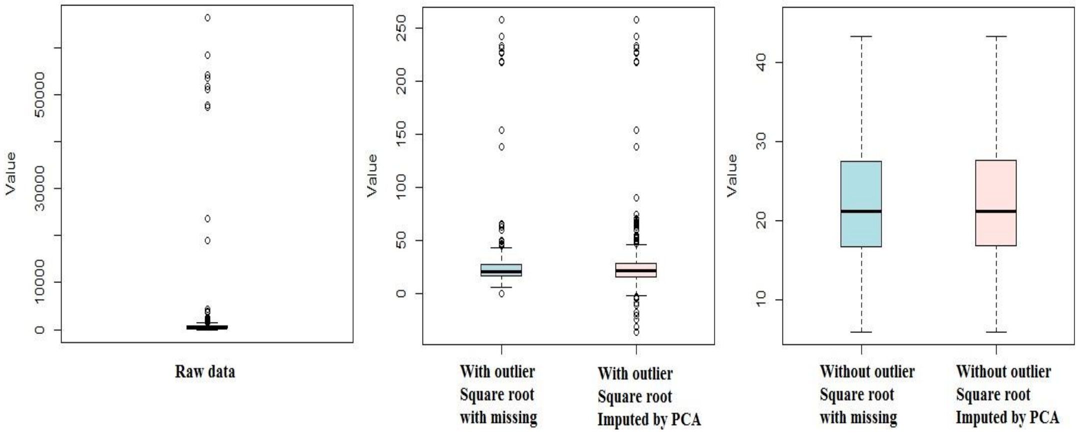
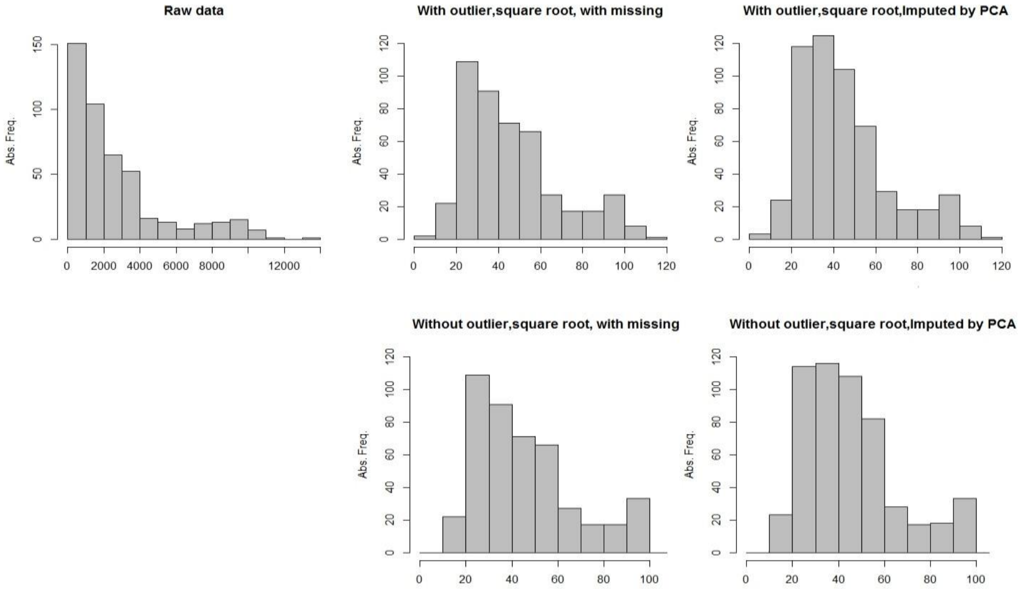


Fig. 11. a) Absolute count, b) Relative frequency histograms and c) boxplot of the Nanoeuk abundance with 15% missing data that is shown for five data sets.

Fig. 12 shows the analysis of Picoeuk abundance. We can see the highest frequency of Picoeuk abundance is between the ranges 0 and 116.05 of observations (Fig 12. a and b). Boxplot of raw data shows outliers at the upper range of the data (above the box), the mean value (46.44) is above the median (42.47), the median line does not evenly divide the box, and the upper tail of the boxplot is longer than the lower tail, then the distribution of which the data were sampled may be skewed to the right (Fig. 12 c).

a



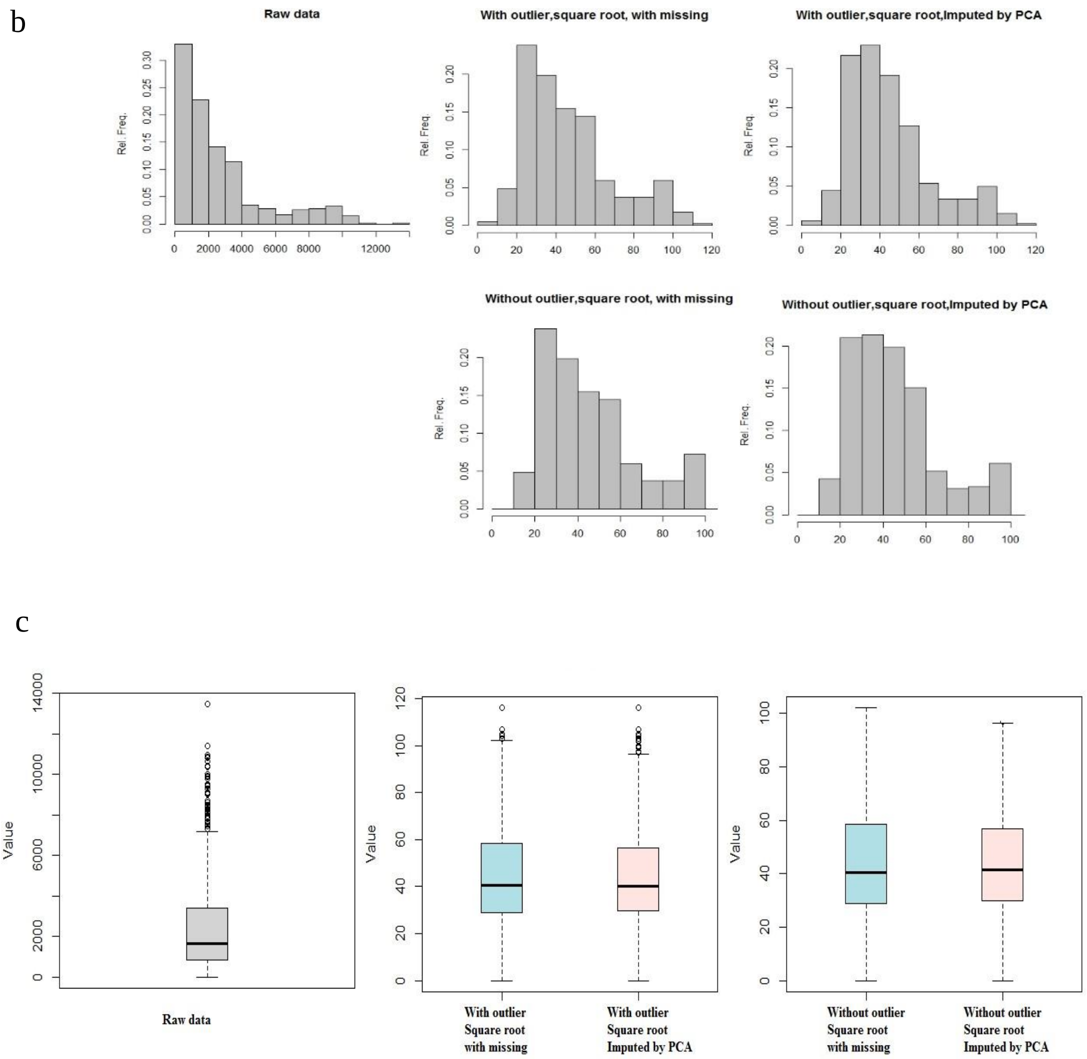


Fig. 12. a) Absolute count, b) Relative frequency histograms and c) boxplot of the Picoeuk abundance with 15% missing data that is shown for five data sets.

Fig. 13 shows the analysis of Synech abundance. We can see the highest frequency of Synech abundance is between the ranges 0 and 514.87 of observations (Fig 13. a and b). Boxplot of raw data shows outliers at the upper range of the data (above the box), the mean value (214.67) is above the median (209.61), the median line does not evenly divide the box, and the upper tail of the boxplot is longer than the lower tail, then the distribution of which the data were sampled may be skewed to the right (Fig. 13 c).

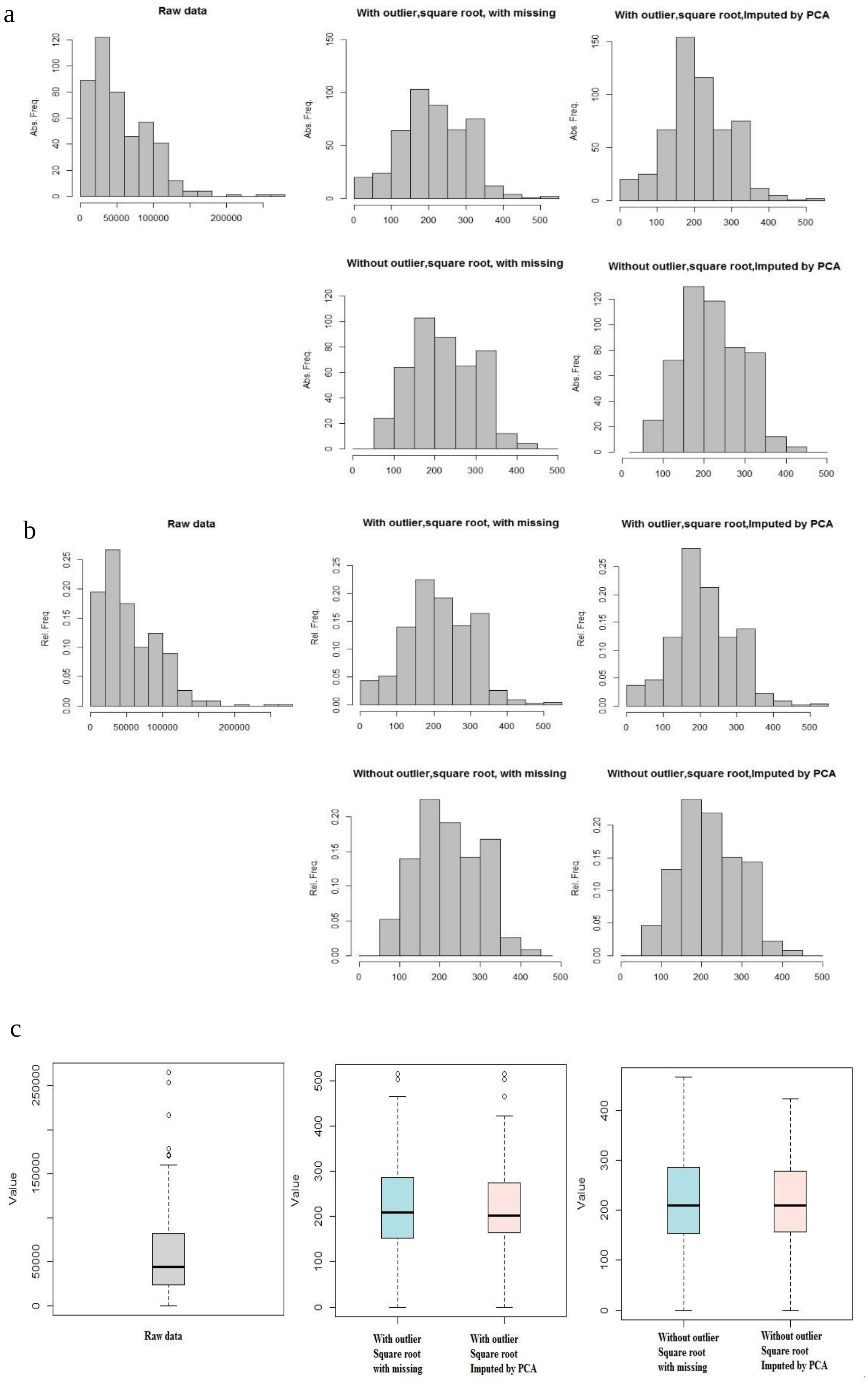
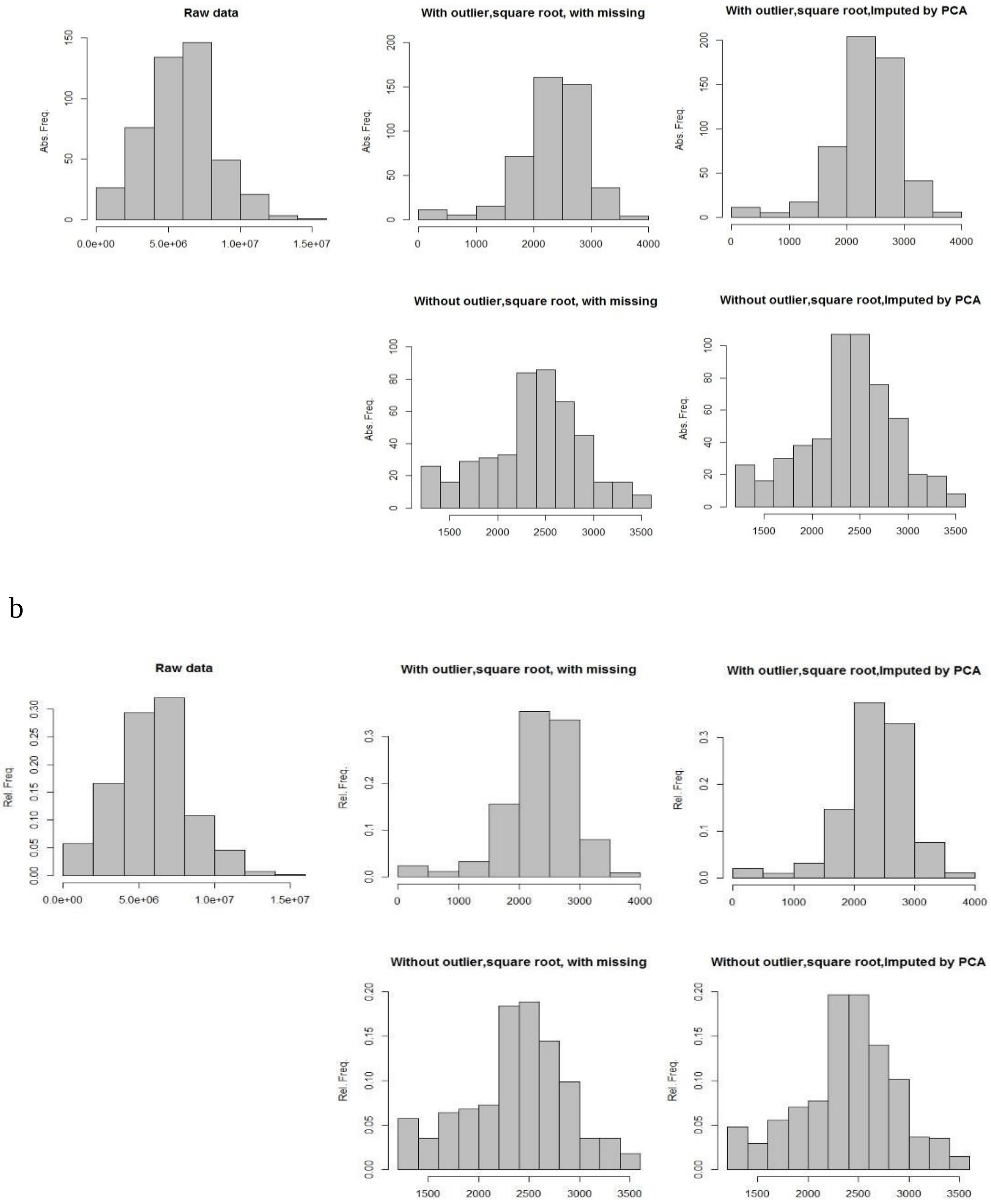


Fig. 13. a) Absolute count, b) Relative frequency histograms and c) boxplot of the Synech abundance with 15% missing data that is shown for five data sets.

Fig. 14 shows the analysis of Virus abundance. We can see the highest frequency of Virus abundance is between the ranges 0 and 4015 of observations (Fig 14a and b). Boxplot of raw data shows outliers at the upper range of the data (above the box), the mean value (2351) and the median (2420), they are almost equal, the median line does evenly divide the box (normal distributions), (Fig. 14 c).



a

b

c

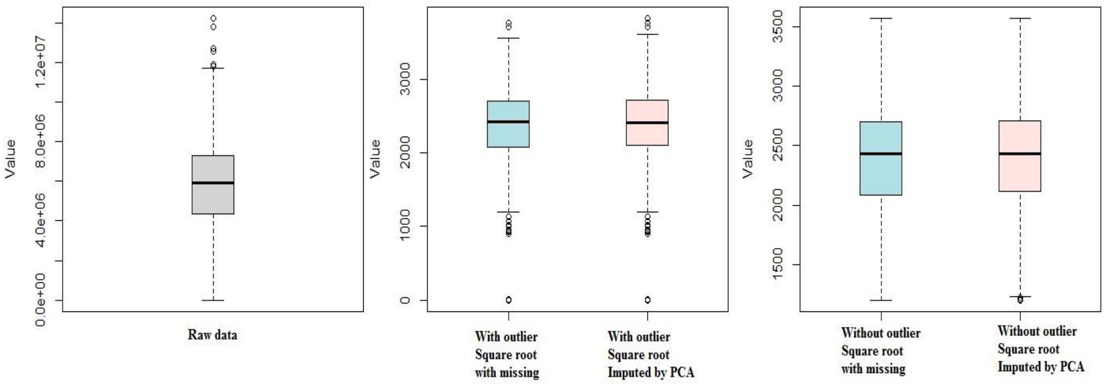
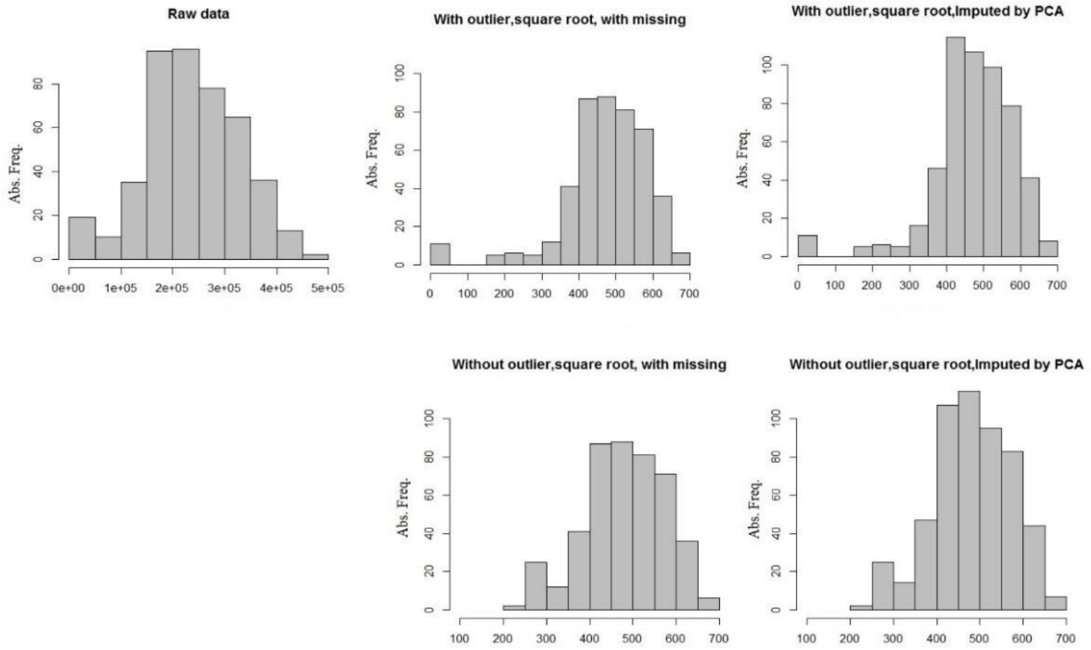


Fig. 14. a) Absolute count, b) Relative frequency histograms and c) boxplot of the Virus abundance with 15% missing data that is shown for five data sets.

Fig. 15 shows the analysis of LNA abundance. We can see the highest frequency of LNA abundance is between the ranges 0 and 732.5 of observations (Fig. 15a and b). Boxplot of raw data shows outliers at the upper range of the data, the mean value (473.6) and the median (481.1), they are almost equal, the median line does evenly divide the box (normal distributions) (Fig. 15 c).

a



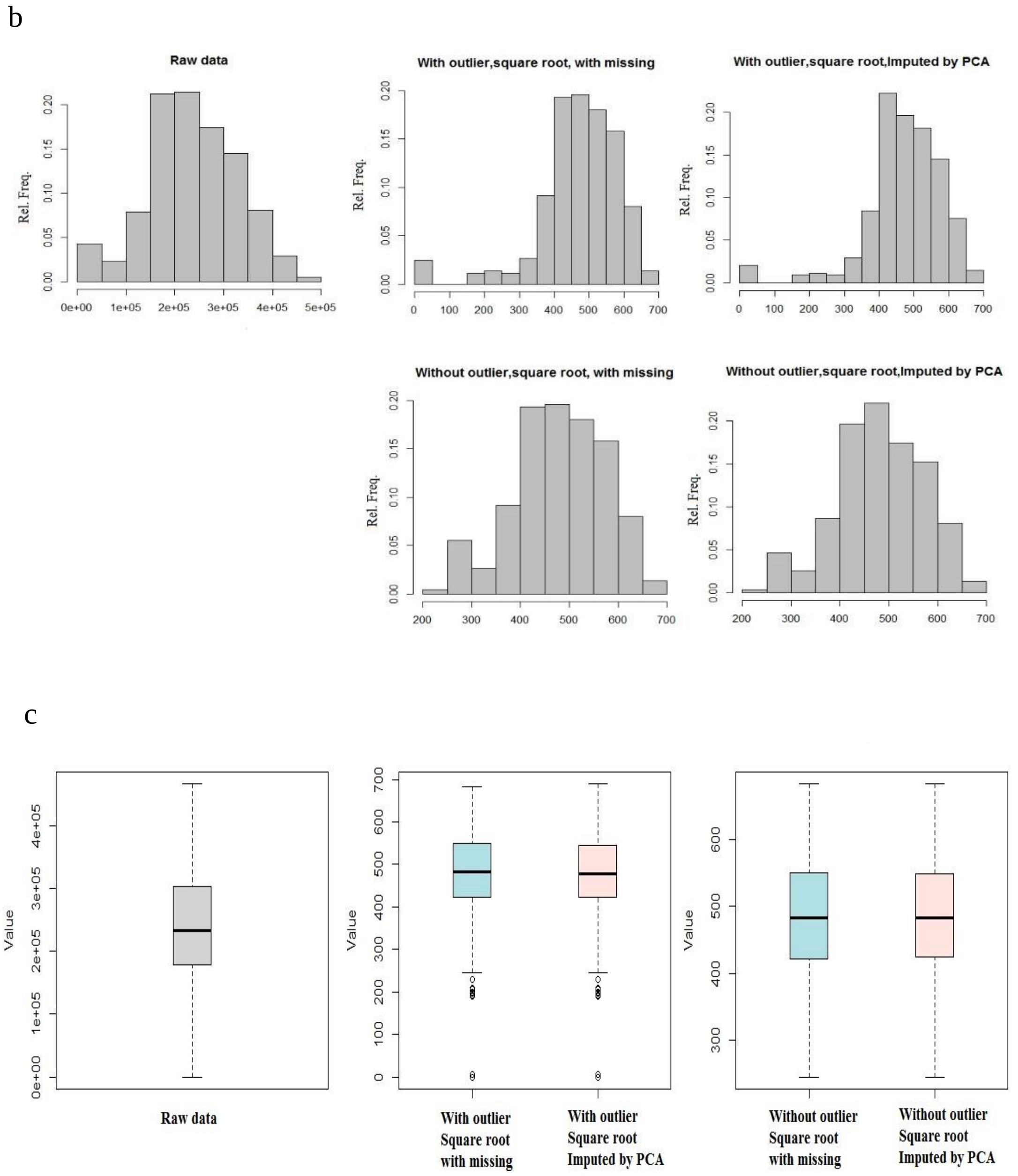


Fig. 15. a) Absolute count, b) Relative frequency histograms and c) boxplot of the LNA abundance with 15% missing data that is shown for five data sets.

Fig. 16 shows the analysis of HNA abundance. We can see the highest frequency of HNA abundance is between the ranges 0 and 732.5 of observations (Fig. 16a and b). Boxplot of raw data shows outliers at the upper range of the data, the mean value (473.6) and the median (481.1), they are almost equal, the median line does evenly divide the box (normal distributions), (Fig. 16 c).

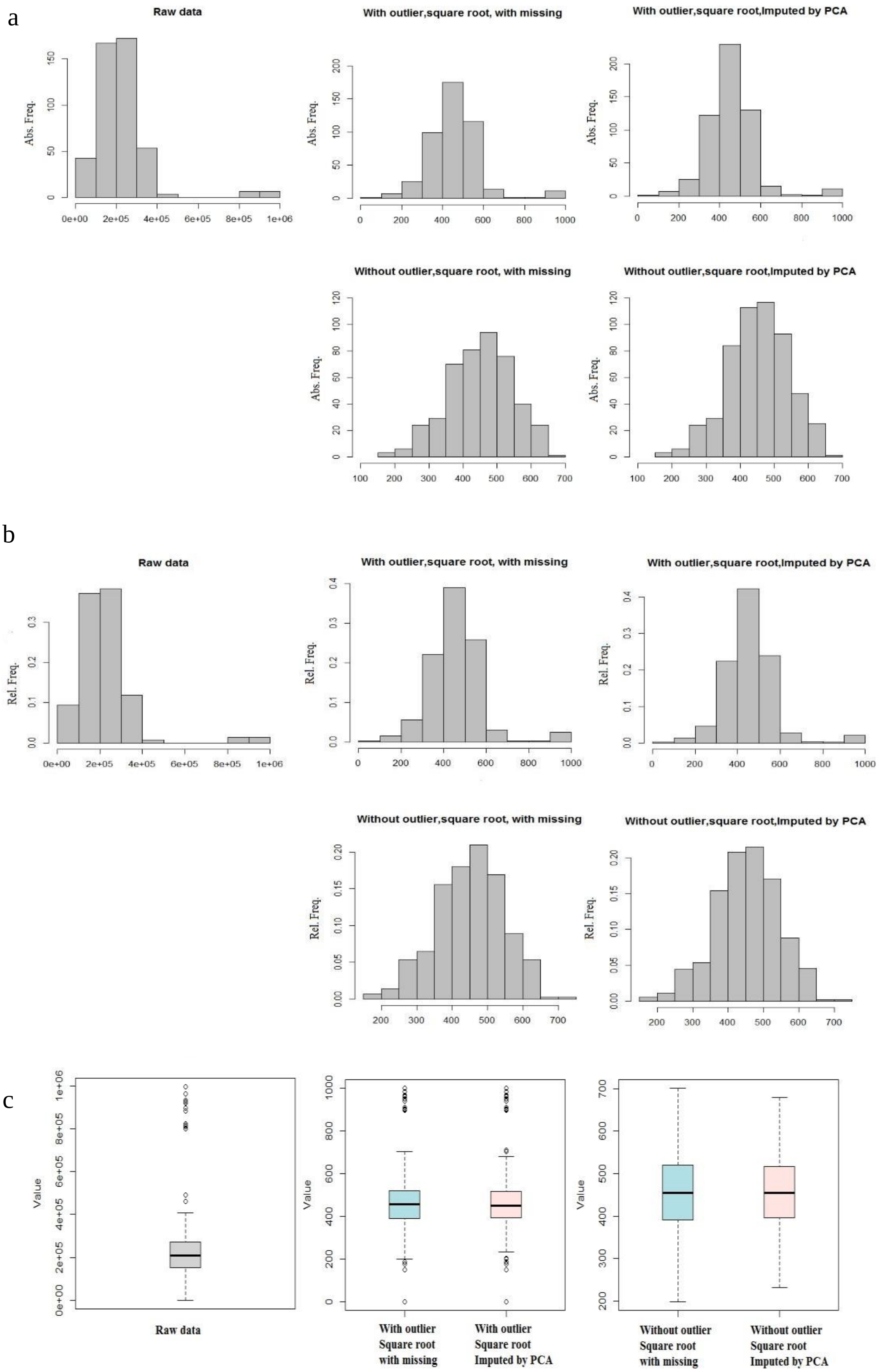
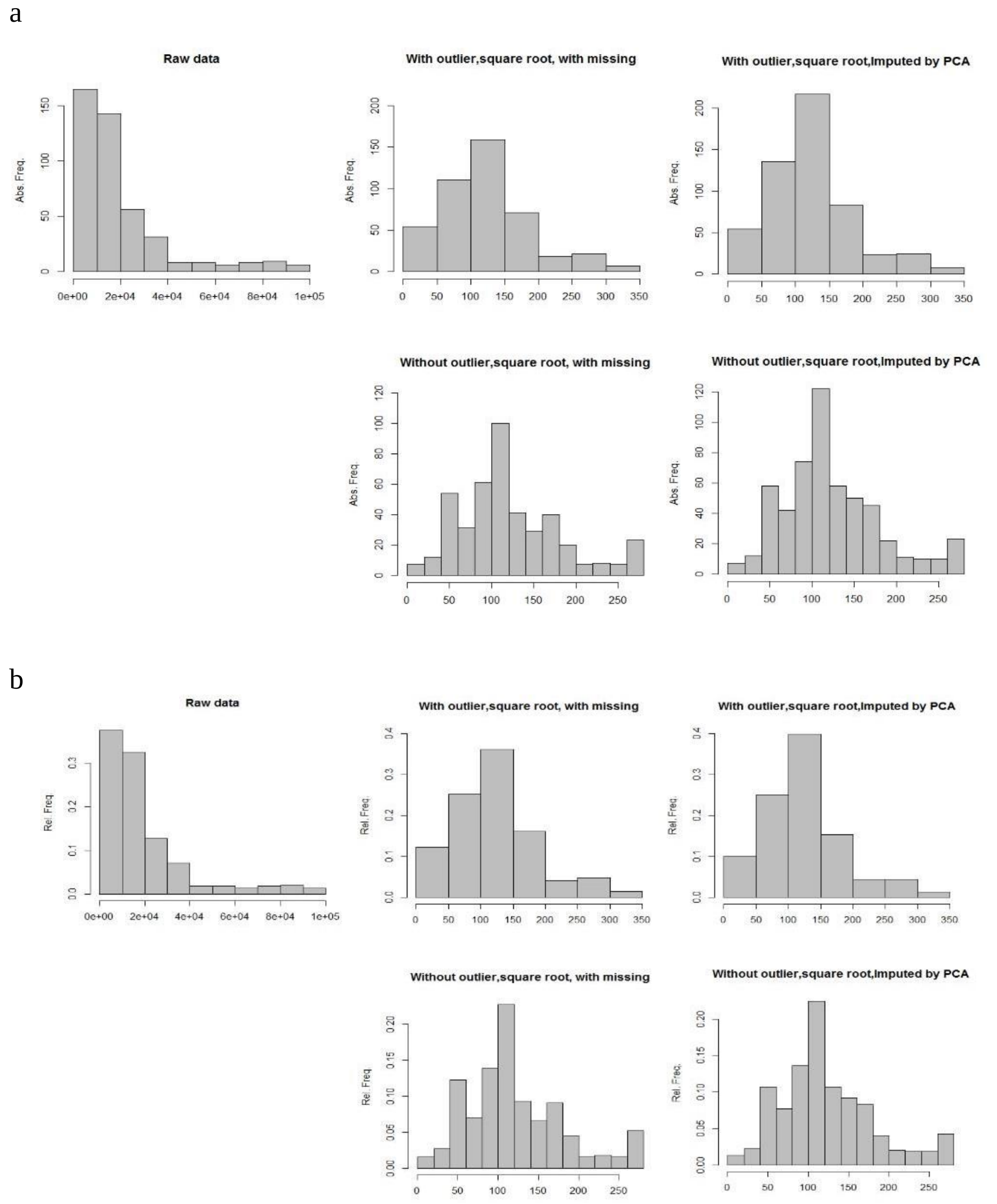


Fig. 16. a) Absolute count, b) Relative frequency histograms and c) boxplot of the HNA abundance with 15% missing data that is shown for five data sets.

Fig. 17 shows the analysis of Prochlor abundance. We can see the highest frequency of Prochlor abundance is between the ranges 0 and 316.02 of observations (Fig. 17a and b). Boxplot of raw data shows outliers at the upper range of the data, the mean value (123.07) is above the median (113.57), the median line does not evenly divide the box, and the upper tail of the boxplot is longer than the lower tail, then the distribution of which the data were sampled may be skewed to the right (Fig. 17 c).



c

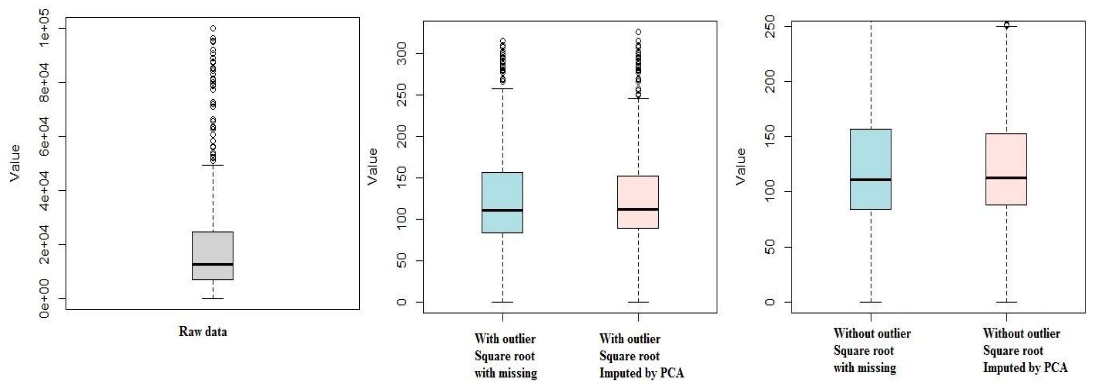
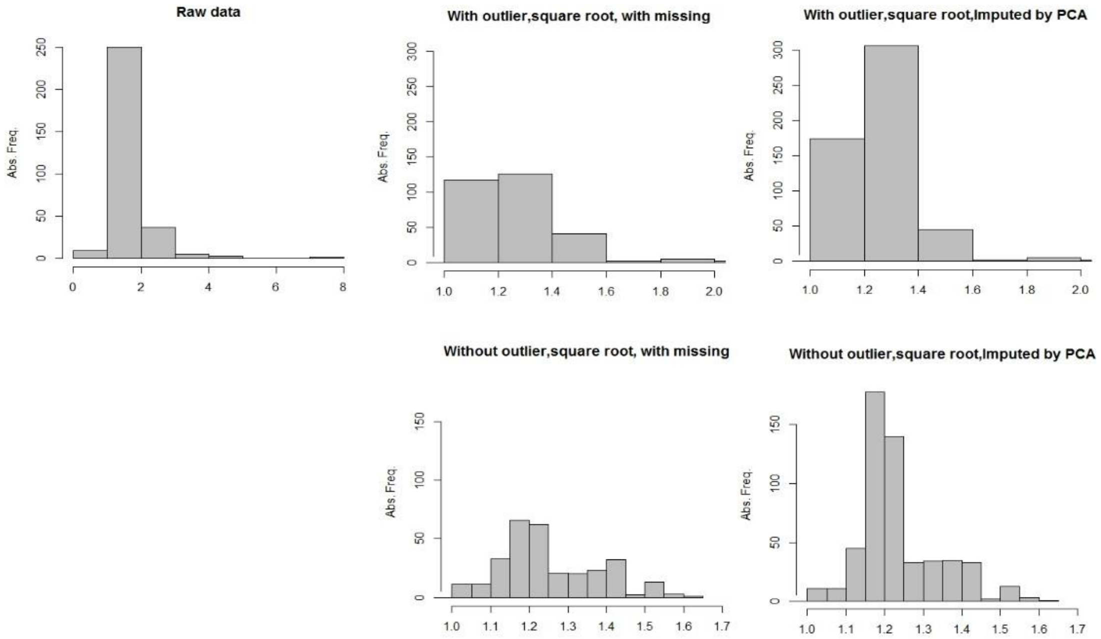


Fig. 17. a) Absolute count, b) Relative frequency histograms and c) boxplot of the Prochlor abundance with 15% missing data that is shown for five data sets.

Fig. 18 shows the analysis of Silicate abundance. We can see the highest frequency of Silicate abundance is between the ranges 0.7817 and 2.7744 of observations (Fig. 18a and b). Boxplot of raw data shows outliers at the upper range of the data, the mean value (1.2745) is above the median (1.2717), the median almost does not line evenly divide the box, and the upper tail of the boxplot is longer than the lower tail, then the distribution of which the data were sampled may be skewed to the right (Fig. 18 c).

a



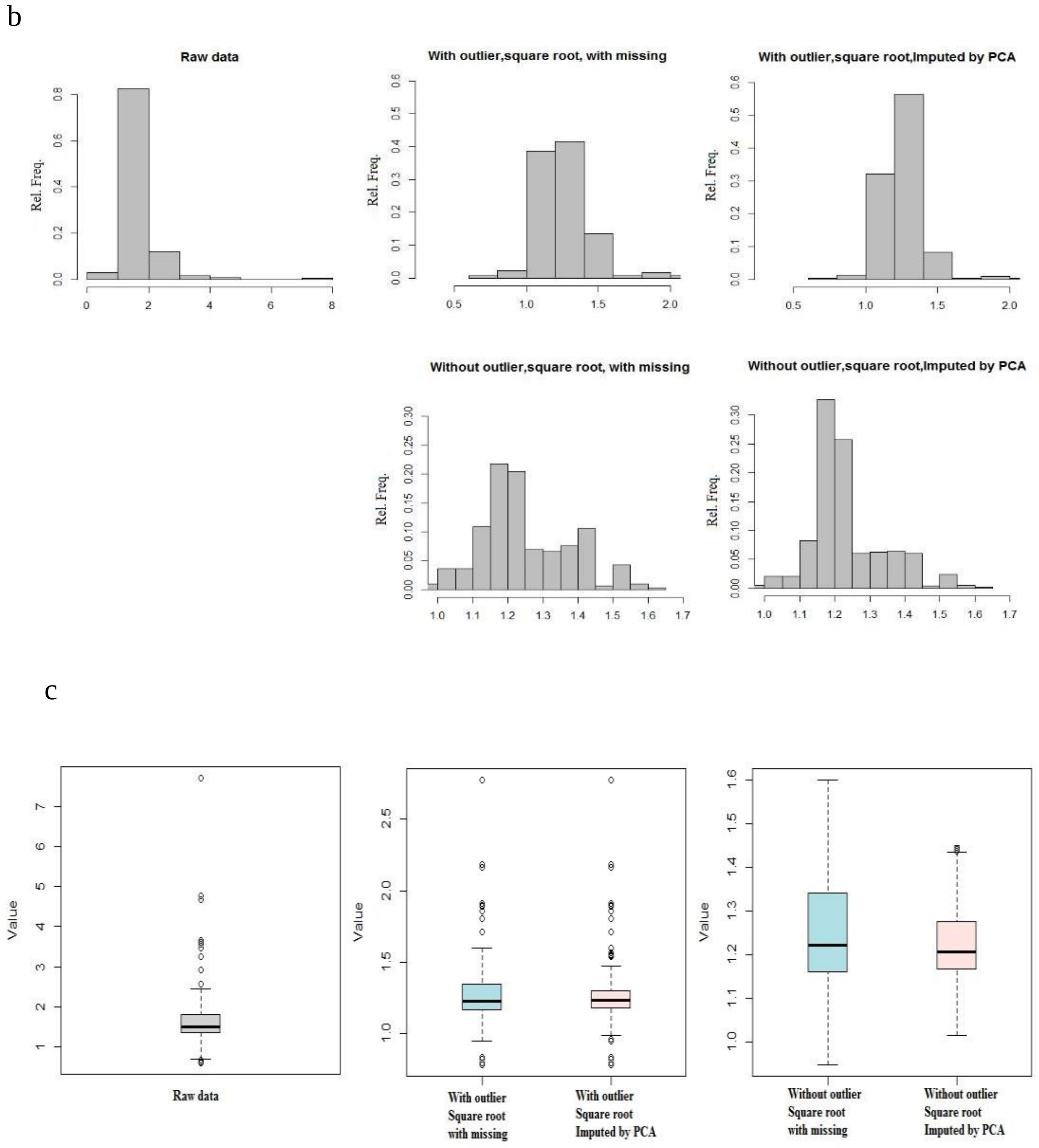
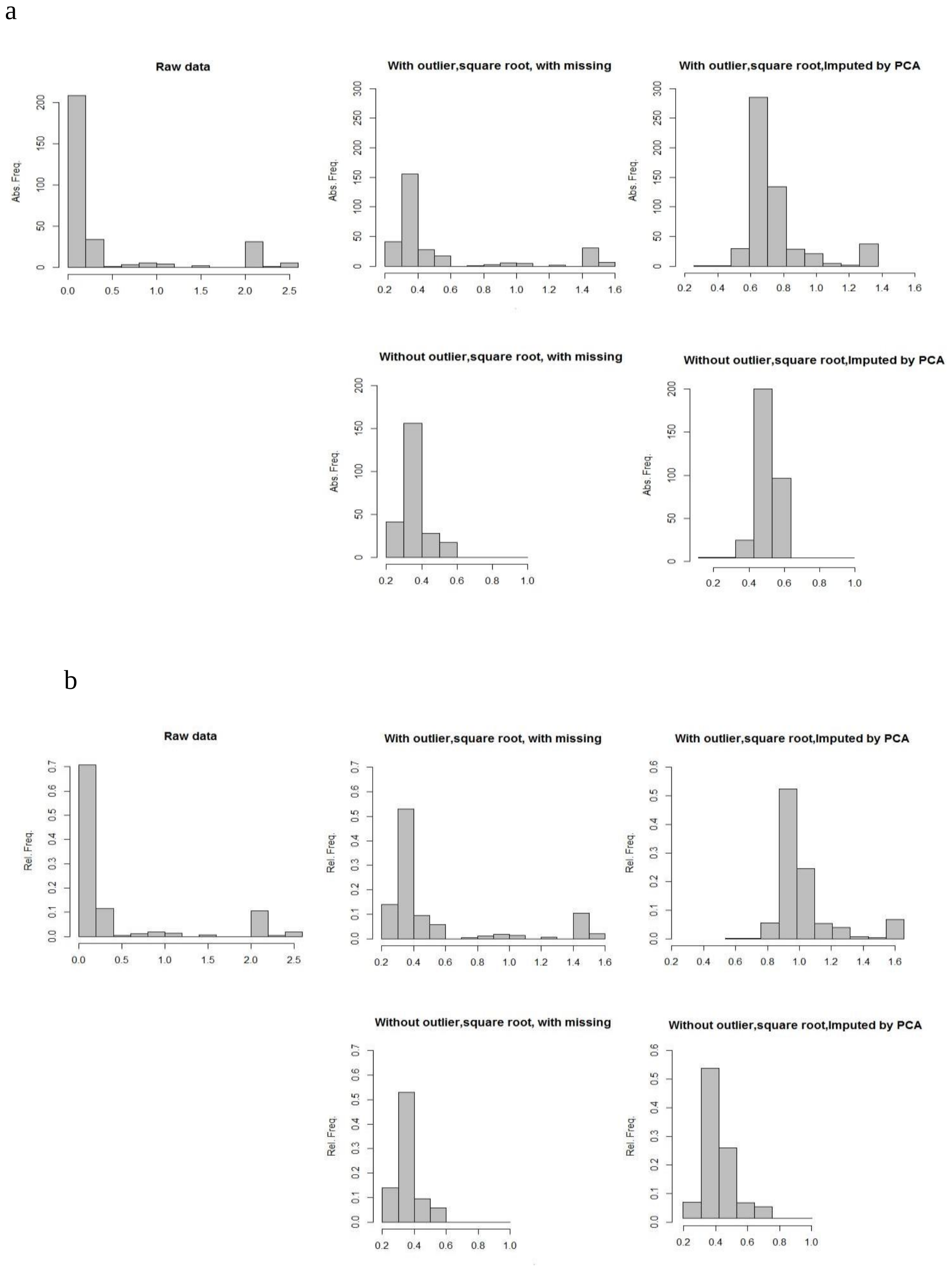


Fig. 18. a) Absolute count, b) Relative frequency histograms and c) boxplot of the Silicate abundance with 15% missing data that is shown for five data sets.

Fig. 19 shows the analysis of Orthop abundance. We can see the highest frequency of Orthop abundance is between the ranges 0.1204 and 1.5897 of observations (Fig. 19a and b). Boxplot of raw data shows outliers at the upper range of the data, the mean value (0.5130) is above the median (0.4240), the median line does not evenly divide the box (almost), and the upper tail of the boxplot is longer than the lower tail, then the distribution of which the data were sampled may be skewed to the right (Fig. 19 c).


a
Raw data
Abs.Freq.
0
50
100
150
200
0.0
0.5
1.0
1.5
2.0
2.5
With outlier,square root, with missing
Abs.Freq.
0
50
100
150
200
250
300
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
With outlier,square root,Imputed by PCA
Abs.Freq.
0
50
100
150
200
250
300
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Without outlier,square root, with missing
Abs.Freq.
0
50
100
150
200
0.2
0.4
0.6
0.8
1.0
Without outlier,square root,Imputed by PCA
Abs.Freq.
0
50
100
150
200
0.2
0.4
0.6
0.8
1.0
b
Raw data
Rel.Freq.
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.0
0.5
1.0
1.5
2.0
2.5
With outlier,square root, with missing
Rel.Freq.
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
With outlier,square root,Imputed by PCA
Rel.Freq.
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Without outlier,square root, with missing
Rel.Freq.
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.2
0.4
0.6
0.8
1.0
Without outlier,square root,Imputed by PCA
Rel.Freq.
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.2
0.4
0.6
0.8
1.0

c

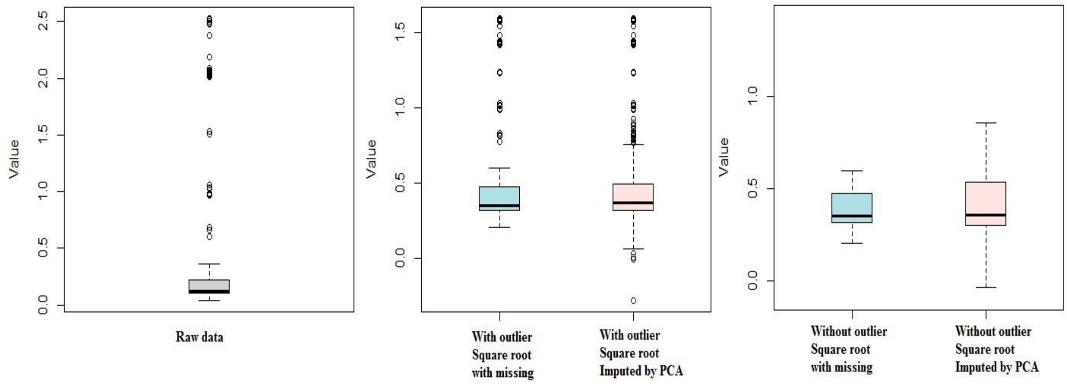
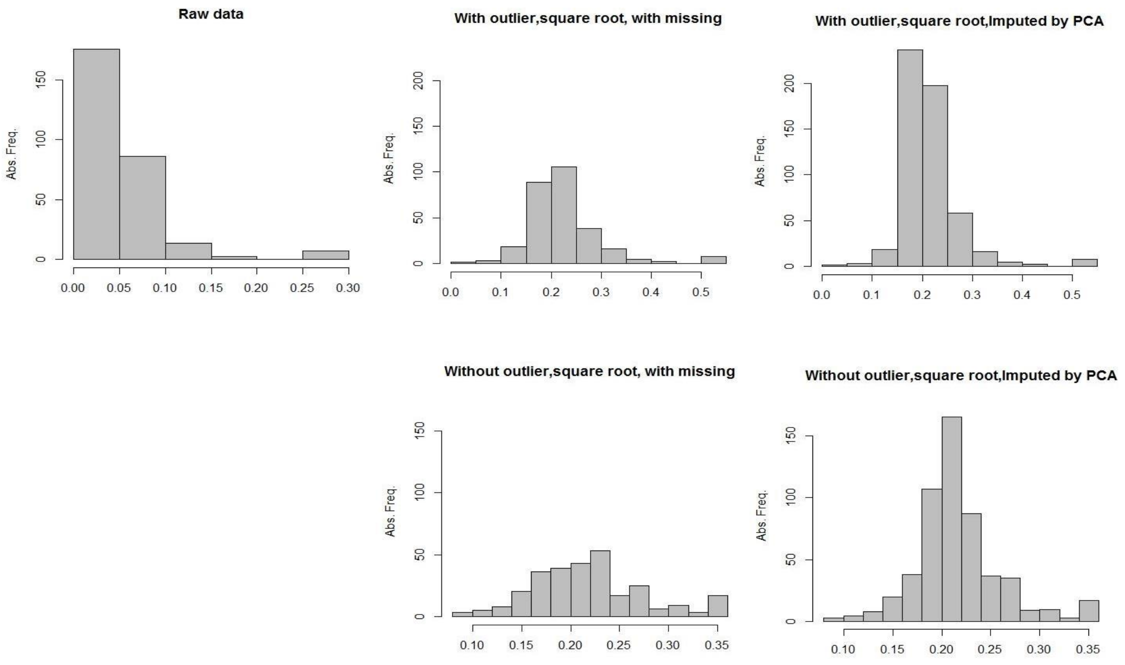


Fig. 19. a) Absolute count, b) Relative frequency histograms and c) boxplot of the Orthop abundance with 15% missing data that is shown for five data sets.

Fig. 20 shows the analysis of Nitrit abundance. We can see the highest frequency of Nitrit abundance is between the ranges 0.03162 and 0.53852 of observations (Fig. 20a and b). Boxplot of raw data shows outliers at the upper range of the data, the mean value(0.22121) is above the median (0.21762), the median line does not evenly divide the box(almost), and the upper tail of the boxplot is longer than the lower tail, then thedistribution of which the data were sampled may be skewed to the right (Fig. 20 c).

a



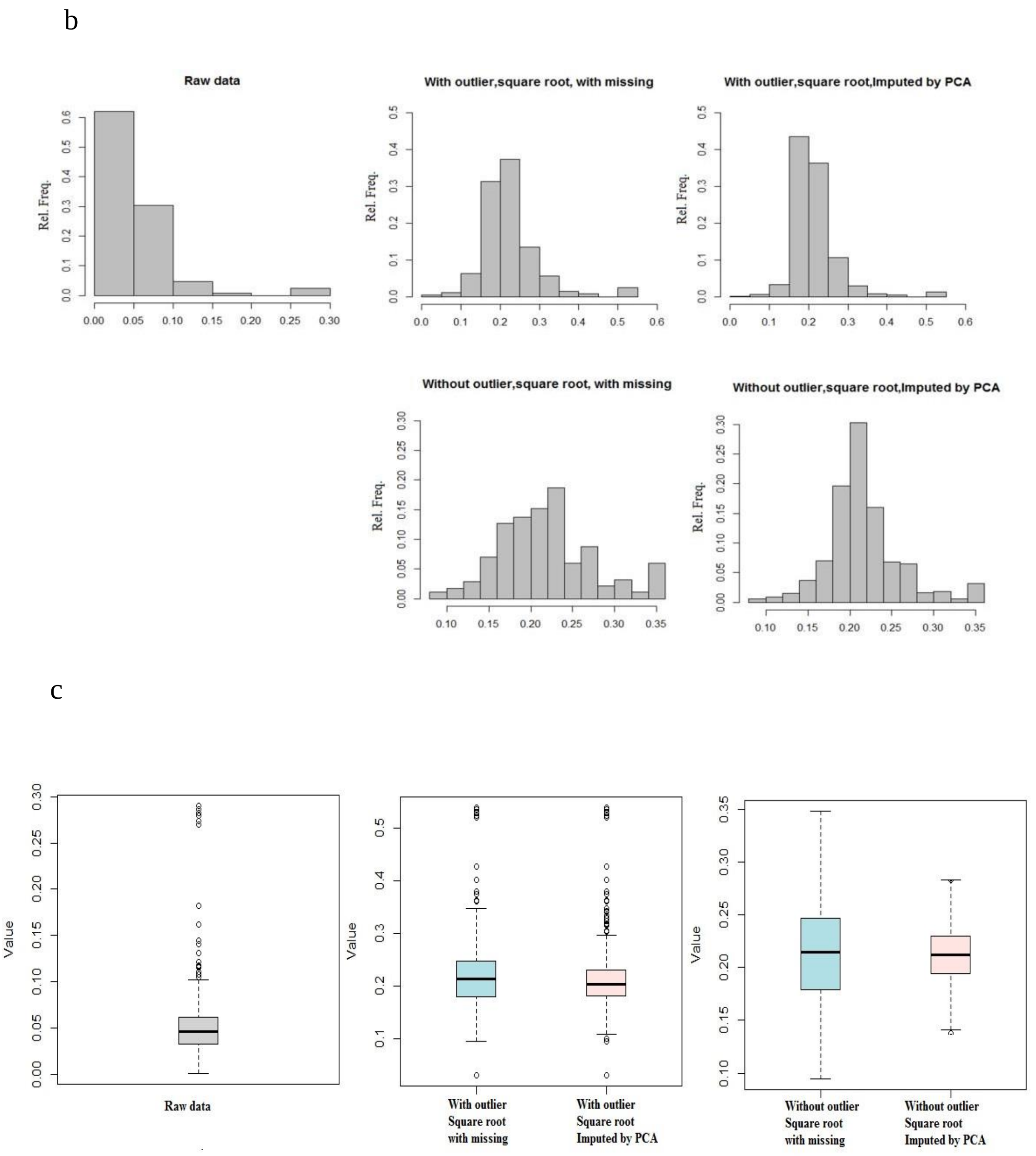


Fig. 20. a) Absolute count, b) Relative frequency histograms and c) boxplot of the Nitrit abundance with 15% missing data that is shown for five data sets.

Fig. 21 shows the analysis of NH3 abundance. We can see the highest frequency of NH3 abundance is between the ranges 0 and 1.0344 of observations (Fig. 21a and b). Boxplot of raw data shows outliers at the upper range of the data, the mean value (0.2497) is above the median (0.2236), the median line does not evenly divide the box (almost), and the upper tail of the boxplot is longer than the lower tail, then the distribution of which the data were sampled may be skewed to the right (Fig. 21 c).

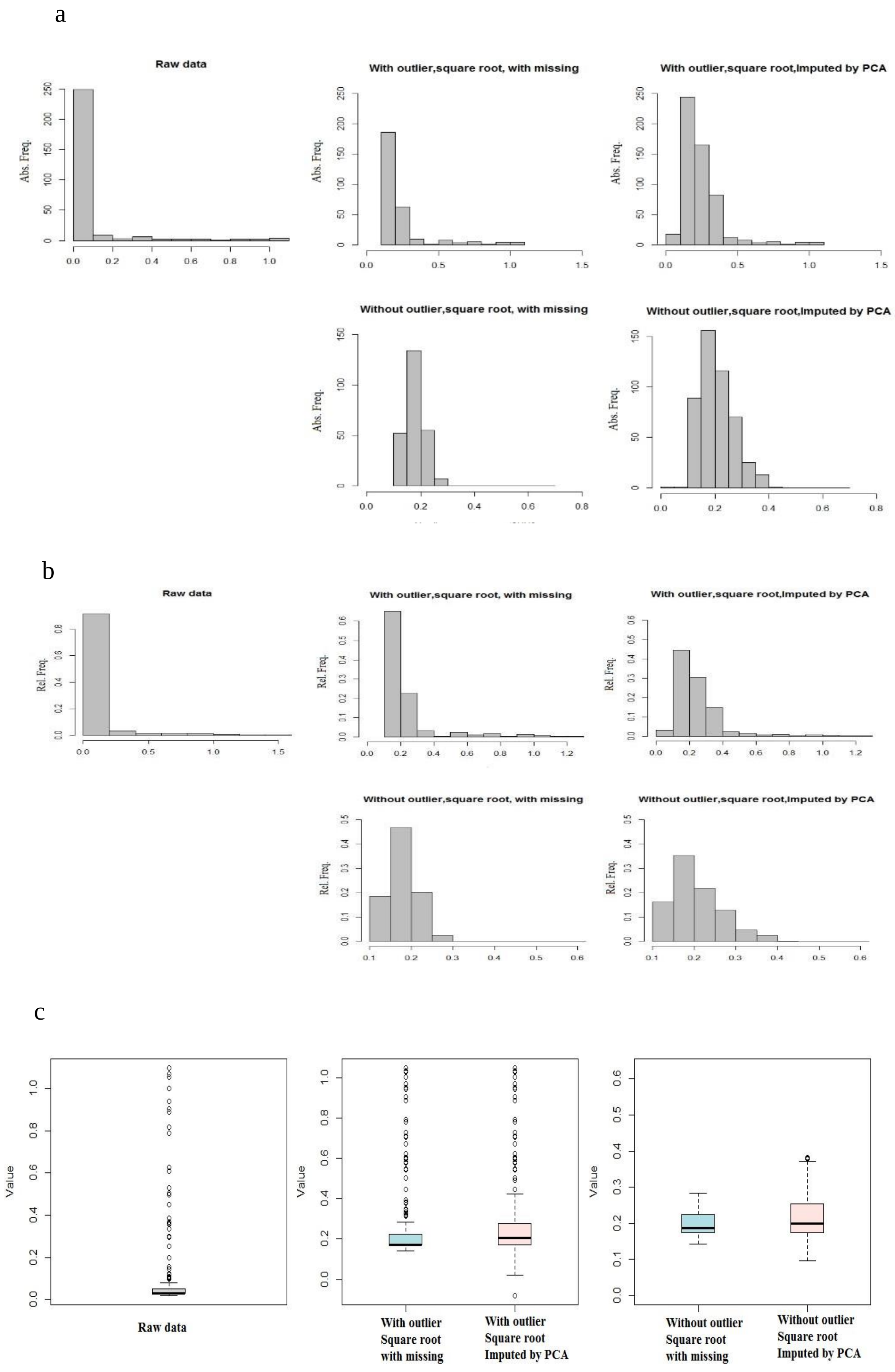
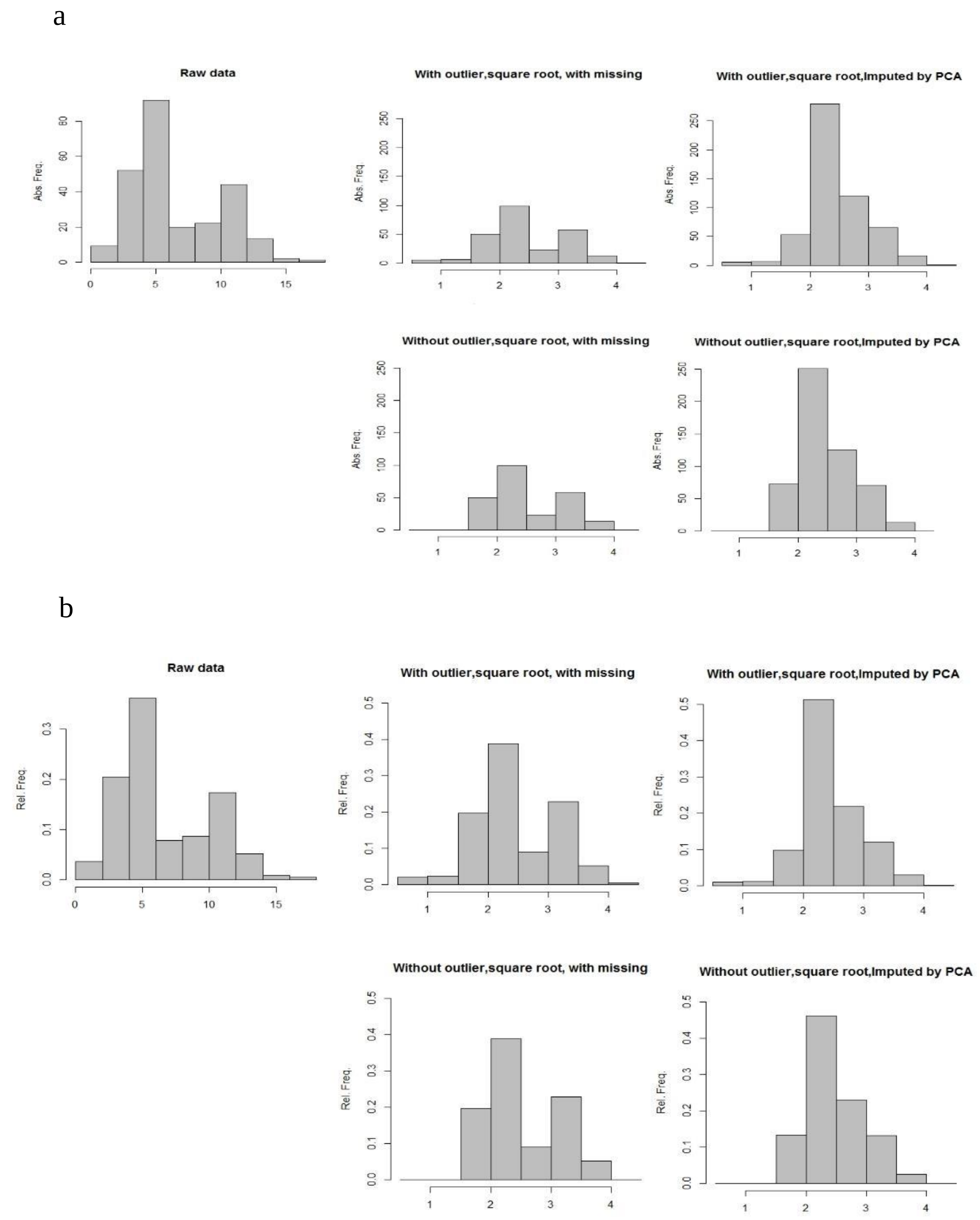


Fig. 21. a) Absolute count, b) Relative frequency histograms and c) boxplot of the NH3 abundance with 15% missing data that is shown for five data sets.

Fig. 22 shows the analysis of TN abundance. We can see the highest frequency of TN abundance is between the ranges 0 and 1.0344 of observations (Fig. 22a and b). Boxplot of raw data shows outliers at the upper range of the data, the mean value (0.2497) is above the median (0.2236), the median line does not evenly divide the box, and the upper tail of the boxplot is longer than the lower tail, then the distribution of which the data were sampled may be skewed to the right (Fig. 22 c).



c

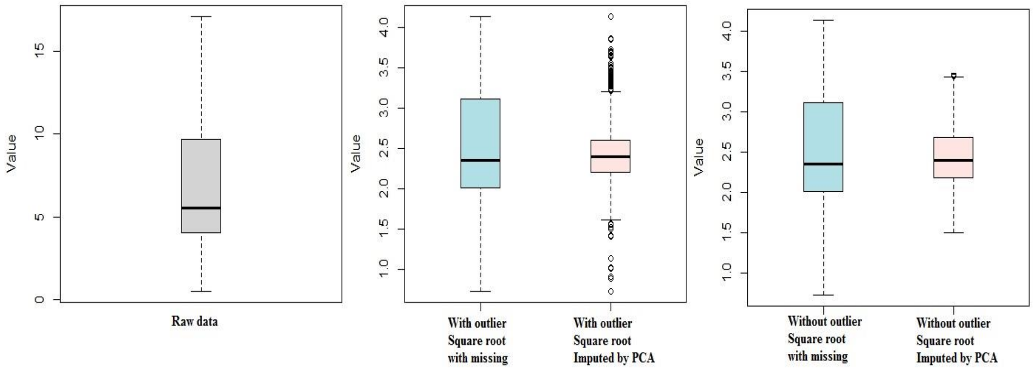
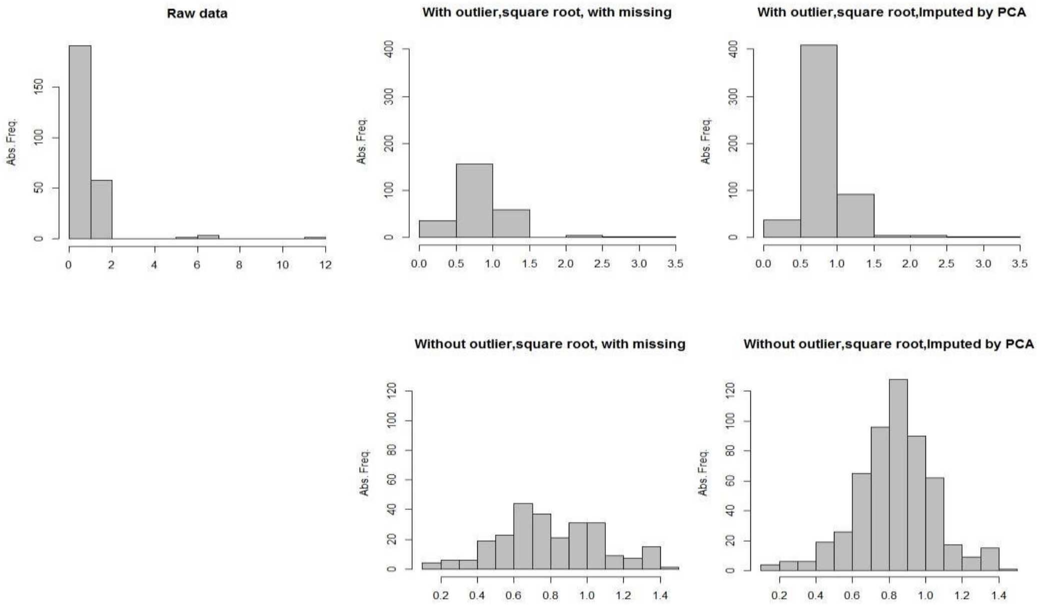
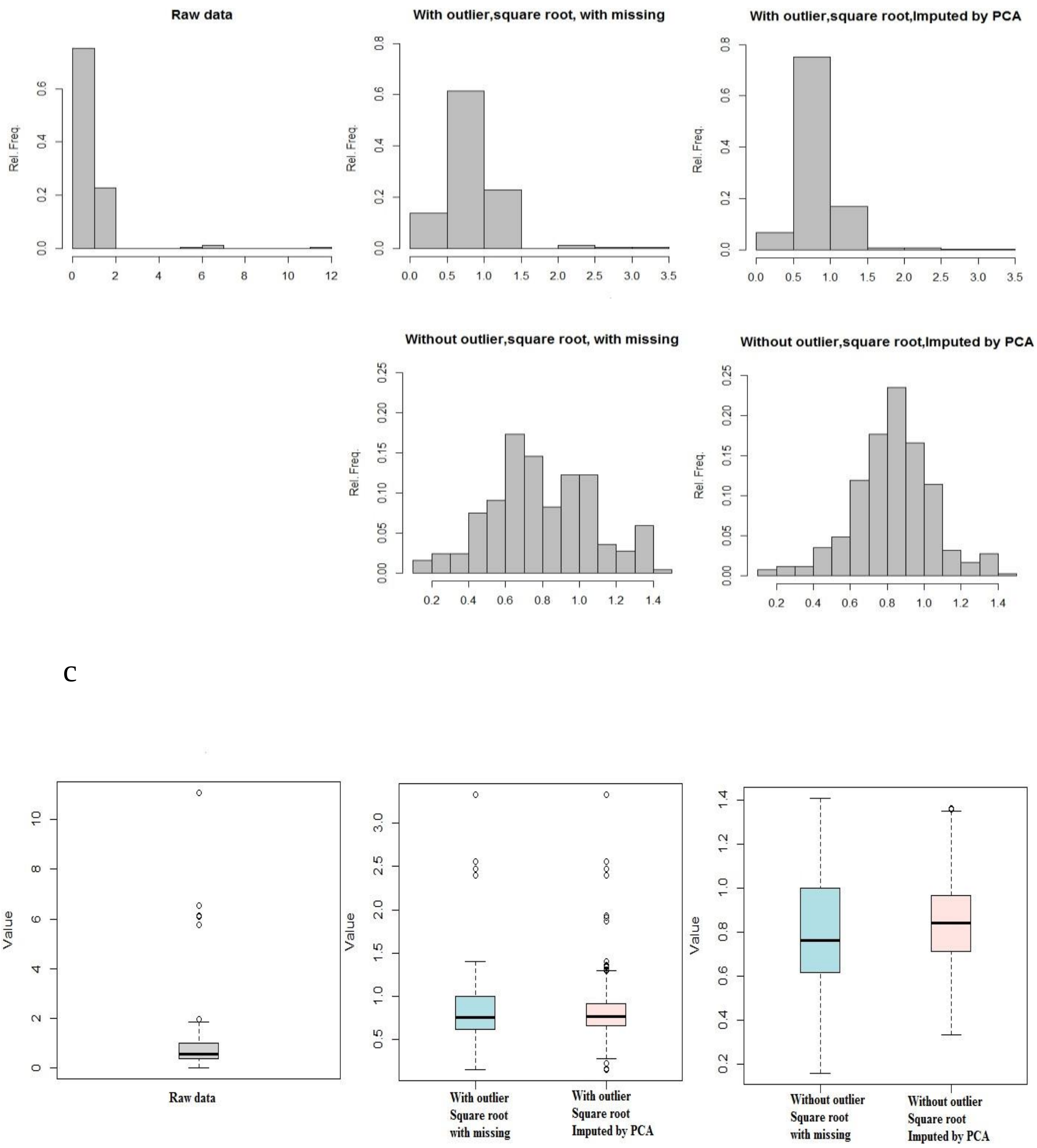


Fig. 22. a) Absolute count, b) Relative frequency histograms and c) boxplot of the TN abundance with 15% missing data that is shown for five data sets.

Fig. 23 shows the analysis of Nitrate abundance. We can see the highest frequency of Nitrate abundance is between the ranges 0.1581 and 3.3274 of observations (Fig.23a and b). Boxplot of raw data shows outliers at the upper range of the data (above the box), the mean value (0.8351) and the median (0.8370), they are almost equal, the median line does evenly divide the box (normal distributions), (Fig. 23 c).

a



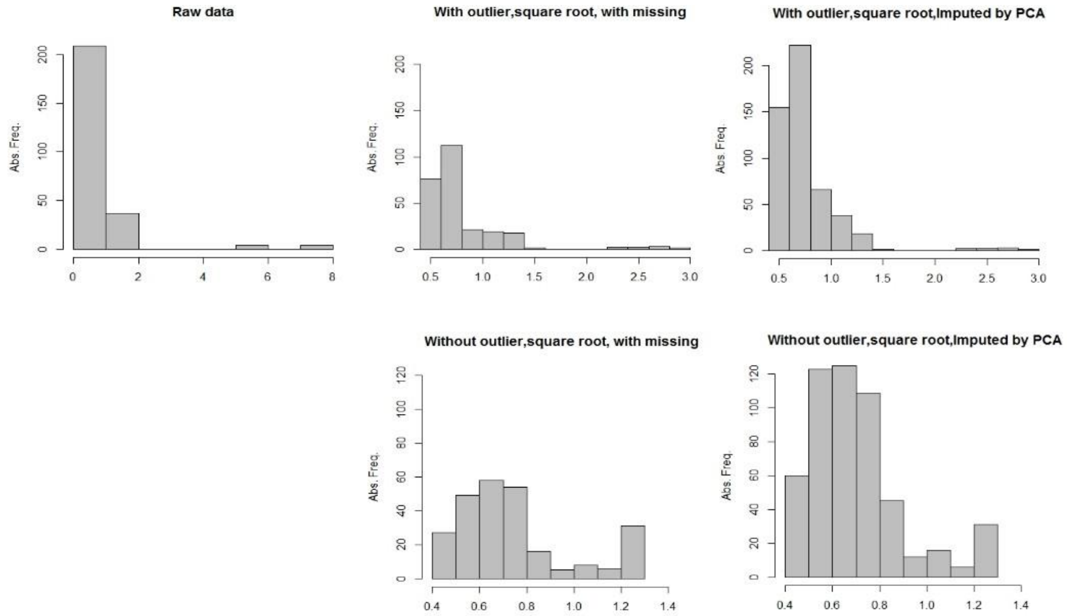

b

c

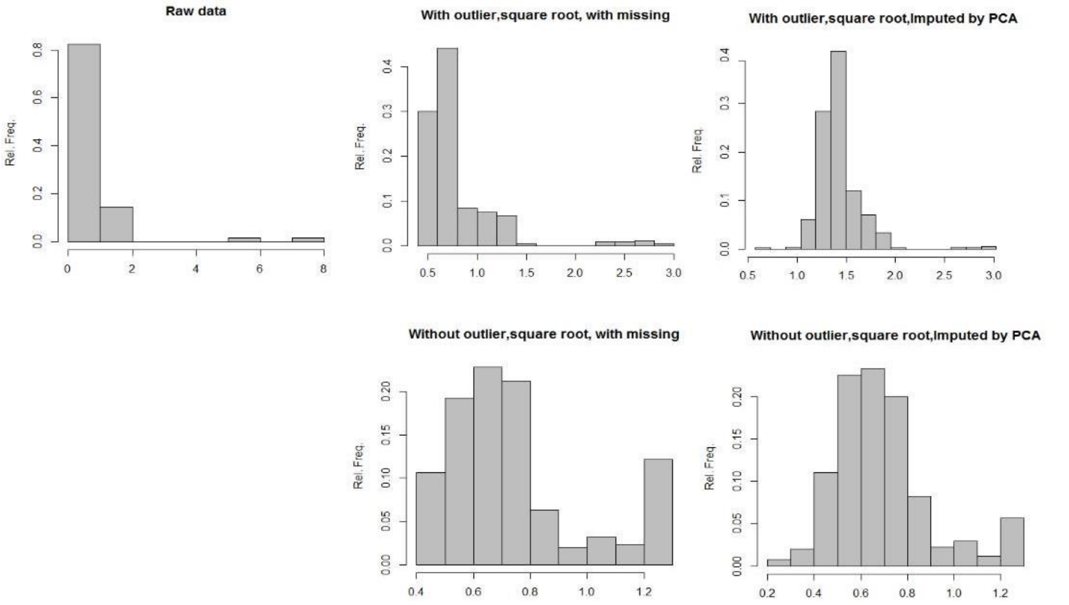
Fig. 23. a) Absolute count, b) Relative frequency histograms and c) boxplot of the Nitrate abundance with 15% missing data that is shown for five data sets.

Fig. 24 shows the analysis of TP abundance. We can see the highest frequency of TP abundance is between the ranges 0.3134 and 2.8073of observations (Fig. 24a and b). Boxplot of raw data shows outliers at the upper range of the data, the mean value (0.7640) is above the median (0.7051), the median line does not evenly divide the box, and the upper tail of the boxplot is longer than the lower tail, then the distribution of which the data were sampled may be skewed to the right (Fig. 24 c).

a


Raw data
Abs.Freq.
0
50
100
150
200
0
2
4
6
8
With outlier,square root, with missing
Abs.Freq.
0
50
100
150
200
0.5
1.0
1.5
2.0
2.5
3.0
With outlier,square root,Imputed by PCA
Abs.Freq.
0
50
100
150
200
0.5
1.0
1.5
2.0
2.5
3.0
Without outlier,square root, with missing
Abs.Freq.
0
20
40
60
80
100
120
0.4
0.6
0.8
1.0
1.2
1.4
Without outlier,square root,Imputed by PCA
Abs.Freq.
0
20
40
60
80
100
120
0.4
0.6
0.8
1.0
1.2
1.4

b


Raw data
Rel.Freq.
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
With outlier,square root, with missing
Rel.Freq.
0.0
0.1
0.2
0.3
0.4
0.5
1.0
1.5
2.0
2.5
3.0
With outlier,square root,Imputed by PCA
Rel.Freq.
0.0
0.1
0.2
0.3
0.4
0.5
1.0
1.5
2.0
2.5
3.0
Without outlier,square root, with missing
Rel.Freq.
0.00
0.05
0.10
0.15
0.20
0.4
0.6
0.8
1.0
1.2
Without outlier,square root,Imputed by PCA
Rel.Freq.
0.00
0.05
0.10
0.15
0.20
0.2
0.4
0.6
0.8
1.0
1.2

c

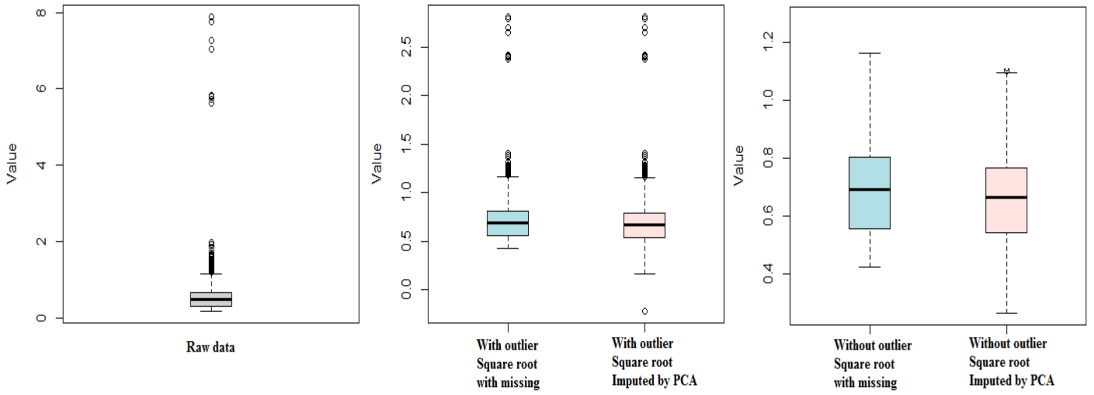
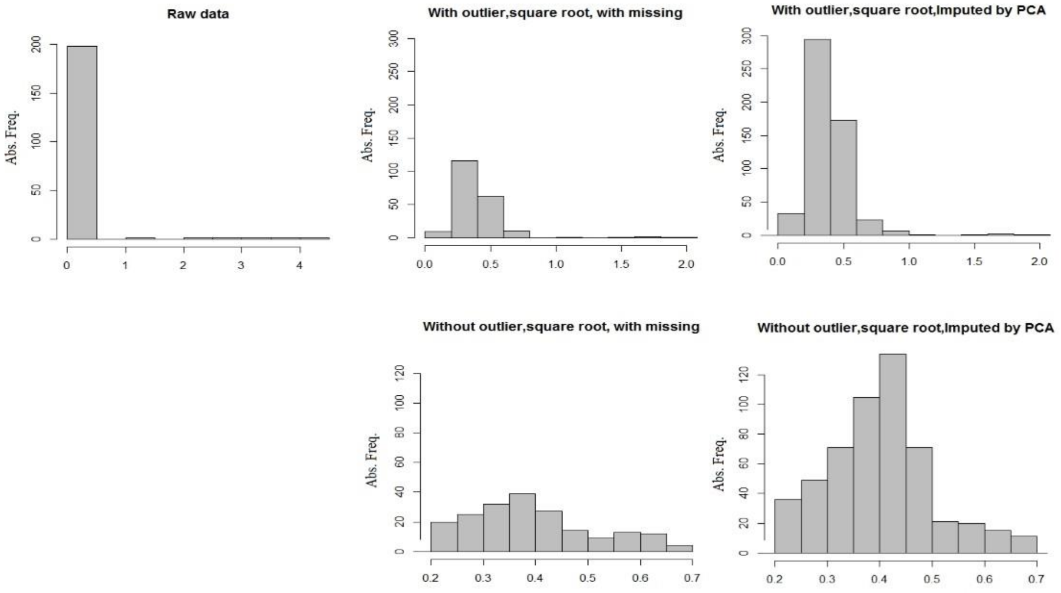


Fig. 24. a) Absolute count, b) Relative frequency histograms and c) boxplot of the TP abundance with 15% missing data that is shown for five data sets.

Fig. 25 shows the analysis of Feof abundance. We can see the highest frequency of Feof abundance is between the ranges 0 and 2.0543 of observations (Fig. 25a and b). Boxplot of raw data shows outliers at the upper range of the data, the mean value (0.4243) is above the median (0.4123), the median line almost does evenly divide the box, and the upper tail of the boxplot is longer than the lower tail, then the distribution of which the data were sampled may be skewed to the right (Fig. 25 c).

a



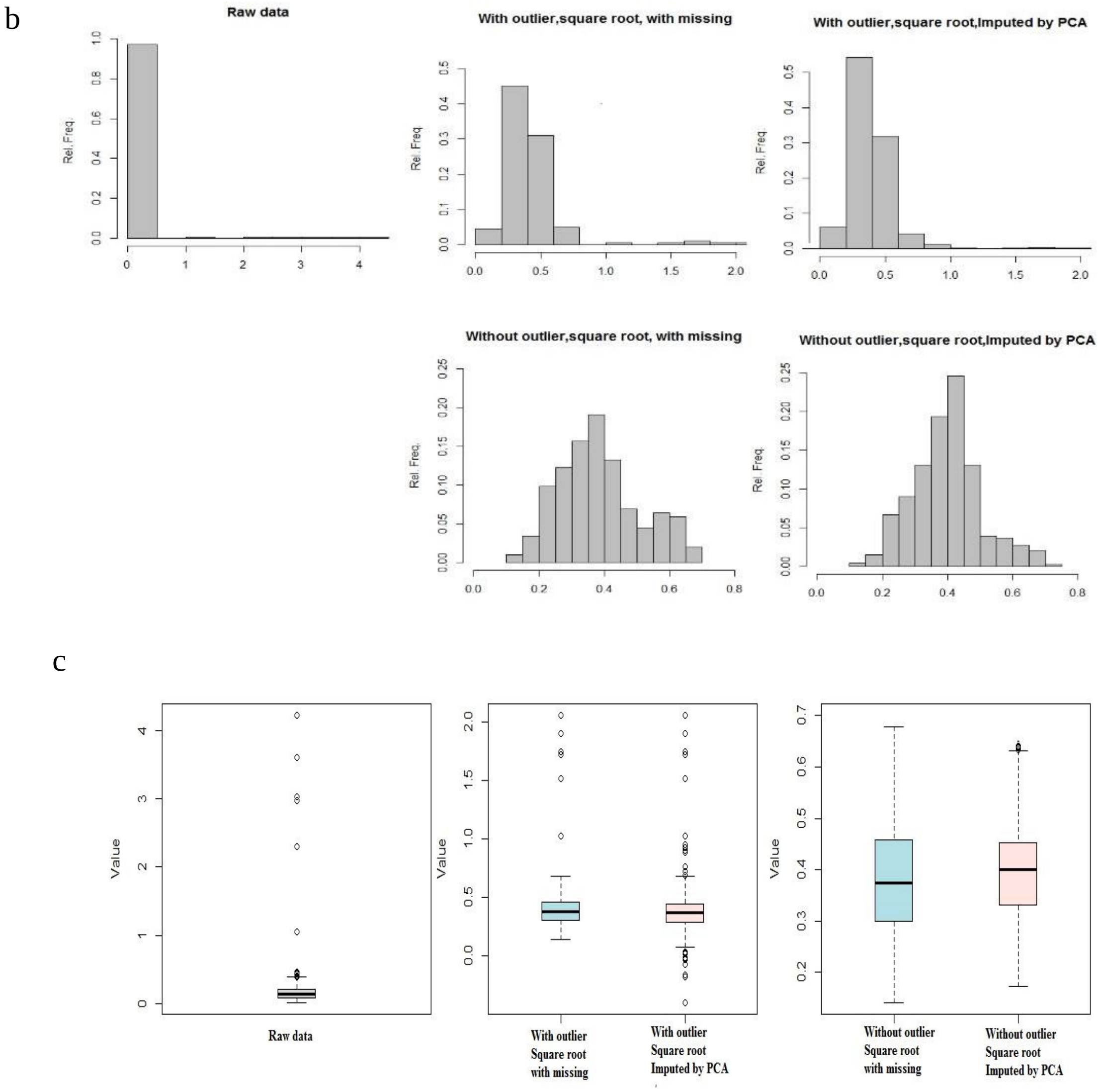


Fig. 25. a) Absolute count, b) Relative frequency histograms and c) boxplot of the Feof abundance with 15% missing data that is shown for five data sets.

Fig. 26 shows the analysis of Cl-a abundance. We can see the highest frequency of Cl-a abundance is between the ranges 0 and 2.3452 of observations (Fig. 26a and b). Boxplot of raw data shows outliers at the upper range of the data, the mean value (0.5361) is above the median (0.5208), the median line almost does evenly divide the box, and the upper tail of the boxplot is longer than the lower tail, then the distribution of which the data were sampled may be skewed to the right (Fig. 26 c).

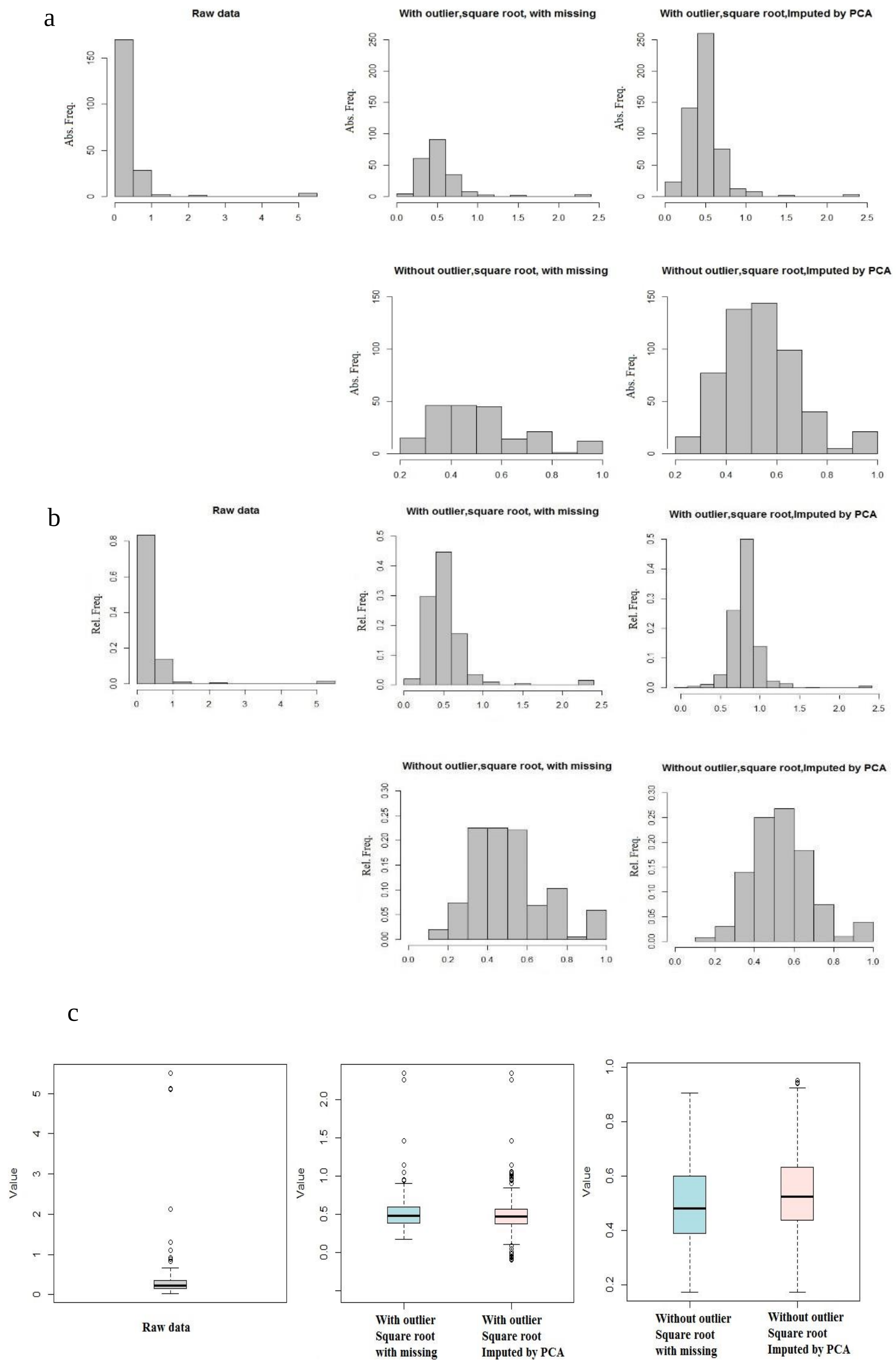
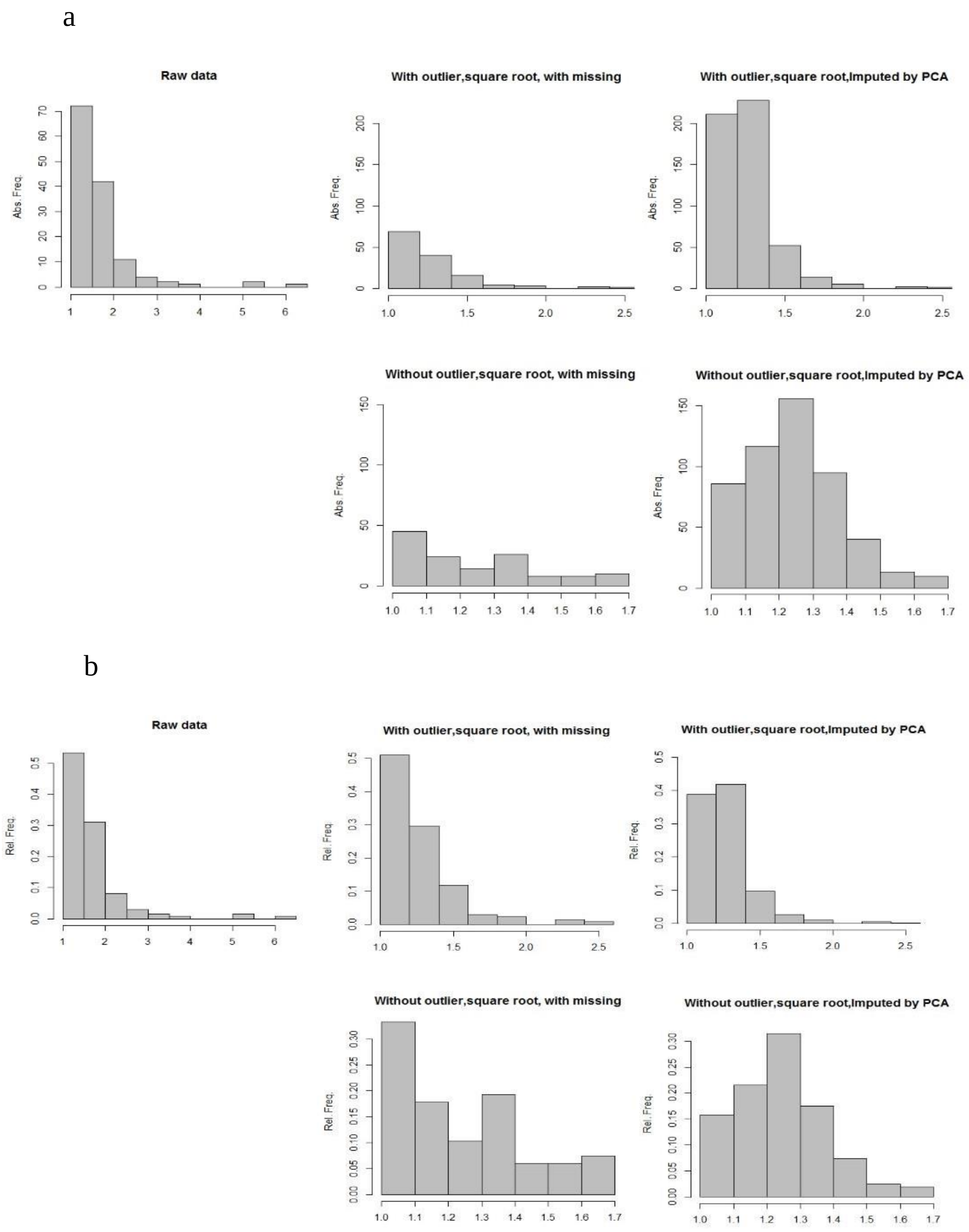


Fig. 26. a) Absolute count, b) Relative frequency histograms and c) boxplot of the Feof abundance with 15% missing data that is shown for five data sets.

Fig. 27 shows the analysis of DOC abundance. We can see the highest frequency of DOC abundance is between the ranges 0.5326 and 2.5120 of observations (Fig. 27a and b). Boxplot of raw data shows outliers at the upper range of the data, the mean value (1.2487) is above the median (1.2429), the median line almost does evenly divide the box, and the upper tail of the boxplot is longer than the lower tail, then the distribution of which the data were sampled may be skewed to the right (Fig. 27 c).



c

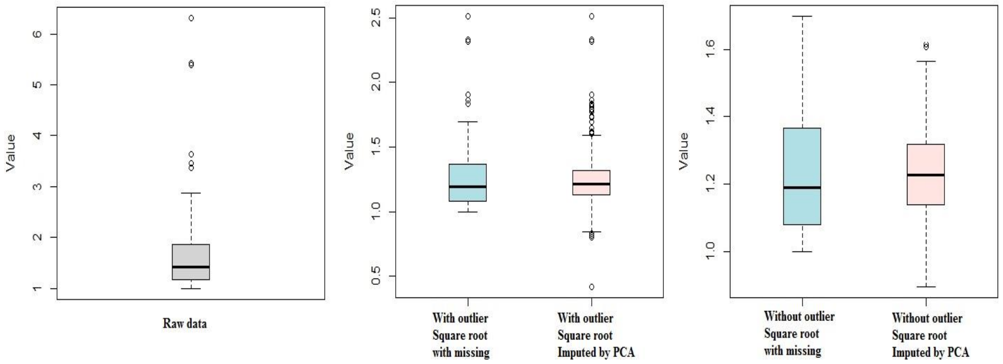
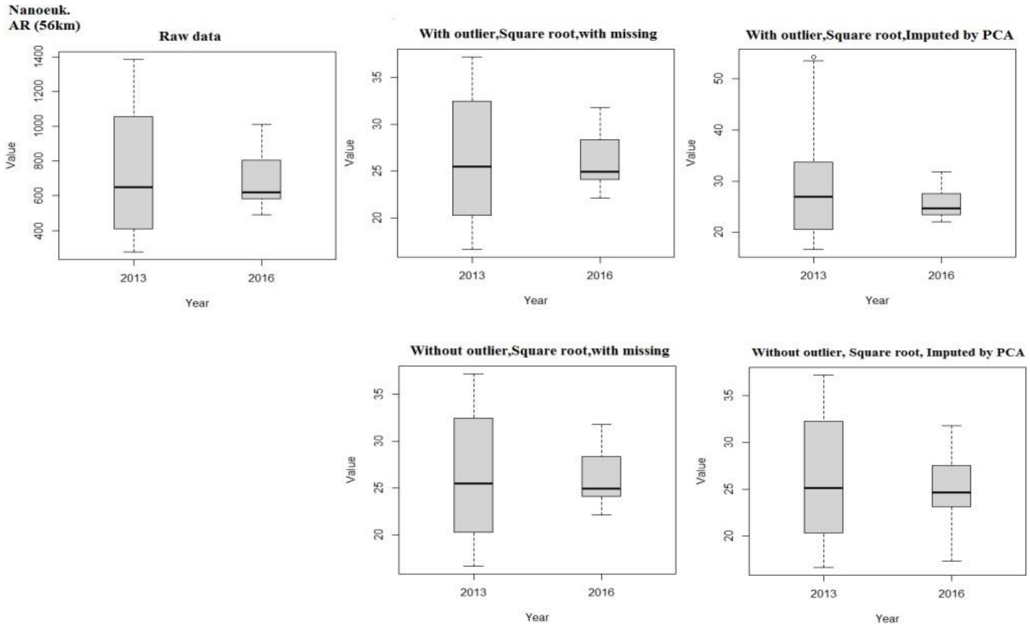


Fig. 27. a) Absolute count, b) Relative frequency histograms and c) boxplot of the DOC abundance with 15% missing data that is shown for five data sets.

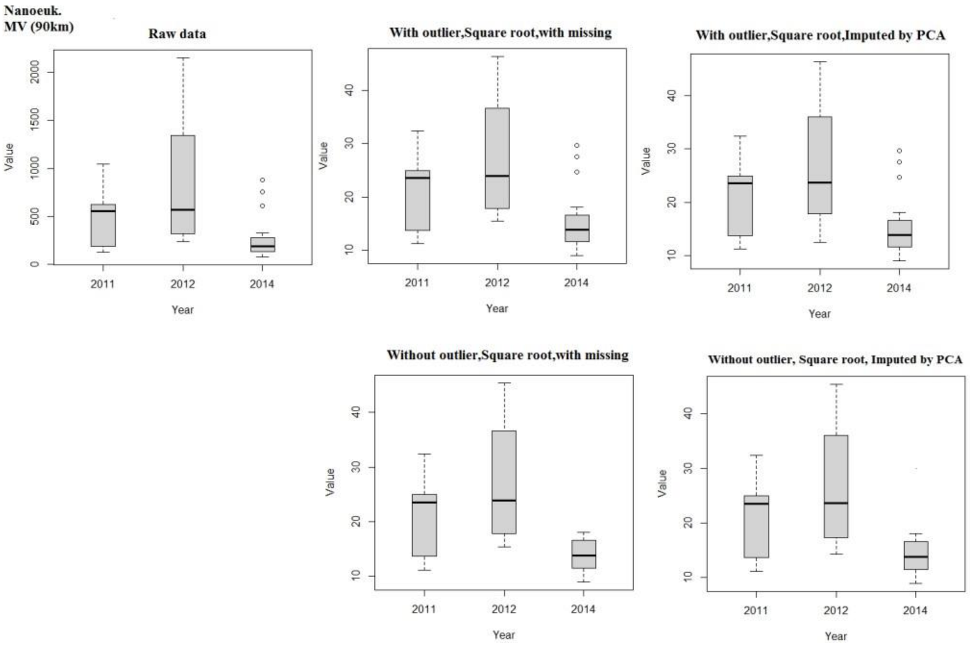
Boxplots of each variable in sites and in each sites per years were presented in Figs. 28-45. The difference of graphs scale between "with outlier" and "without outlier" defined to better show the shape of the distribution.

Fig. 28 shows the distribution of Nanoeuk. in different years and sites. The abundance of Nanoeuk in AR and 2013 is higher than 2016 (Fig. 28a). The abundance of Nanoeuk in MV and 2012 is higher than 2011 and 2014 (Fig. 28b). The abundance of Nanoeuk in PAB and 2014 is higher than 2012 and 2016 (Fig. 28c). The abundance of Nanoeuk in PL and 2013 is higher than 2012 and 2014 and 2016 (Fig. 28d). The abundance of Nanoeuk in SG and 2014 is higher than 2012 and 2016 (Fig. 28e). The abundance of Nanoeuk in TIM and 2013 is higher than 2012 and 2014 (Fig. 28f).

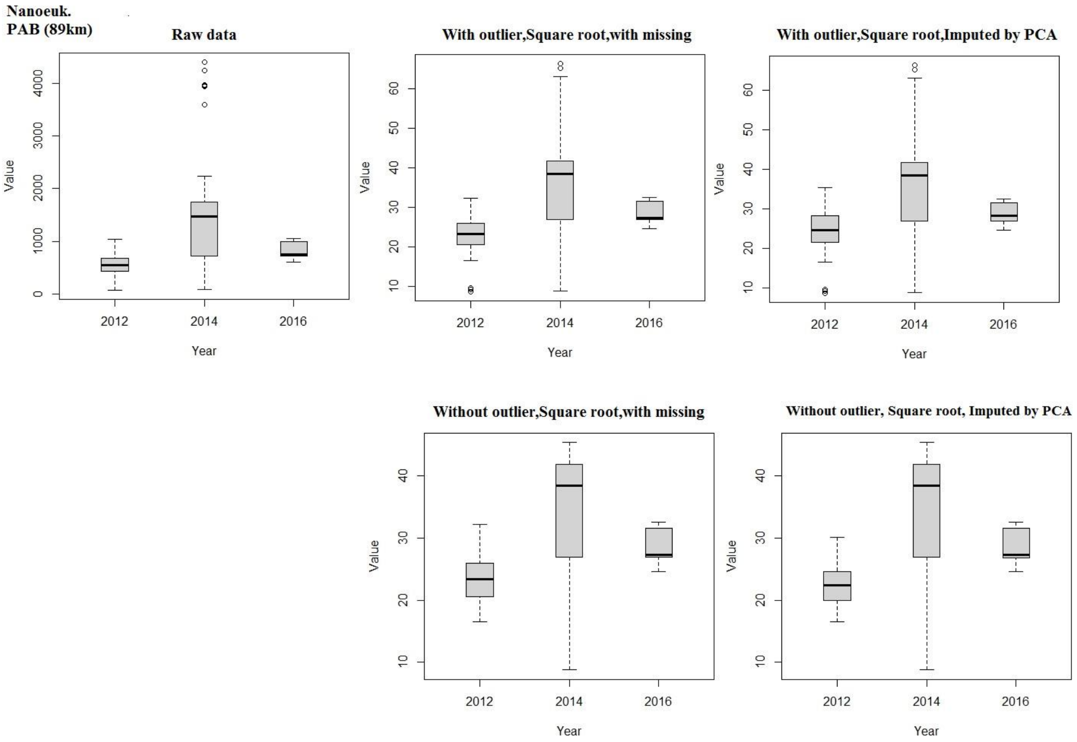
a



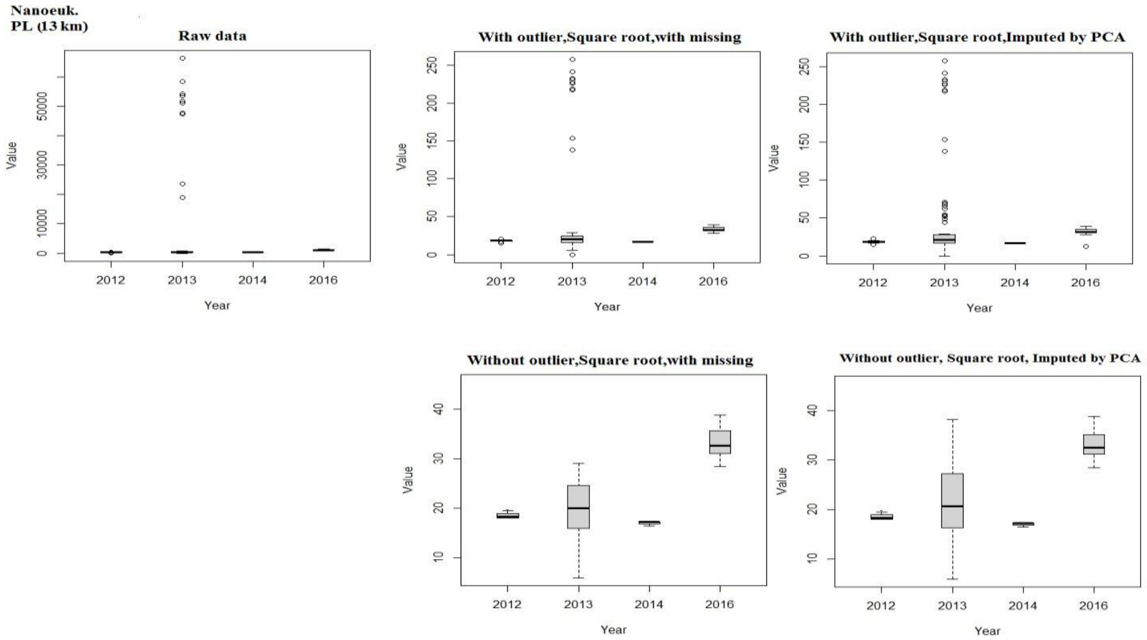
b


Nanoeuk.
MV (90km)
Raw data
With outlier,Square root,with missing
With outlier,Square root,Imputed by PCA
Without outlier,Square root,with missing
Without outlier, Square root, Imputed by PCA
Value
Year
2011
2012
2014
0
500
1000
1500
2000
10
20
30
40

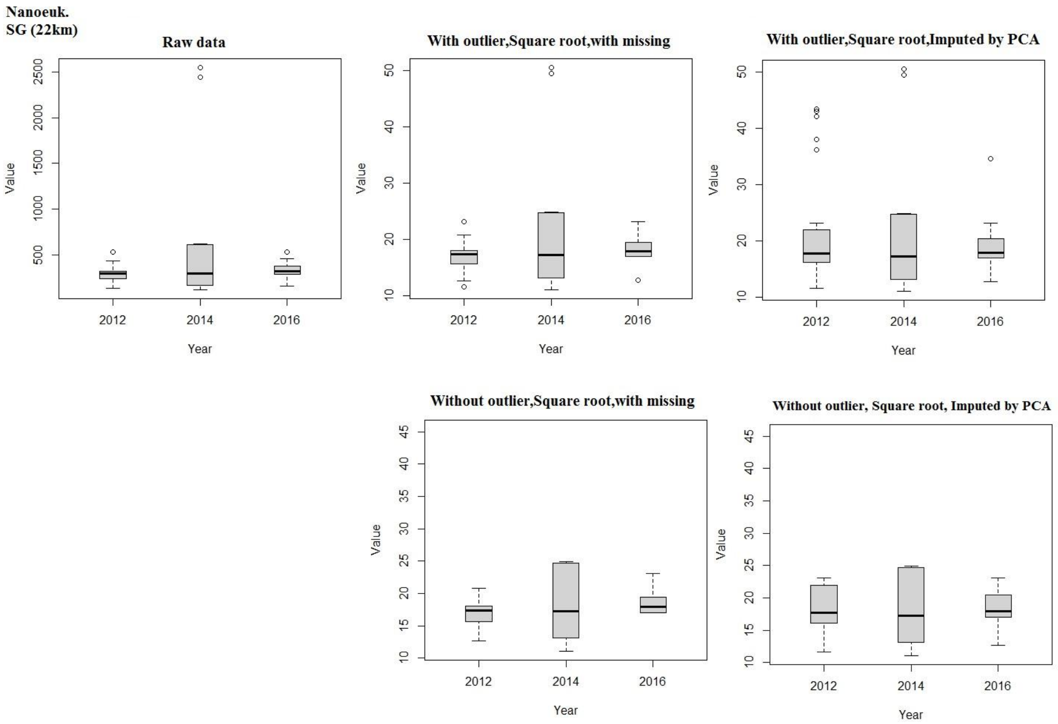
c


Nanoeuk.
PAB (89km)
Raw data
With outlier,Square root,with missing
With outlier,Square root,Imputed by PCA
Without outlier,Square root,with missing
Without outlier, Square root, Imputed by PCA
Value
Year
2012
2014
2016
0
1000
2000
3000
4000
10
20
30
40
50
60

d


Nanoeuk.
PL (13 km)
Raw data
Value
0
10000
30000
50000
2012
2013
2014
2016
Year
With outlier,Square root,with missing
Value
0
50
100
150
200
250
2012
2013
2014
2016
Year
With outlier,Square root,Imputed by PCA
Value
0
50
100
150
200
250
2012
2013
2014
2016
Year
Without outlier,Square root,with missing
Value
10
20
30
40
2012
2013
2014
2016
Year
Without outlier, Square root, Imputed by PCA
Value
10
20
30
40
2012
2013
2014
2016
Year

e


Nanoeuk.
SG (22km)
Raw data
Value
500
1000
1500
2000
2500
2012
2014
2016
Year
With outlier,Square root,with missing
Value
10
20
30
40
50
2012
2014
2016
Year
With outlier,Square root,Imputed by PCA
Value
10
20
30
40
50
2012
2014
2016
Year
Without outlier,Square root,with missing
Value
10
15
20
25
30
35
40
45
2012
2014
2016
Year
Without outlier, Square root, Imputed by PCA
Value
10
15
20
25
30
35
40
45
2012
2014
2016
Year

f

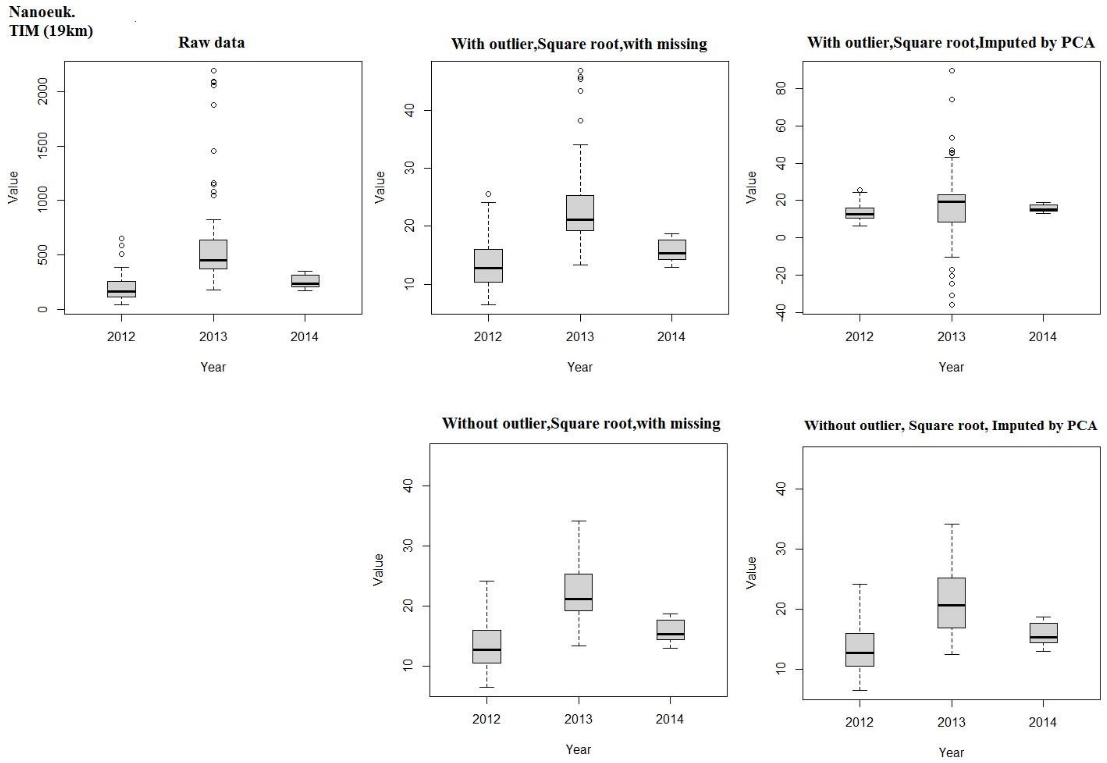
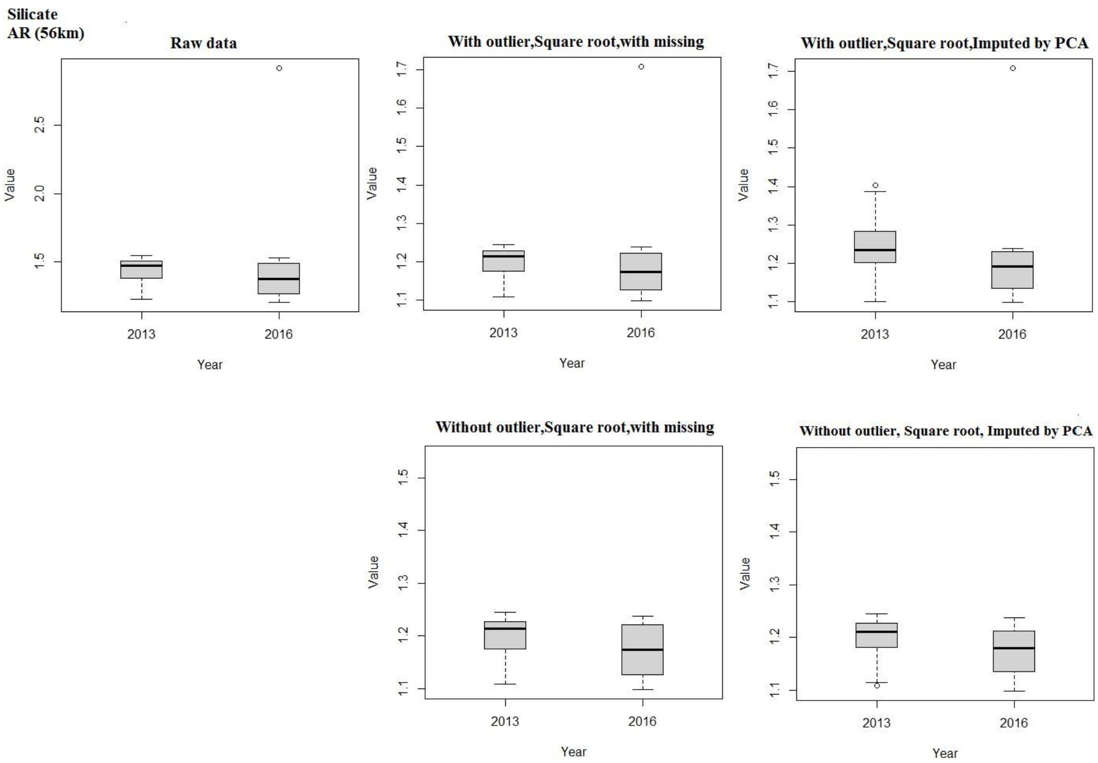


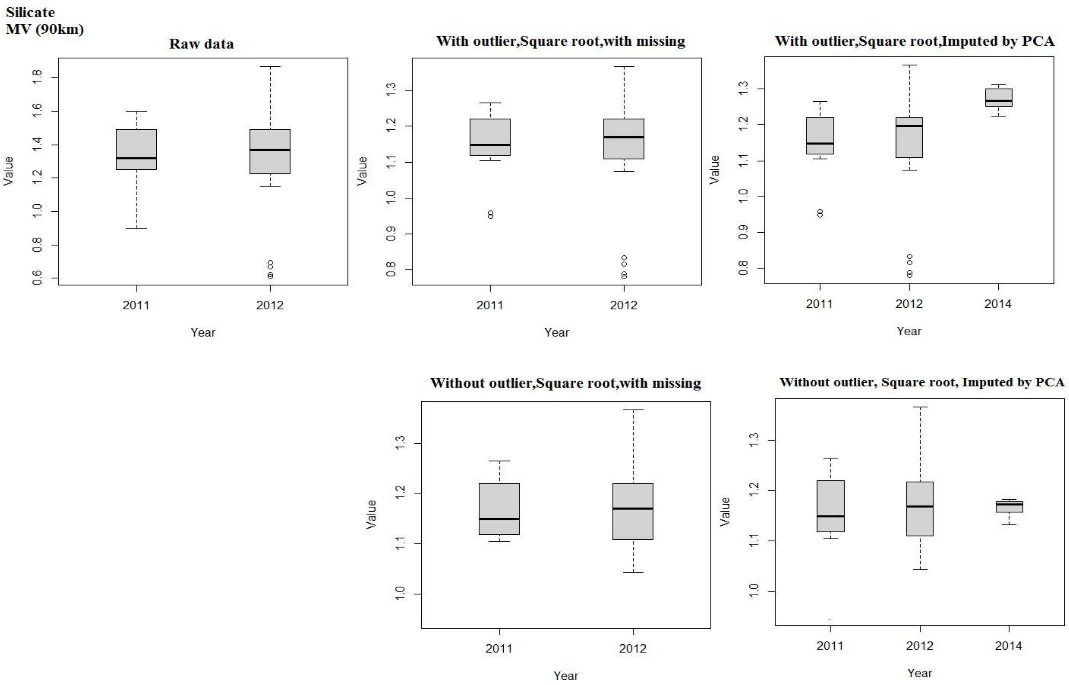
Fig. 28. Distribution of Nanoeuk. (15.81% missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

Fig. 29 shows the distribution of Silicate. in different years and sites. The abundance of Silicate in AR and 2016 is higher than 2013 (Fig. 29a). The abundance of Silicate in MV and 2012 is higher than 2011 (Fig. 29b). The abundance of Silicate in PAB and 2016 is higher than 2012 (Fig. 29c). The abundance of Silicate in PL and 2016 is higher than 2013 and 2012 (Fig. 29d). The abundance of Silicate in SG and 2012 is higher than 2016(Fig. 29e). The abundance of Silicate in TIM and 2013 is higher than 2012 (Fig. 29f).

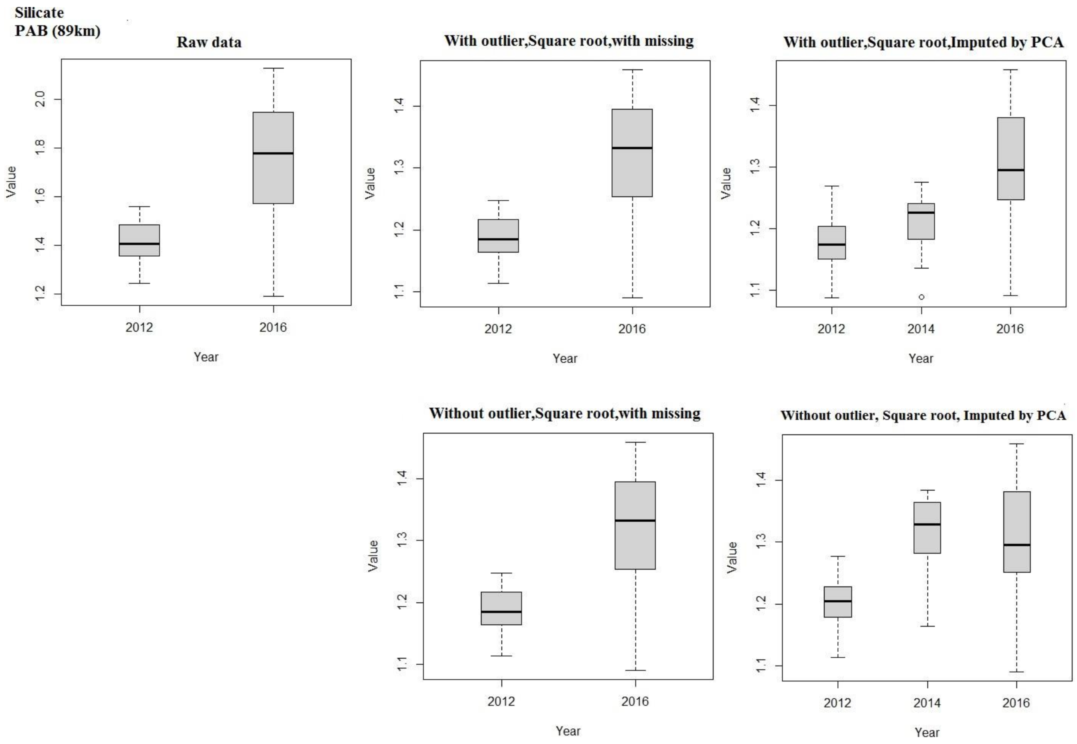
a


Silicate
AR (56km)
Raw data
With outlier,Square root,with missing
With outlier,Square root,Imputed by PCA
Without outlier,Square root,with missing
Without outlier, Square root, Imputed by PCA
Value
Year
2013
2016
1.5
2.0
2.5
1.1
1.2
1.3
1.4
1.5
1.6
1.7

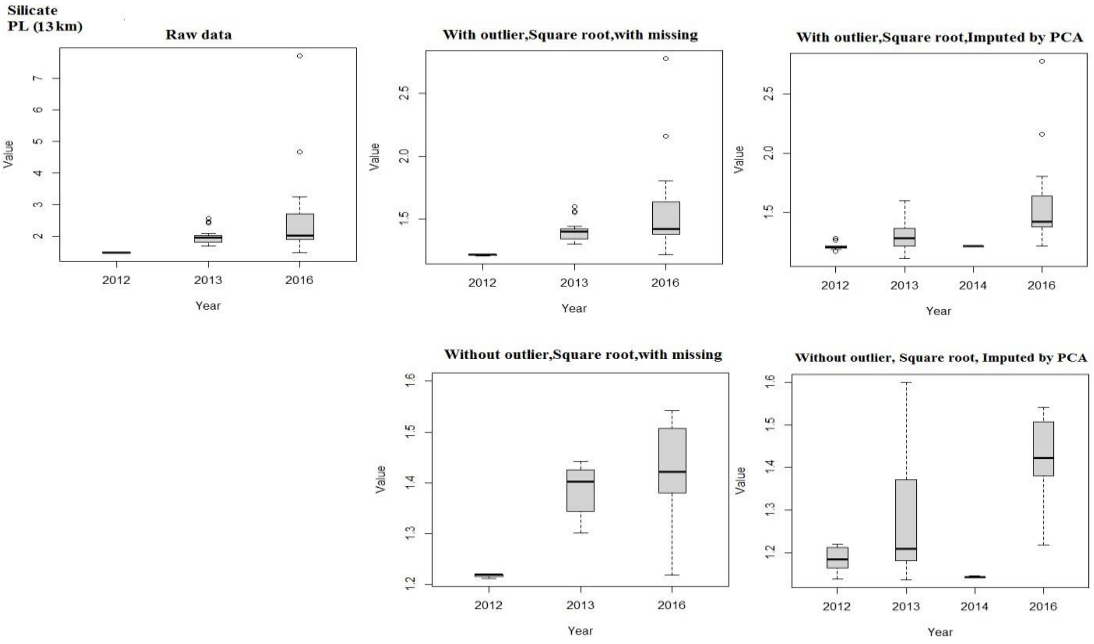
b


Silicate
MV (90km)
Raw data
With outlier,Square root,with missing
With outlier,Square root,Imputed by PCA
Without outlier,Square root,with missing
Without outlier, Square root, Imputed by PCA
Value
Year
2011
2012
2014
0.6
0.8
1.0
1.2
1.4
1.6
1.8
0.9
1.1
1.3

c


Silicate
PAB (89km)
Raw data
With outlier,Square root,with missing
With outlier,Square root,Imputed by PCA
Without outlier,Square root,with missing
Without outlier, Square root, Imputed by PCA
Value
Year
2012
2014
2016
1.2
1.4
1.6
1.8
2.0
1.1
1.3

d


Silicate
PL (13km)
Raw data
With outlier,Square root,with missing
With outlier,Square root,Imputed by PCA
Without outlier,Square root,with missing
Without outlier, Square root, Imputed by PCA
Value
Year
2012
2013
2014
2016
2
3
4
5
6
7
1.5
2.0
2.5
1.2
1.3
1.4
1.6

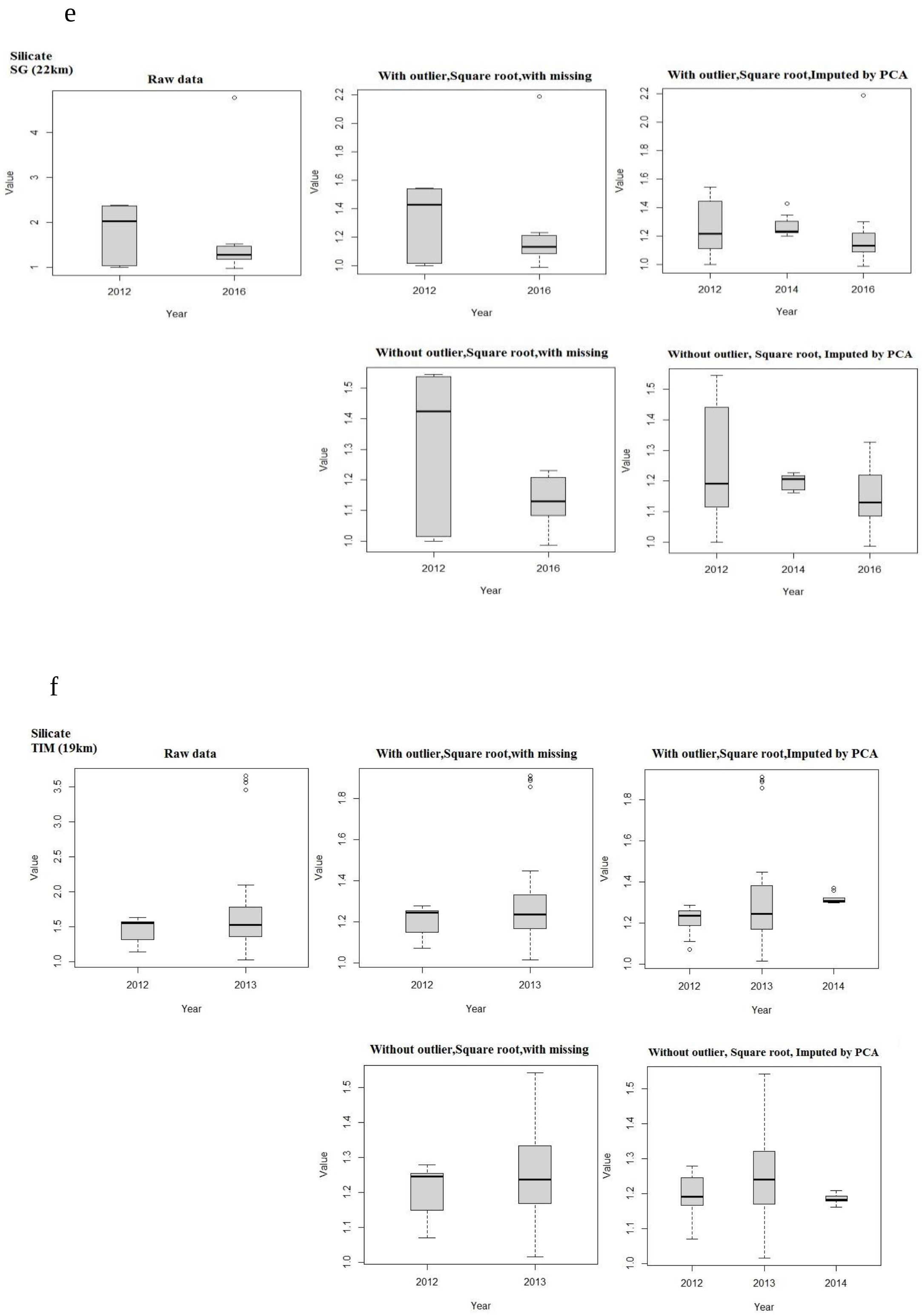
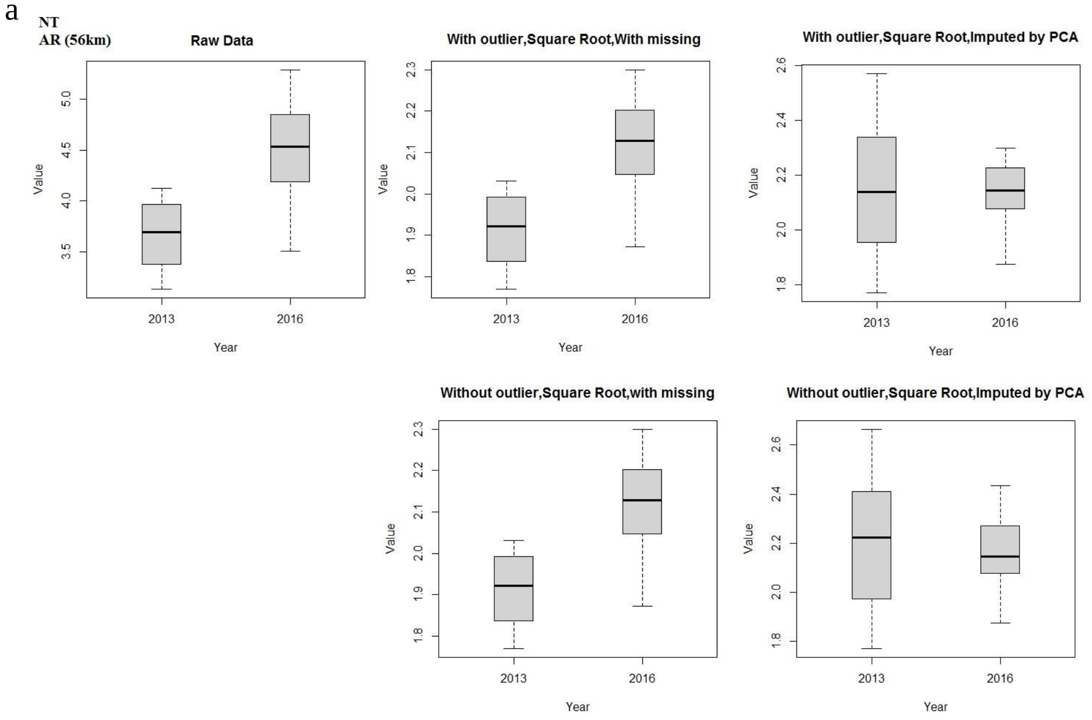
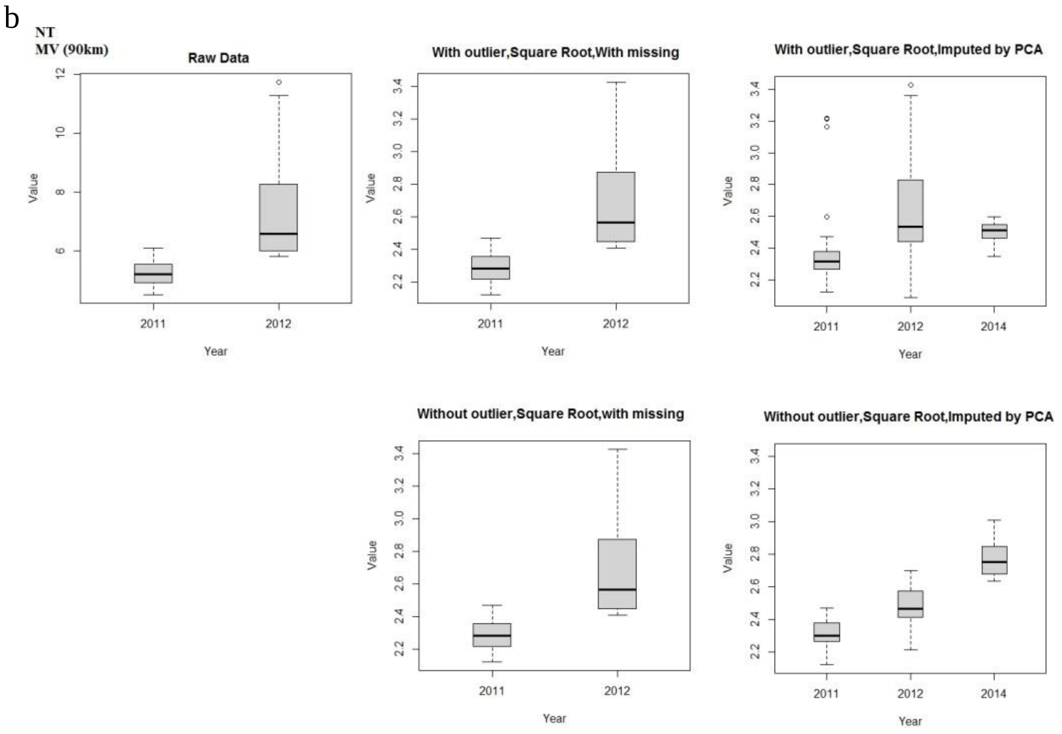


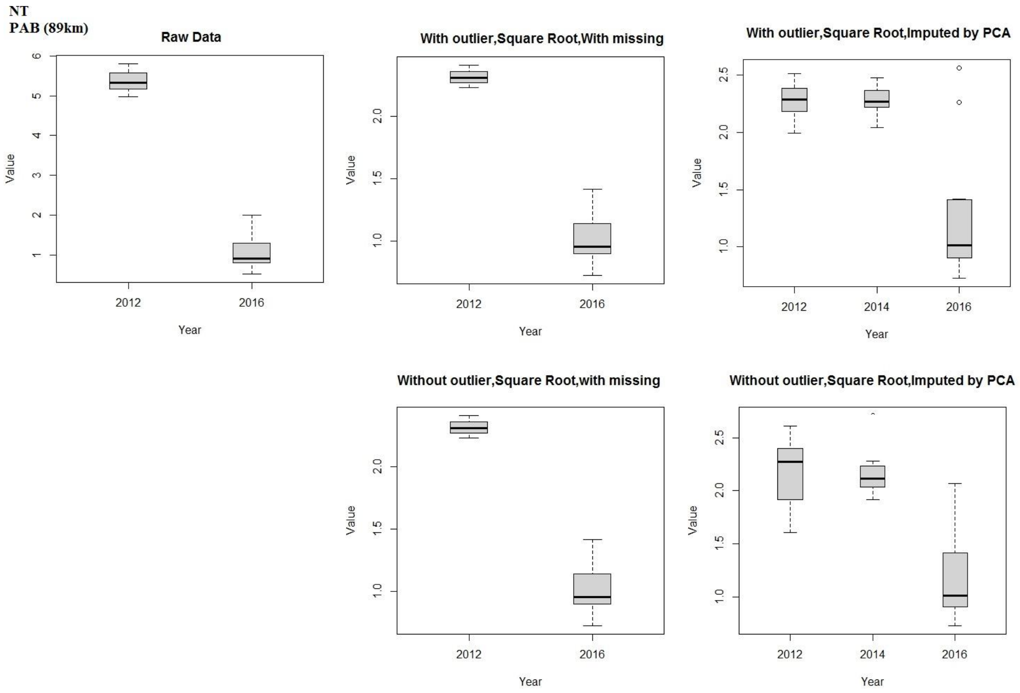
Fig. 29. Distribution of Silicate (44.3 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

Fig. 30 shows the distribution of TN. in different years and sites. The abundance of TN in AR and 2016 is higher than 2013 (Fig. 30a). The abundance of TN in MV and 2012 is higher than 2011 (Fig. 30b). The abundance of TN in PAB and 2016 is higher than 2012 (Fig. 30c). The abundance of TN in PL and 2016 is higher than 2013 and 2012 (Fig. 30d). The abundance of TN in SG and 2016 is higher than 2012 (Fig. 30e). The abundance of TN in TIM and 2013 is higher than 2012 (Fig. 30f).

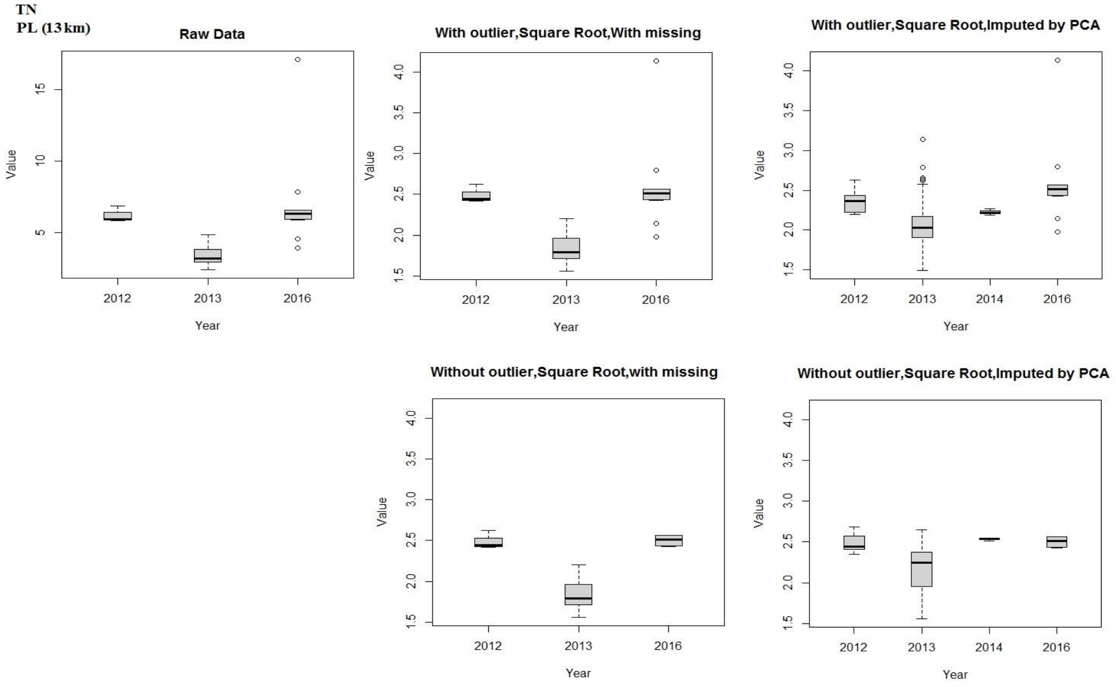




c


NT
PAB (89km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2014
2016
1
2
3
4
5
6
1.0
1.5
2.0
2.5

d


TN
PL (13km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2013
2014
2016
5
10
15
1.5
2.0
2.5
3.0
3.5
4.0

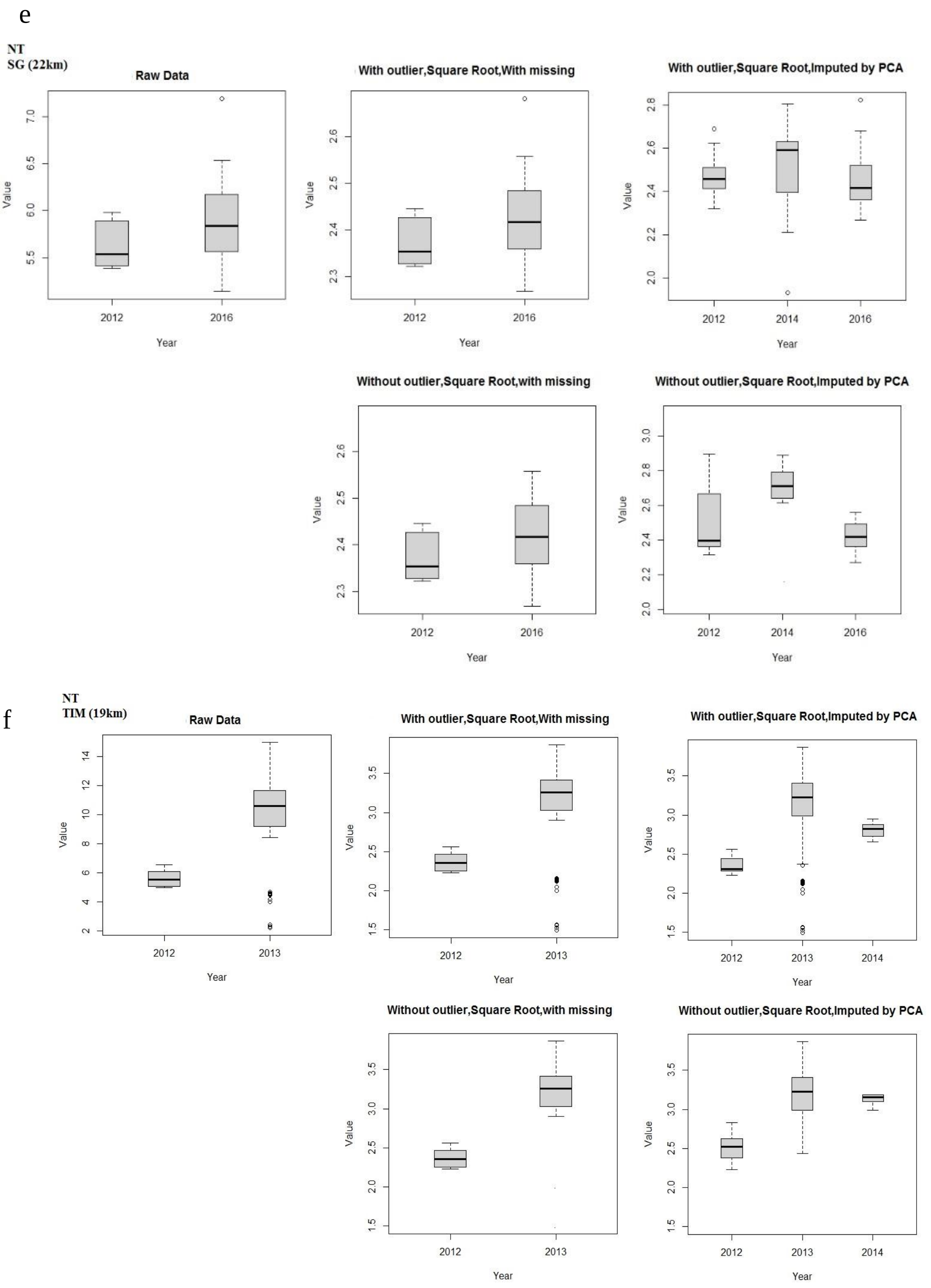
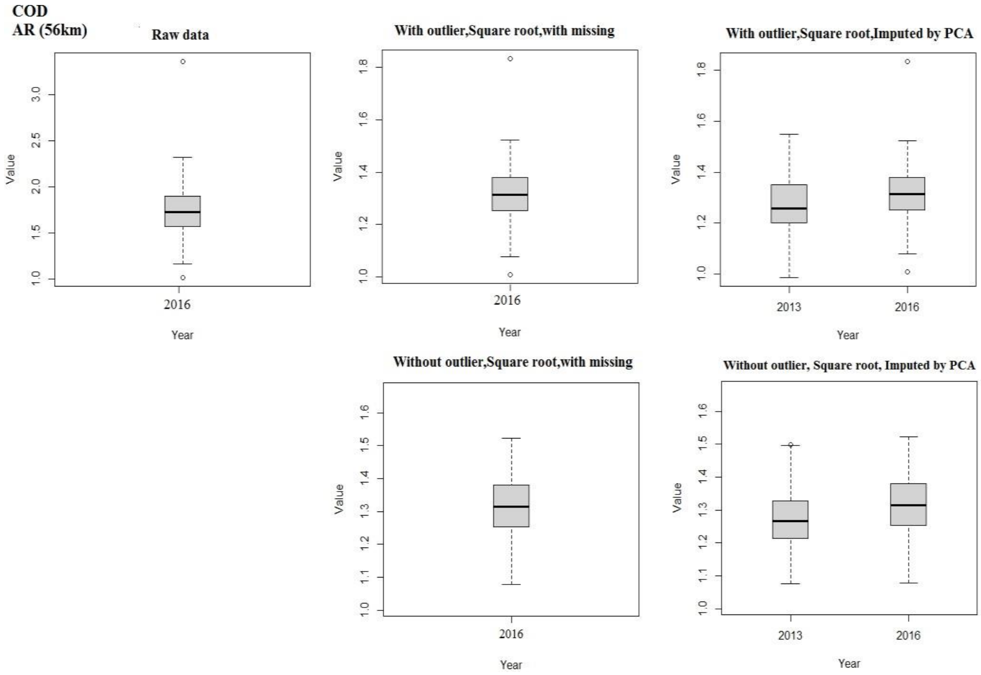


Fig. 30. Distribution of TN. (53.12 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

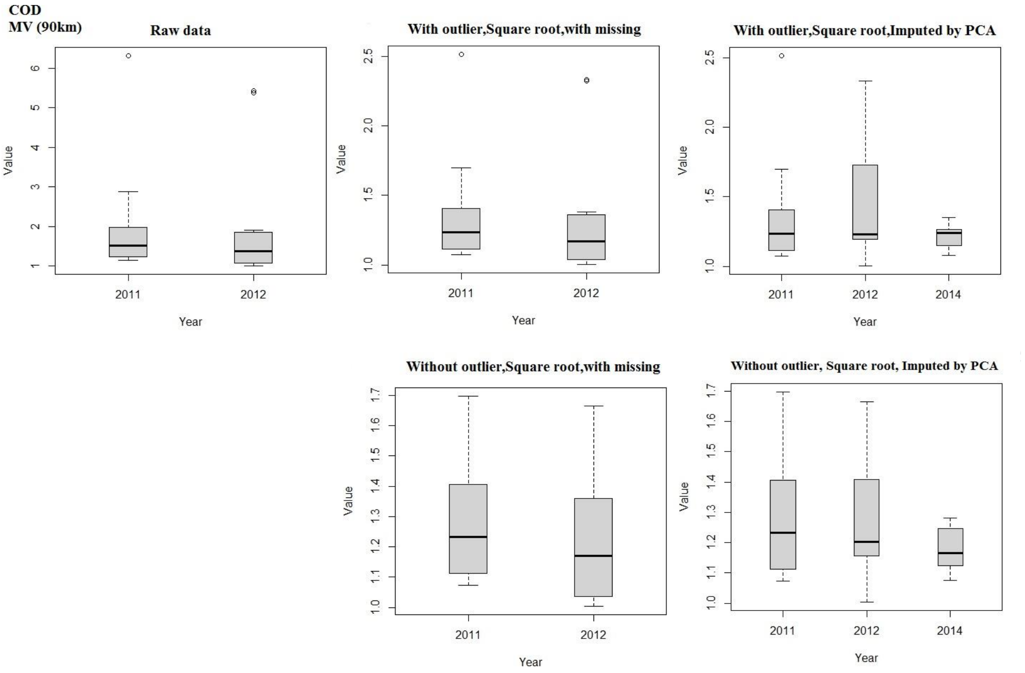
Fig. 31 shows the distribution of DOC. in different years and sites. The abundance of DOC in AR and 2016 (Fig. 31a). The abundance of DOC in MV and 2011 is higher than 2012 (Fig. 31b). The abundance of DOC in PAB and 2016 is higher than 2012 (Fig. 31c).

The abundance of DOC in PL and 2016 is higher than 2012 (Fig. 31d). The abundance of DOC in SG and 2016 is higher than 2012 (Fig. 31e). The abundance of DOC in TIM and 2012 (Fig. 31f).

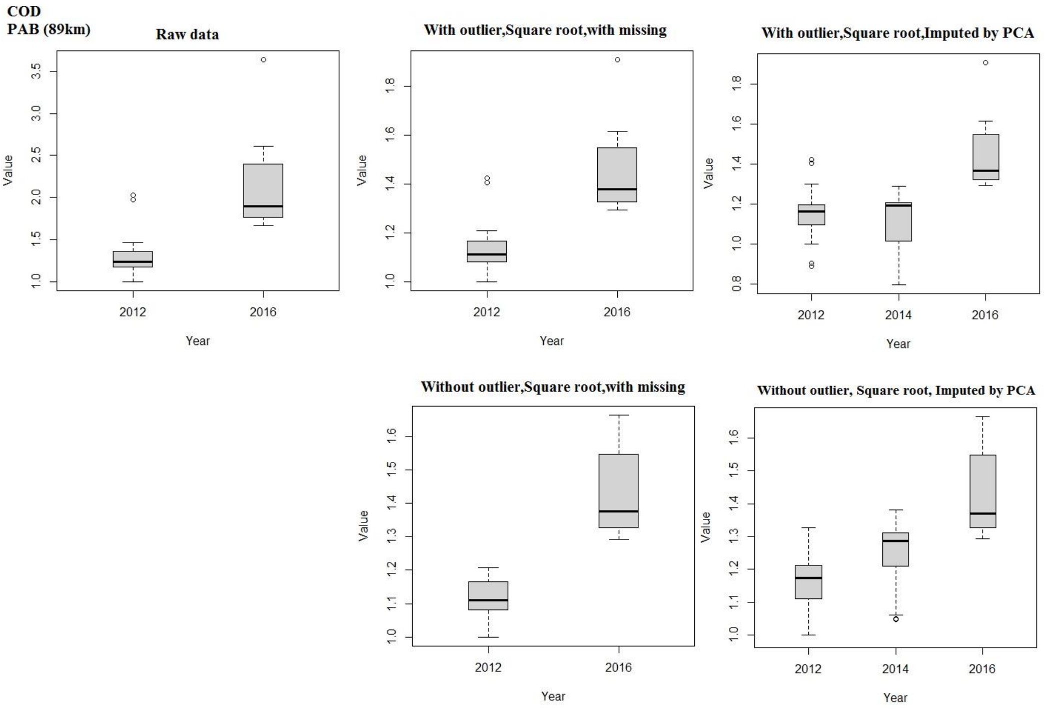
a



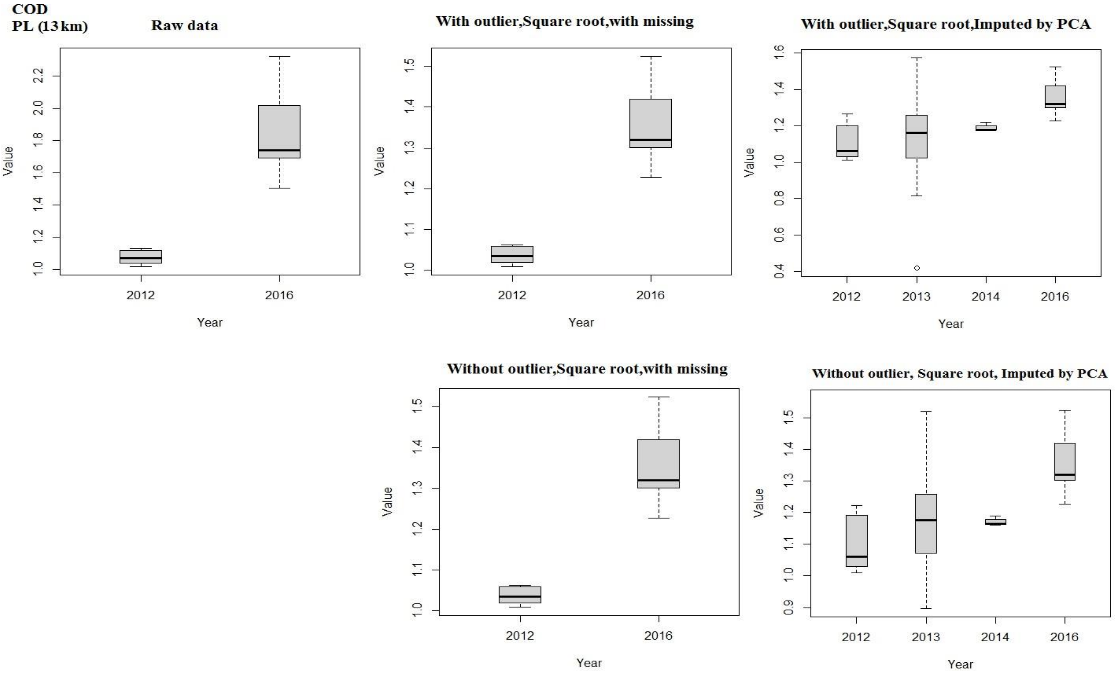
b



c


COD
PAB (89km)
Raw data
With outlier,Square root,with missing
With outlier,Square root,Imputed by PCA
Without outlier,Square root,with missing
Without outlier, Square root, Imputed by PCA
Value
Year
2012
2014
2016
1.0
1.5
2.0
2.5
3.0
3.5
1.2
1.4
1.6
1.8
0.8
1.1
1.3
1.5

d


COD
PL (13km)
Raw data
With outlier,Square root,with missing
With outlier,Square root,Imputed by PCA
Without outlier,Square root,with missing
Without outlier, Square root, Imputed by PCA
Value
Year
2012
2013
2014
2016
1.0
1.2
1.4
1.6
1.8
2.0
2.2
1.1
1.3
1.5
0.4
0.6
0.8
0.9

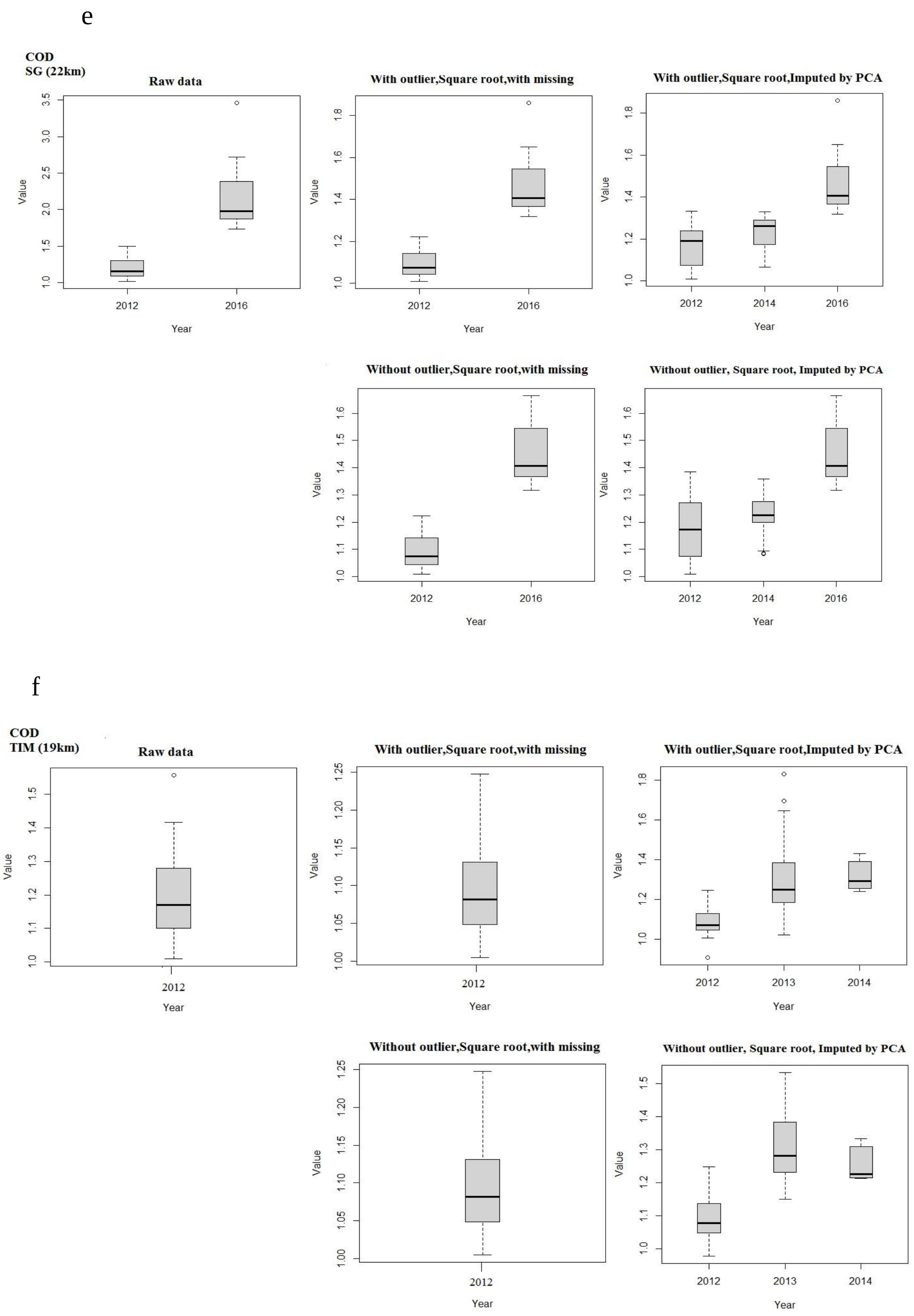
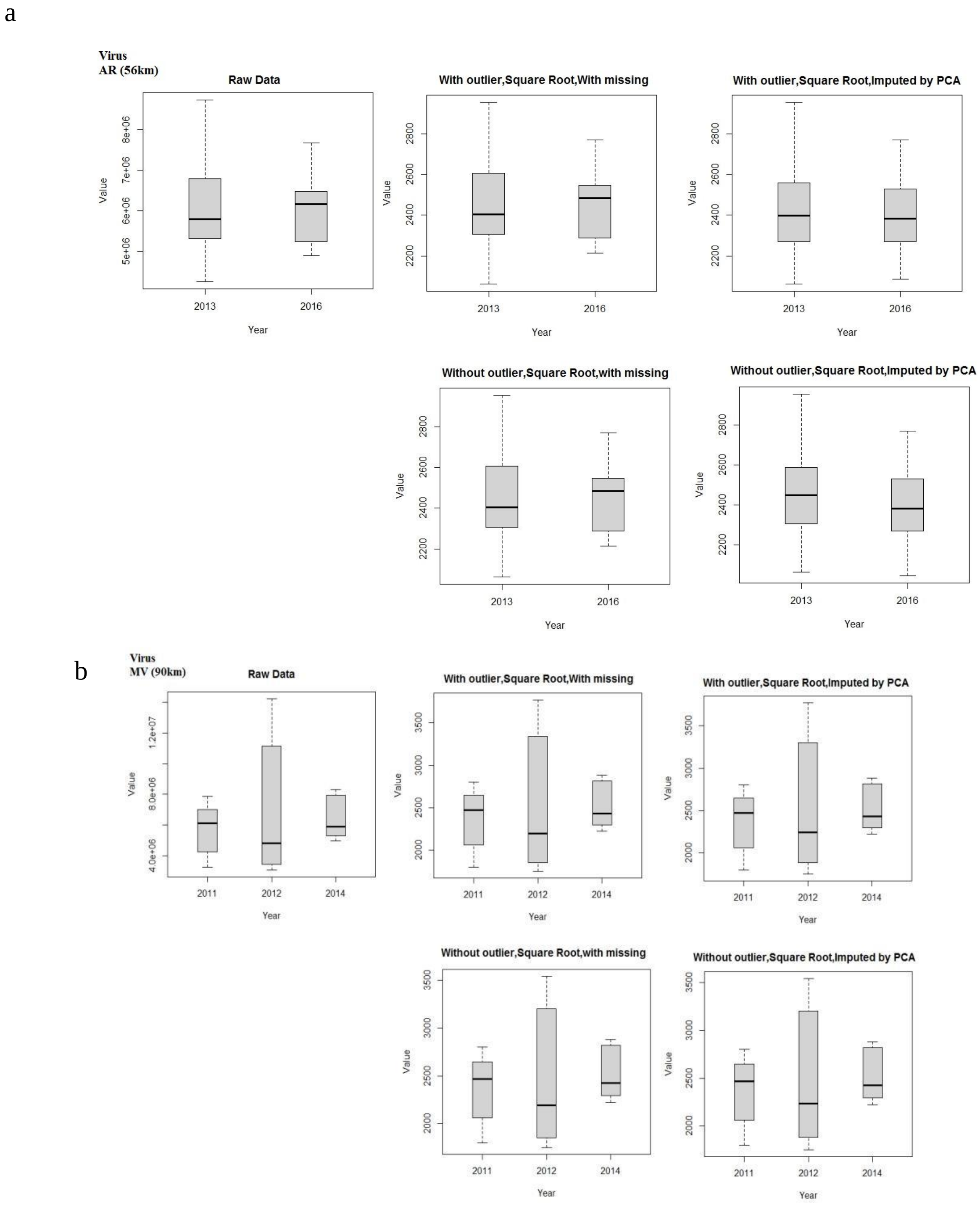


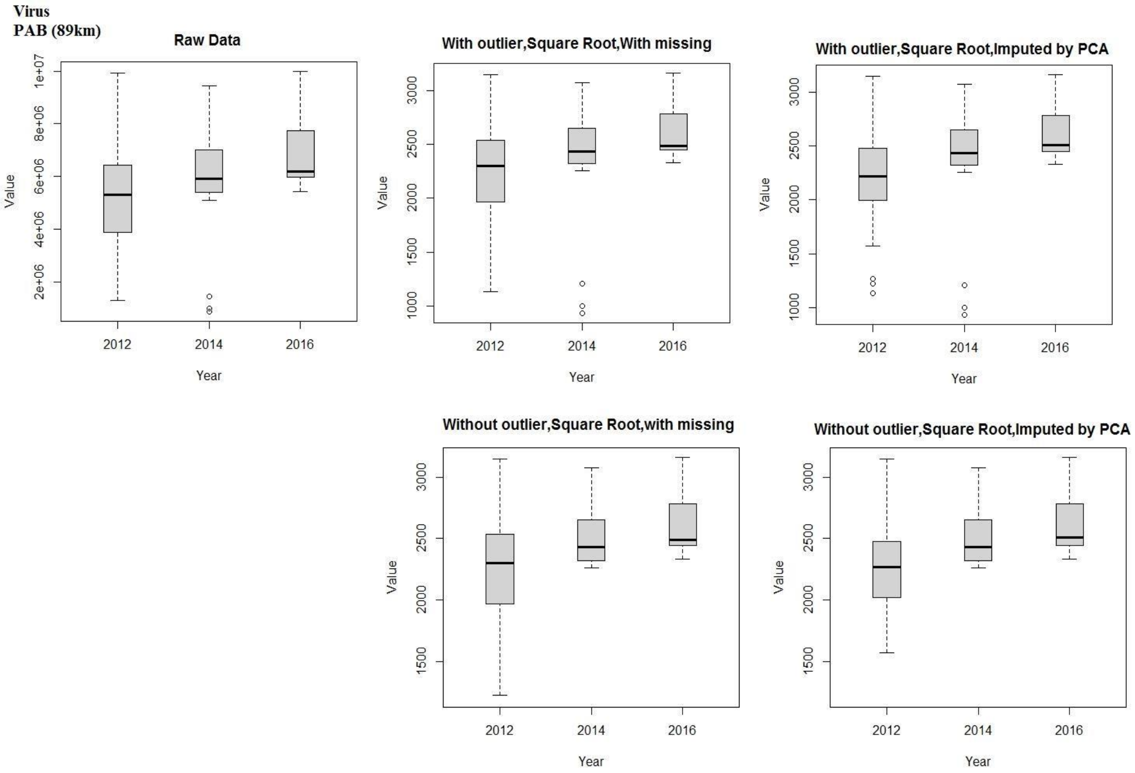
Fig. 31. Distribution of DOC. (75.18 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

Fig. 32 shows the distribution of Virus in different years and sites. The abundance of Virus in AR and 2013 is higher than 2016 (Fig. 32a). The abundance of Virus in MV and

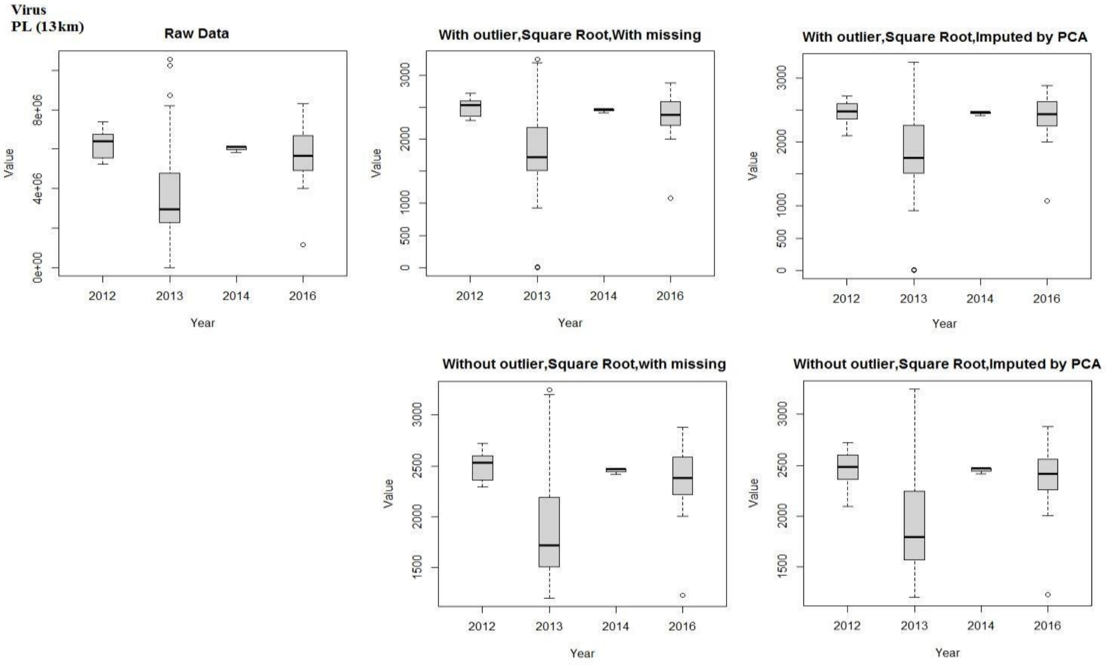
2012 is higher than 2011 and 2014 (Fig. 32b). The abundance of Virus in PAB and 2012 is higher than 2014 and 2016 (Fig. 32c). The abundance of Virus in PL and 2013 is higher than 2012 and 2014 and 2016 (Fig. 32d). The abundance of Virus in SG and 2012 is higher than 2014 and 2016 (Fig. 32e). The abundance of Virus in TIM and 2012 is higher than 2013 and 2014 (Fig. 32f).



c


Virus
PAB (89km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2014
2016
2e+06
4e+06
6e+06
8e+06
1e+07
1000
1500
2000
2500
3000

d


Virus
PL (13km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2013
2014
2016
0e+00
4e+06
8e+06
0
500
1000
1500
2000
2500
3000

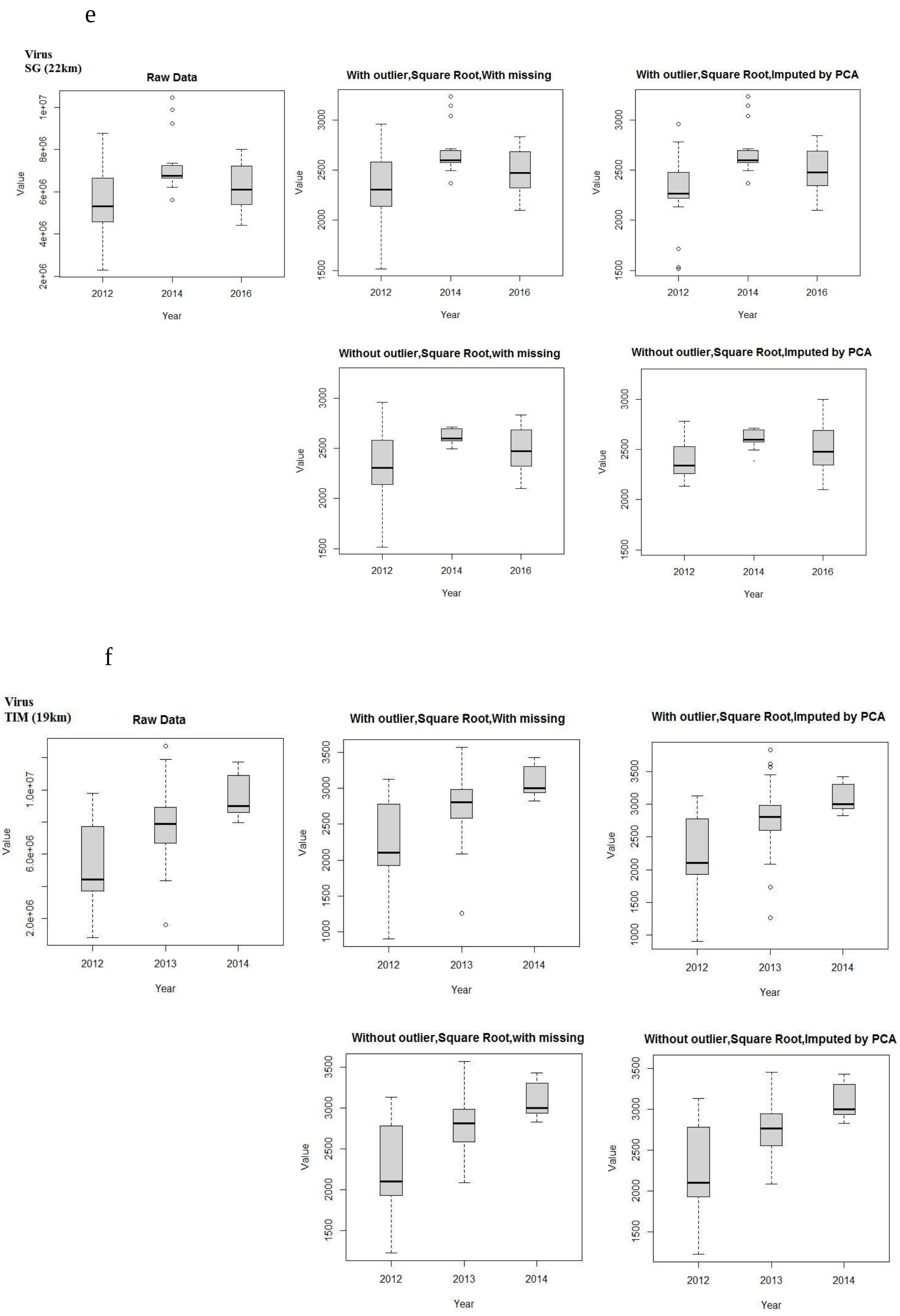
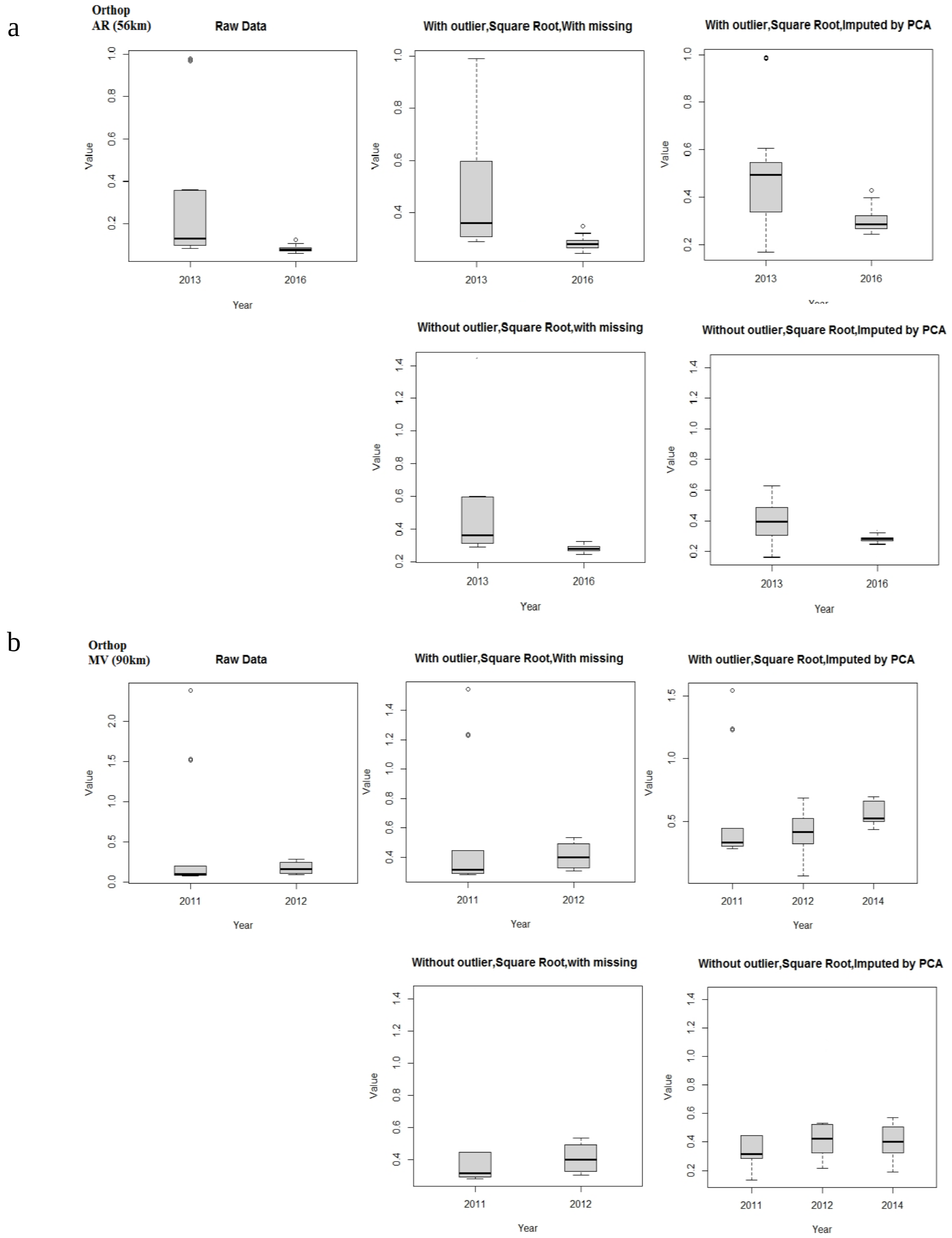
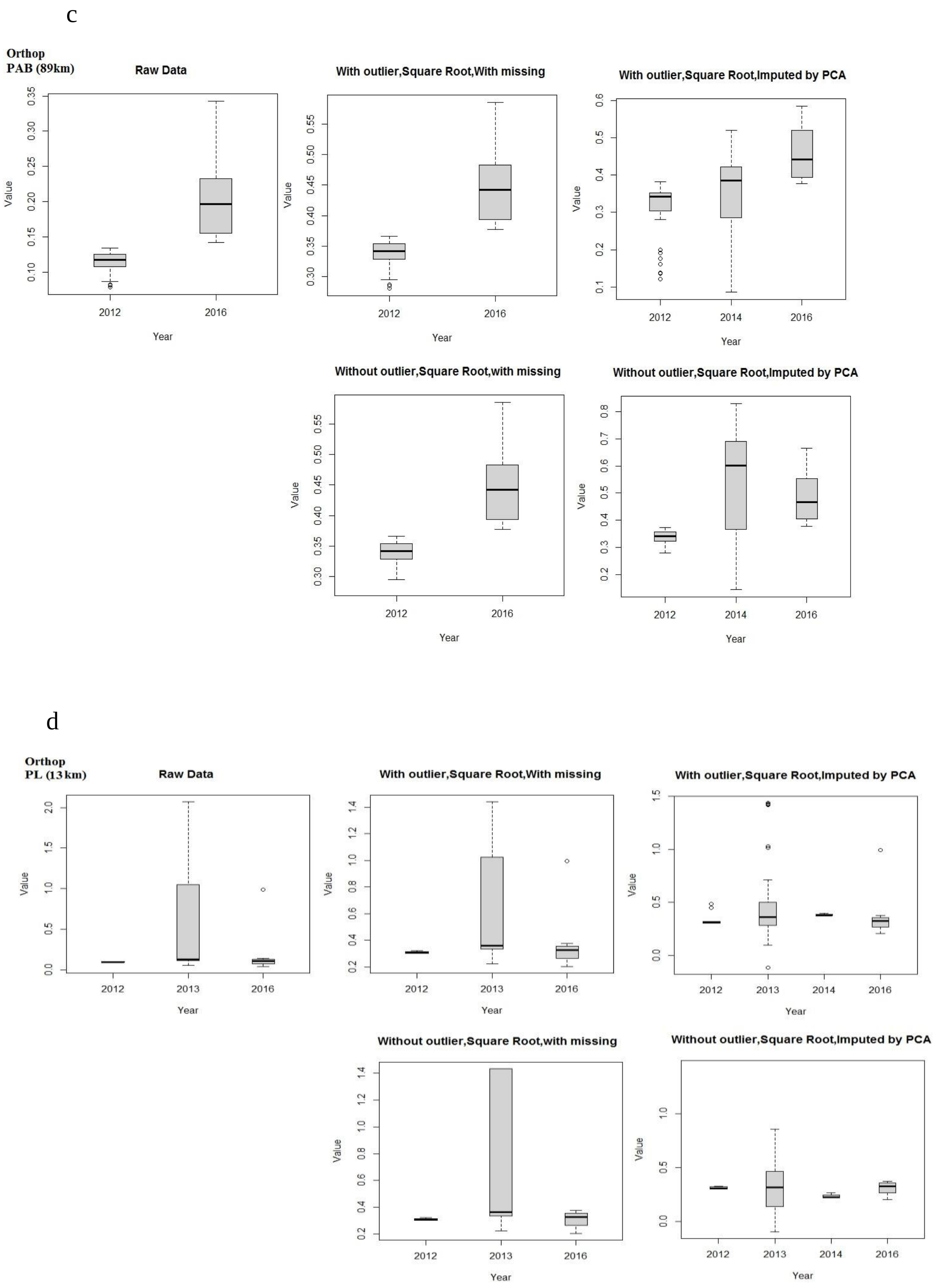


Fig. 32 Distribution of Virus. (16.18 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

Fig. 33 shows the distribution of Orthop in different years and sites. The abundance of Orthop in AR and 2013 is higher than 2016 (Fig. 33a). The abundance of Orthop in MV and 2012 is higher than 2011 and 2014 (Fig. 33b). The abundance of Orthop in PAB and 2016 is higher than 2012 (Fig. 33c). The abundance of Orthop in PL and 2013 is higher than 2016 and 2012 (Fig. 33d). The abundance of Orthop in SG and 2012 is higher than 2016 (Fig. 33e). The abundance of Orthop in TIM and 2013 is higher than 2012 (Fig. 33f).




c
Orthop
PAB (89km)
Raw Data
Value
0.10
0.15
0.20
0.25
0.30
0.35
2012
2016
Year
With outlier,Square Root,With missing
Value
0.30
0.35
0.40
0.45
0.50
0.55
2012
2016
Year
With outlier,Square Root,Imputed by PCA
Value
0.1
0.2
0.3
0.4
0.5
0.6
2012
2014
2016
Year
Without outlier,Square Root,with missing
Value
0.30
0.35
0.40
0.45
0.50
0.55
2012
2016
Year
Without outlier,Square Root,Imputed by PCA
Value
0.2
0.3
0.4
0.5
0.6
0.7
0.8
2012
2014
2016
Year
d
Orthop
PL (13km)
Raw Data
Value
0.0
0.5
1.0
1.5
2.0
2012
2013
2016
Year
With outlier,Square Root,With missing
Value
0.2
0.4
0.6
0.8
1.0
1.2
1.4
2012
2013
2016
Year
With outlier,Square Root,Imputed by PCA
Value
0.0
0.5
1.0
1.5
2012
2013
2014
2016
Year
Without outlier,Square Root,with missing
Value
0.2
0.4
0.6
0.8
1.0
1.2
1.4
2012
2013
2016
Year
Without outlier,Square Root,Imputed by PCA
Value
0.0
0.5
1.0
2012
2013
2014
2016
Year

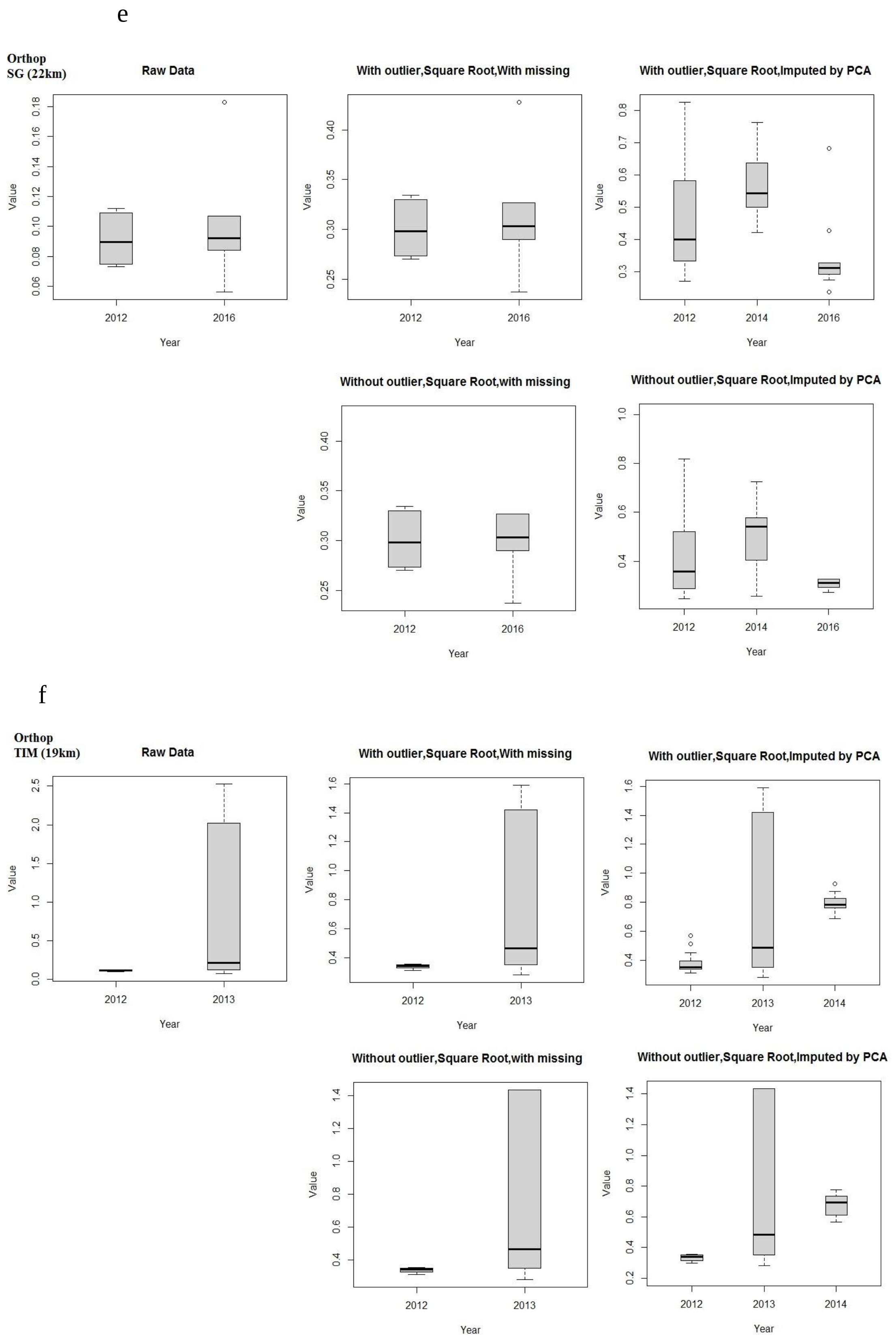
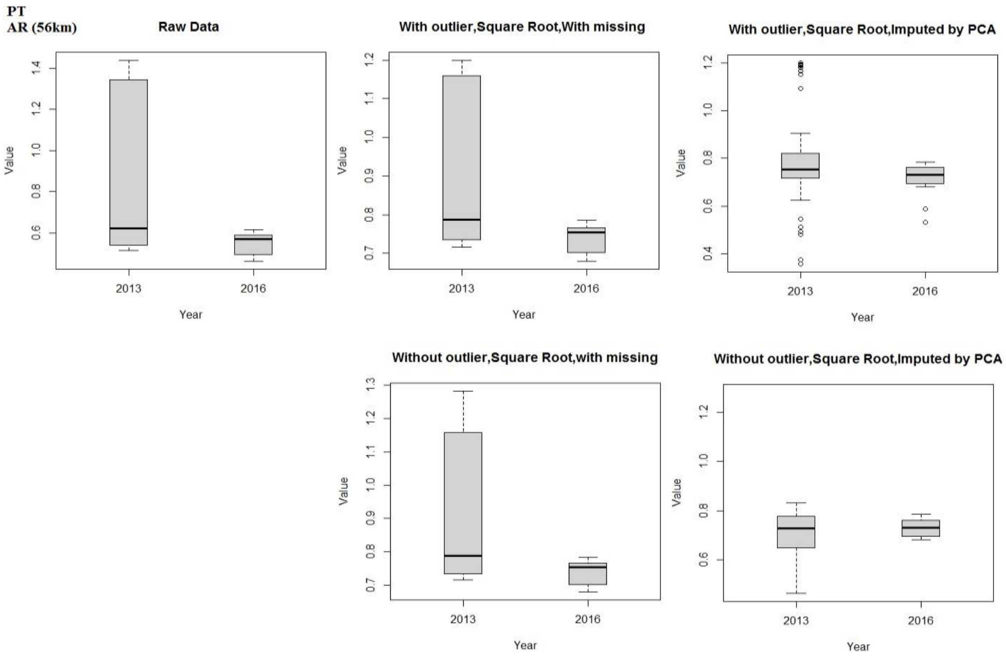


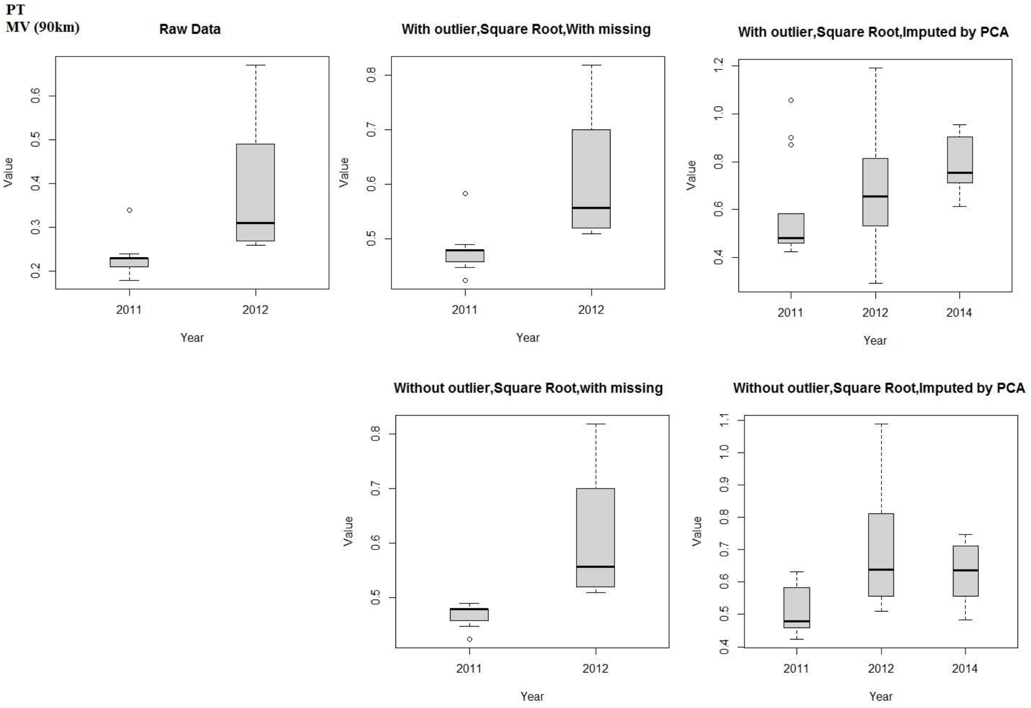
Fig. 33 Distribution of Orthop. (45.96 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

Fig. 34 shows the distribution of TP in different years and sites. The abundance of TP in AR and 2013 is higher than 2016 (Fig. 34a). The abundance of TP in MV and 2012 is higher than 2011 (Fig. 34b). The abundance of TP in PAB and 2016 is higher than 2012 (Fig. 34c). The abundance of TP in PL and 2013 is higher than 2016 and 2012 (Fig. 34d). The abundance of TP in SG and 2012 is higher than 2016 (Fig. 34e). The abundance of TP in TIM and 2013 is higher than 2012 (Fig. 34f).

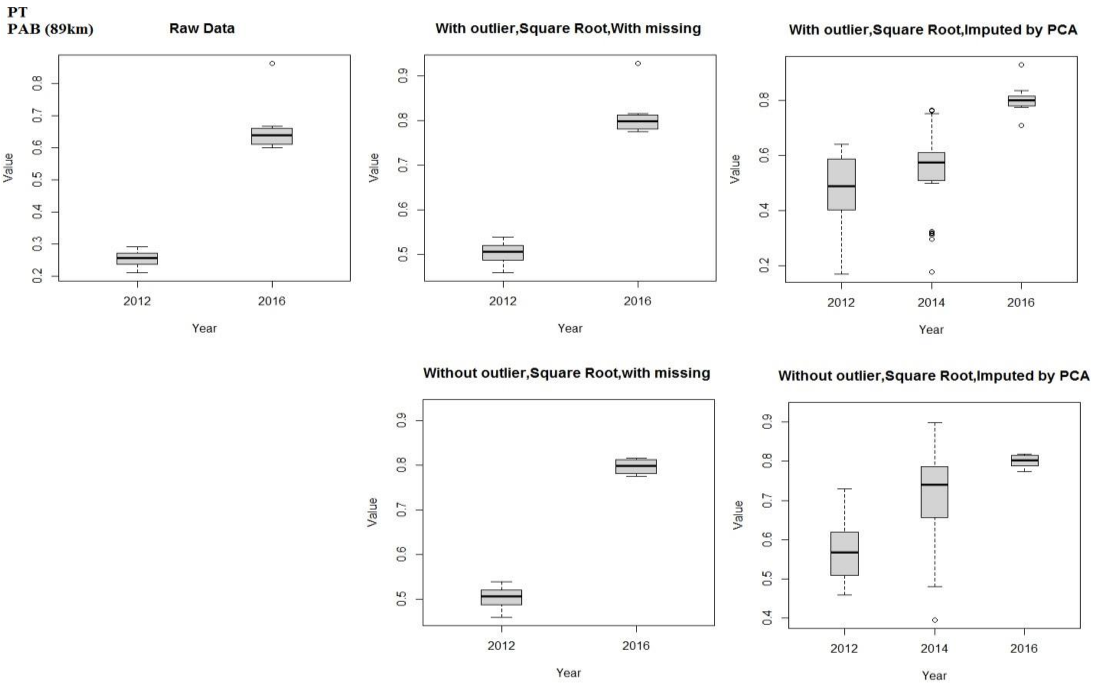
a



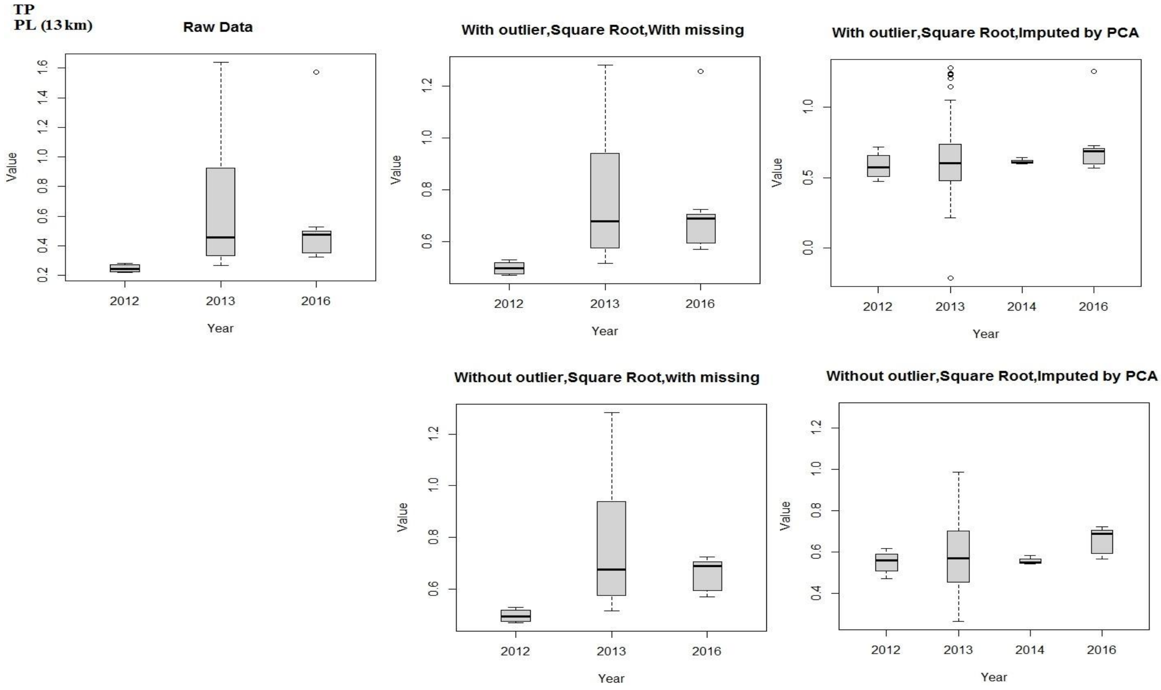
b



c


PT
PAB (89km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2014
2016

d


TP
PL (13 km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2013
2014
2016

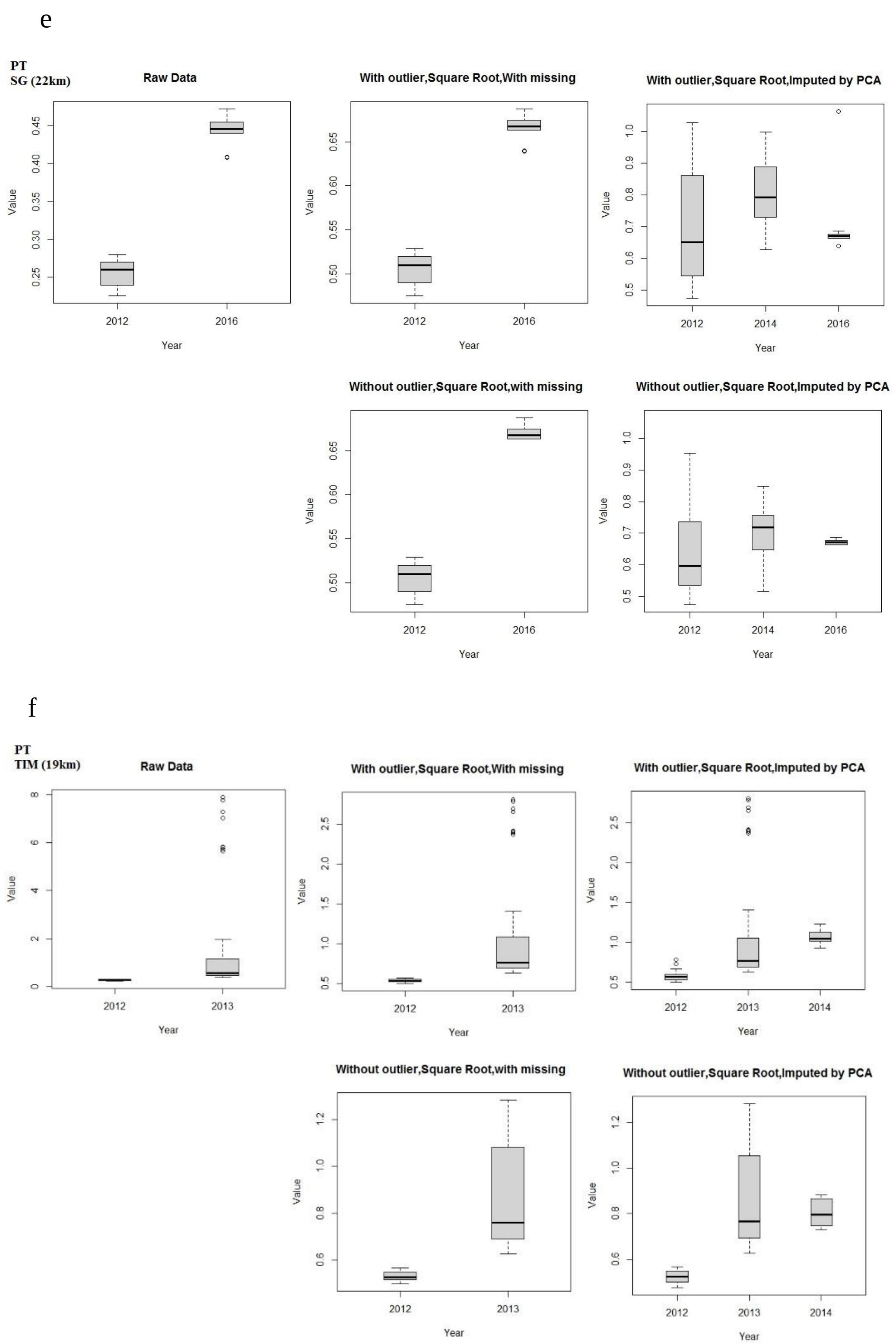
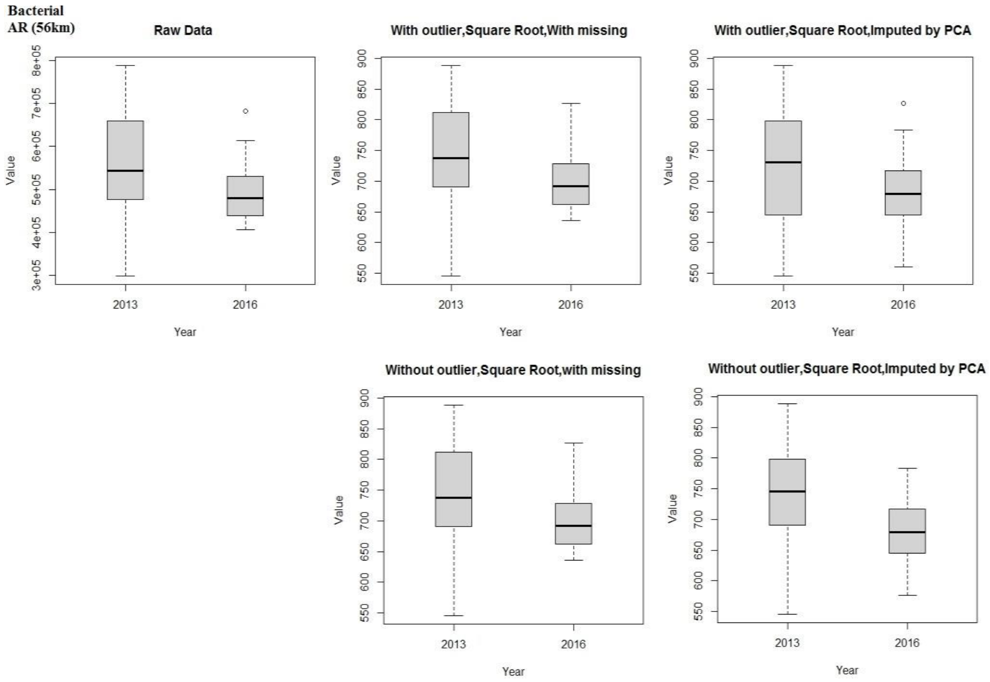


Fig. 34 Distribution of Orthop. (51.33 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

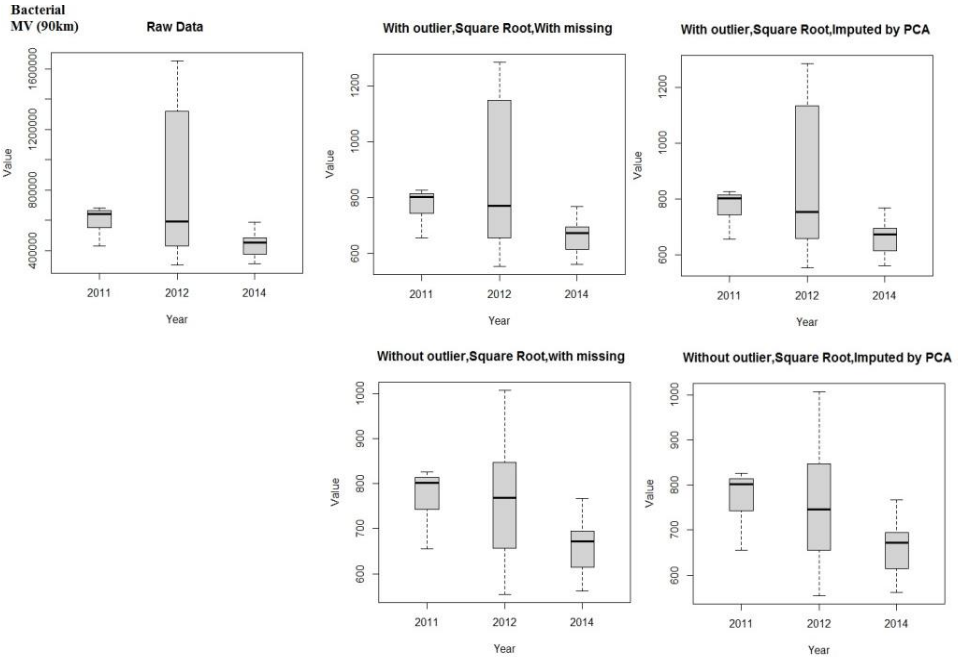
Fig. 35 shows the distribution of Bacterial in different years and sites. The abundance of Bacterial in AR and 2013 is higher than 2016 (Fig. 35a). The abundance of Bacterial in

MV and 2012 is higher than 2011 and 2014 (Fig. 35b). The abundance of Bacterial in PAB and 2014 is higher than 2012 and 2016 (Fig. 35c). The abundance of Bacterial in PL and 2013 is higher than 2012 and 2014 and 2016 (Fig. 35d). The abundance of Bacterial in SG and 2014 is higher than 2012 and 2016 (Fig. 35e). The abundance of Bacterial in TIM and 2013 is higher than 2012 and 2014 (Fig. 35f).

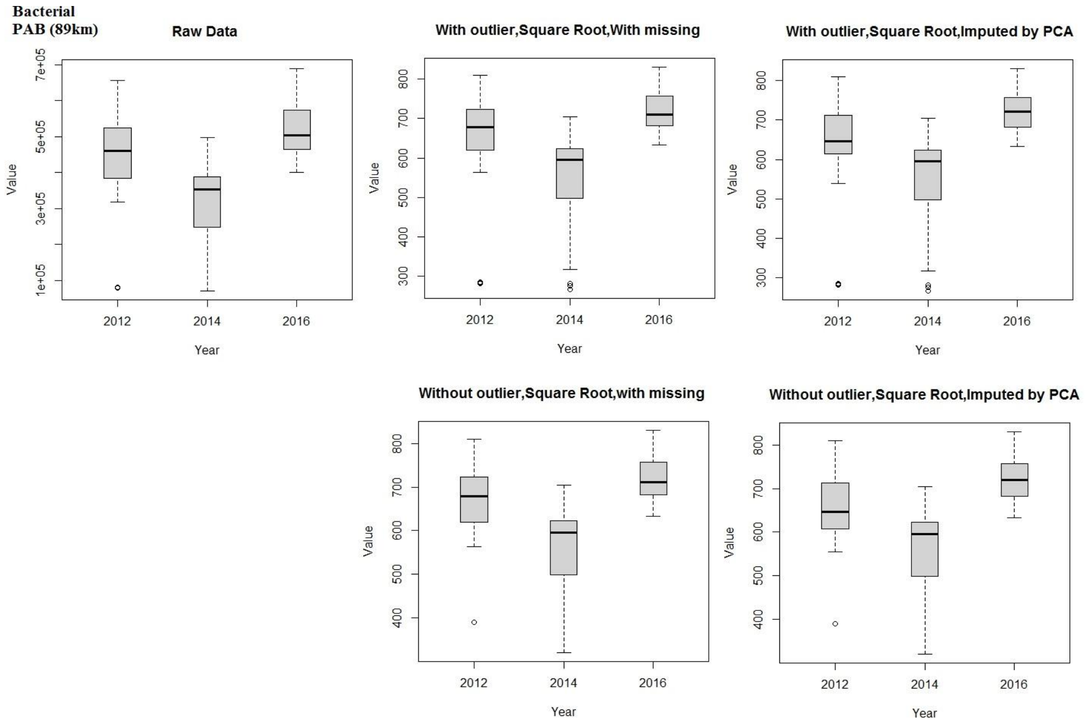
a



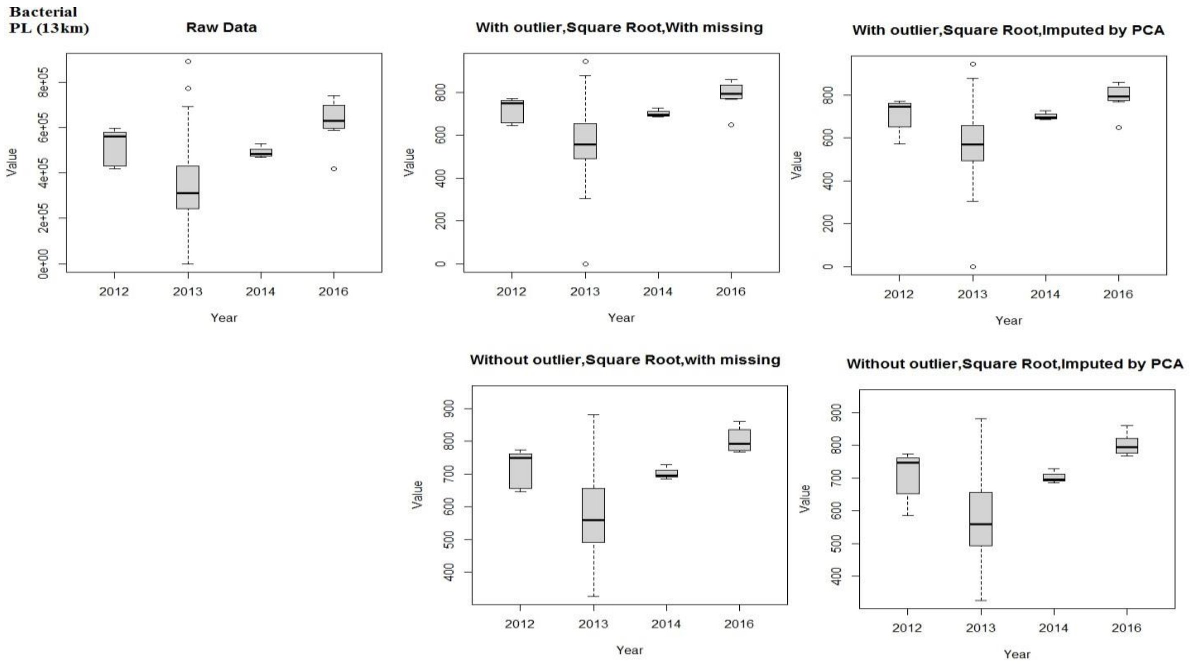
b



c


Bacterial
PAB (89km)
Raw Data
Value
1e+05
3e+05
5e+05
7e+05
2012
2014
2016
Year
With outlier,Square Root,With missing
Value
300
400
500
600
700
800
2012
2014
2016
Year
With outlier,Square Root,Imputed by PCA
Value
300
400
500
600
700
800
2012
2014
2016
Year
Without outlier,Square Root,with missing
Value
400
500
600
700
800
2012
2014
2016
Year
Without outlier,Square Root,Imputed by PCA
Value
400
500
600
700
800
2012
2014
2016
Year

d


Bacterial
PL (13km)
Raw Data
Value
0e+00
2e+05
4e+05
6e+05
8e+05
2012
2013
2014
2016
Year
With outlier,Square Root,With missing
Value
0
200
400
600
800
2012
2013
2014
2016
Year
With outlier,Square Root,Imputed by PCA
Value
0
200
400
600
800
2012
2013
2014
2016
Year
Without outlier,Square Root,with missing
Value
400
500
600
700
800
900
2012
2013
2014
2016
Year
Without outlier,Square Root,Imputed by PCA
Value
400
500
600
700
800
900
2012
2013
2014
2016
Year

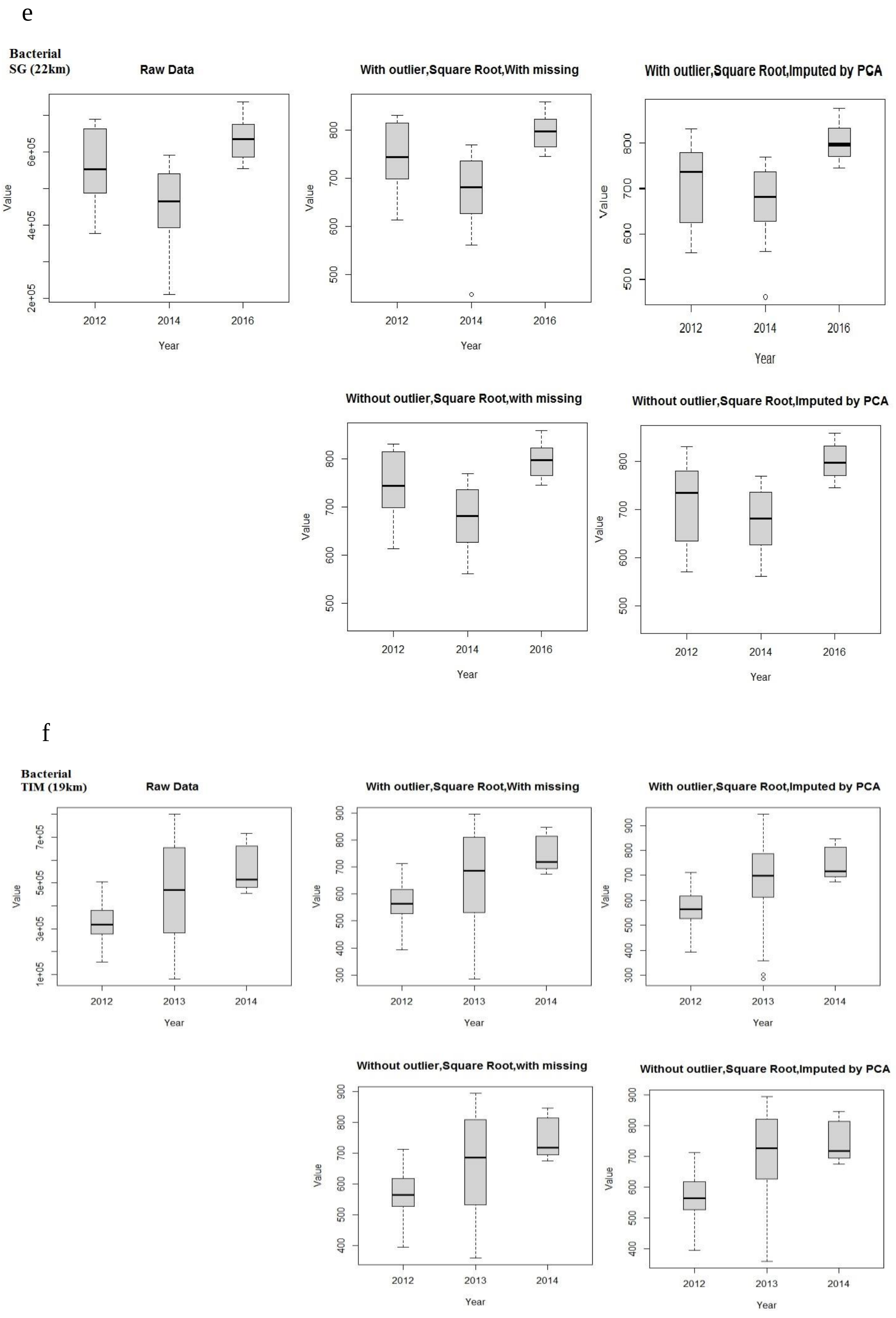
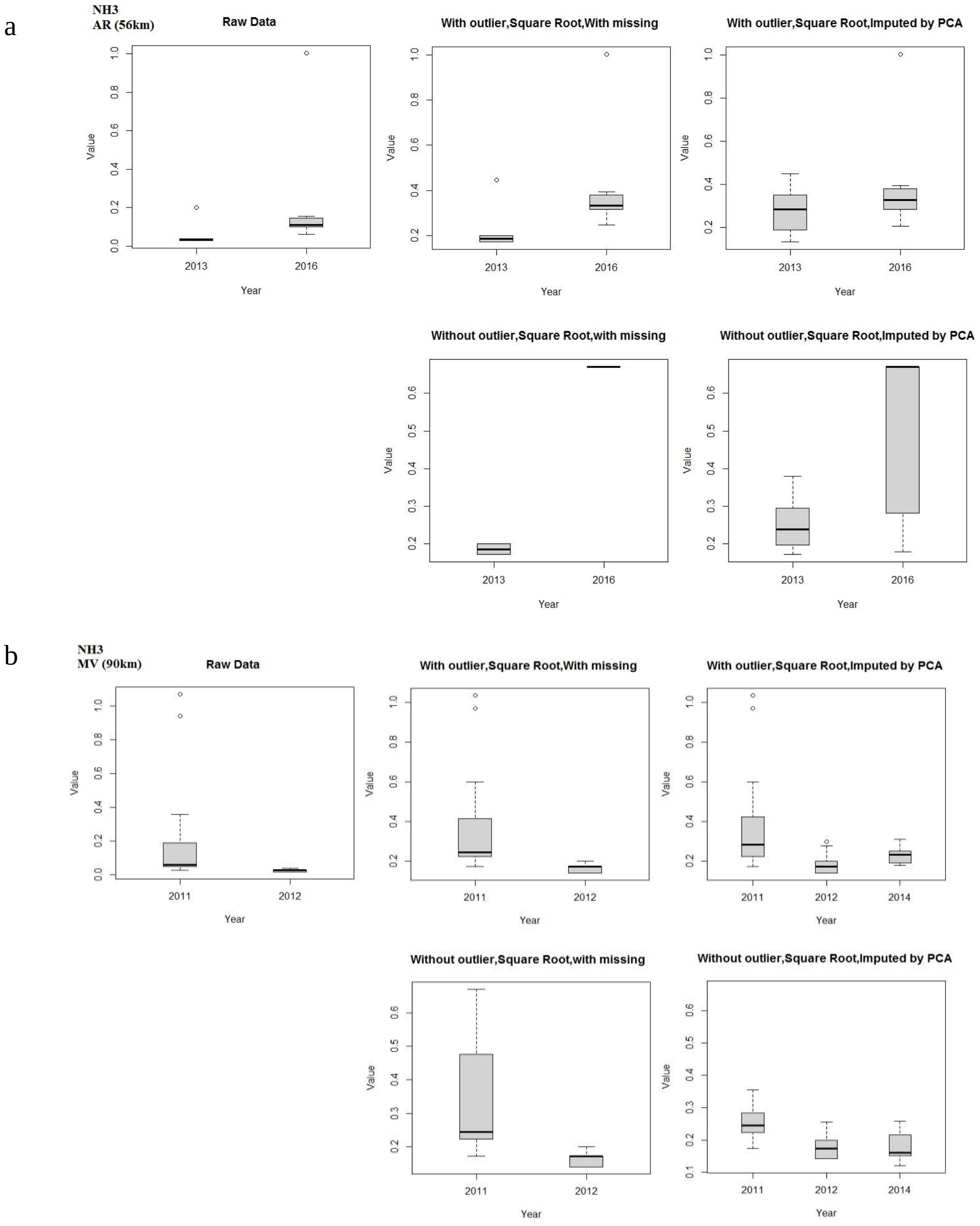
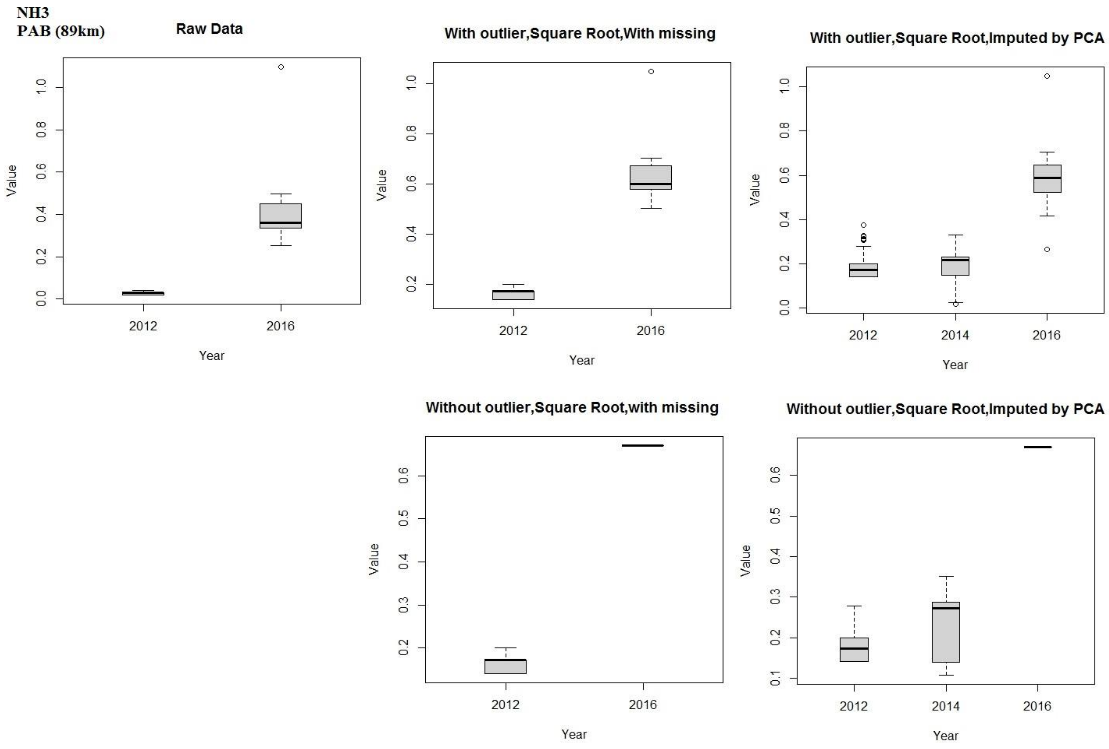


Fig. 35 Distribution of Bacterial (15.44 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

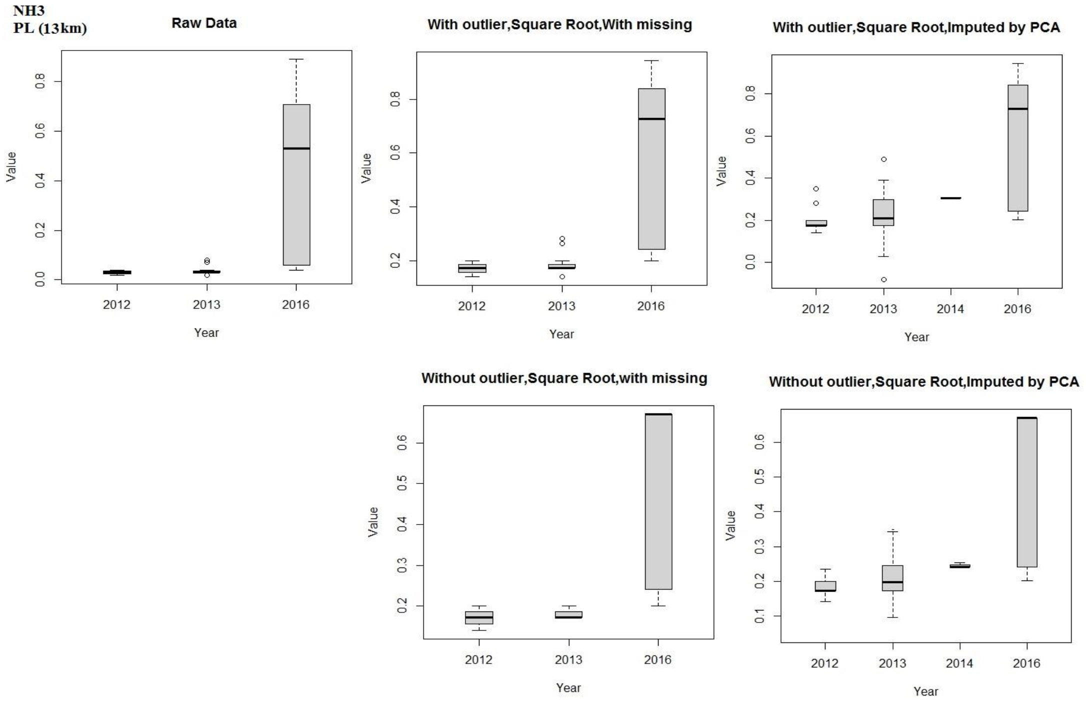
Fig. 36 shows the distribution of NH3 in different years and sites. The abundance of NH3 in AR and 2016 is higher than 2013 (Fig. 36a). The abundance of NH3 in MV and 2011 is higher than 2012 (Fig. 36b). The abundance of NH3 in PAB and 2016 is higher than 2012 (Fig. 36c). The abundance of NH3 in PL and 2016 is higher than 2012 and 2013 (Fig.36d). The abundance of NH3 in SG and 2016 is higher than 2012 (Fig. 36e). The abundance of NH3 in TIM and 2013 is higher than 2012 (Fig. 36f).



c


NH3
PAB (89km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2014
2016
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.8
1.0

d


NH3
PL (13km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2013
2014
2016
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.8

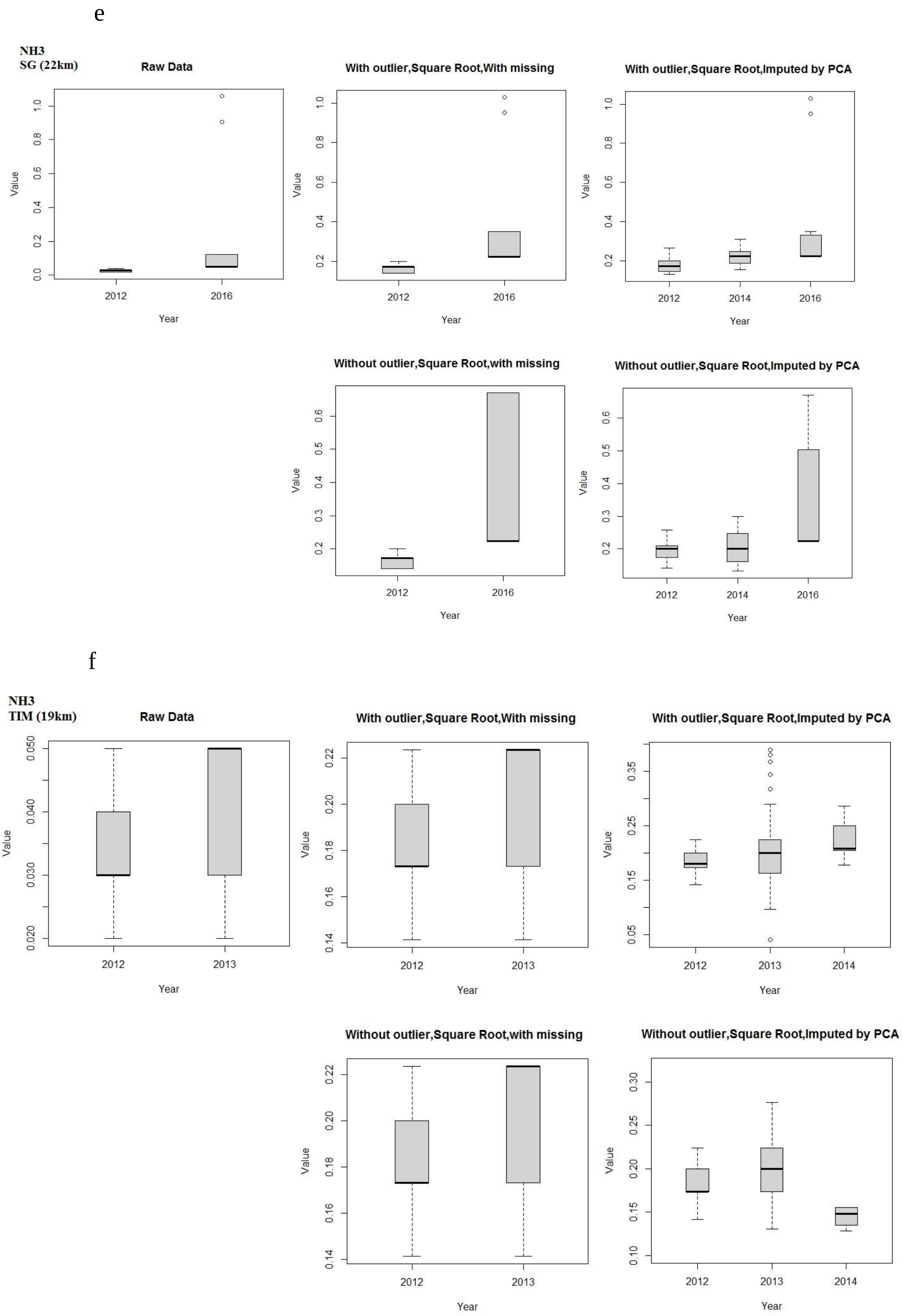
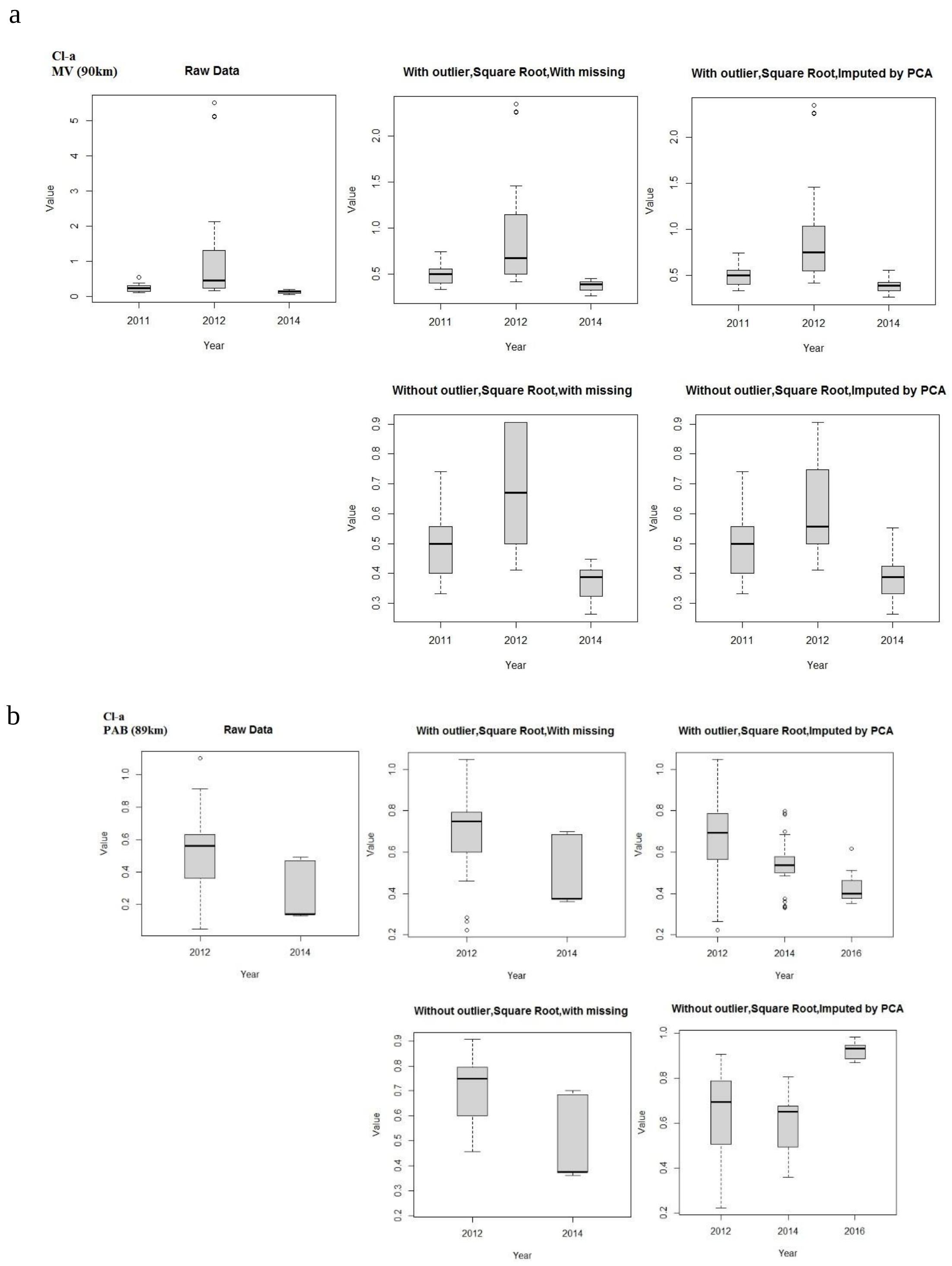
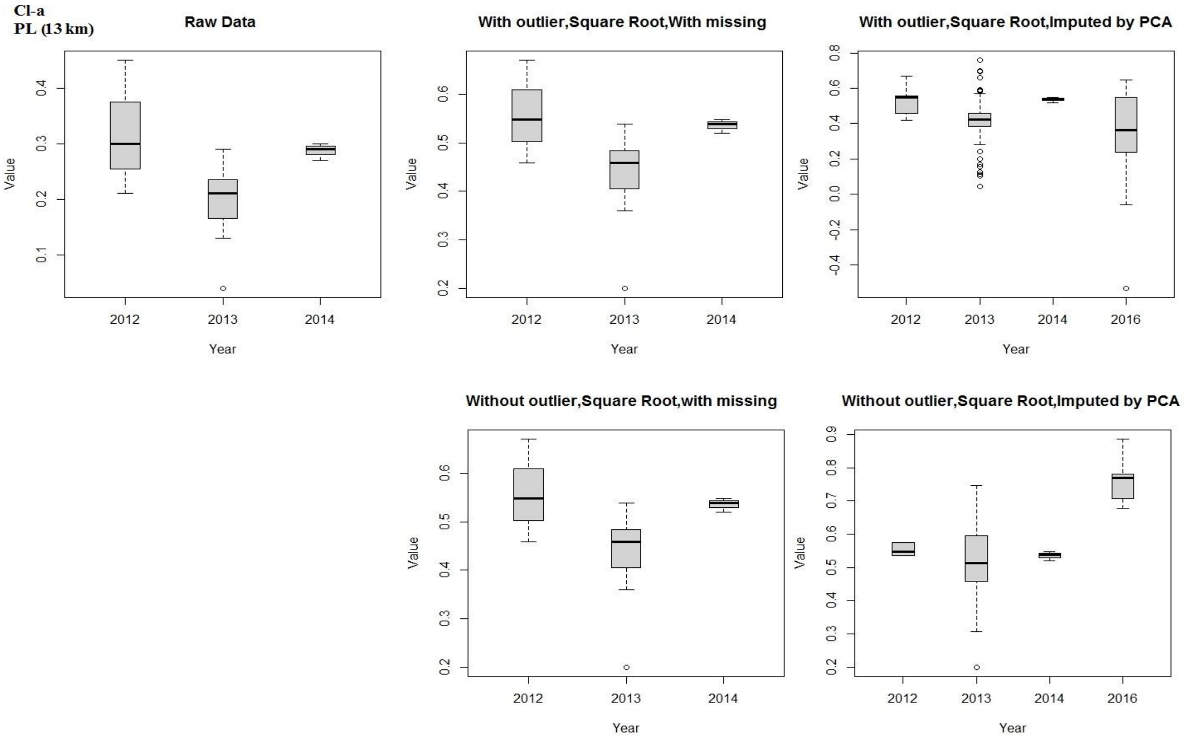


Fig. 36 Distribution of NH3 (48.16 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

Fig. 37 shows the distribution of Cl-a in different years and sites. The abundance of Cl-a in MV and 2012 is higher than 2011 and 2014 (Fig. 37a). The abundance of Cl-a in PAB and 2012 is higher than 2014 (Fig. 37b). The abundance of Cl-a in PL and 2012 is higher than 2013 and 2014 (Fig. 37c). The abundance of Cl-a in SG and 2012 is higher than 2014 (Fig. 37d). The abundance of Cl-a in TIM and 2012 is higher than 2014 (Fig. 37e).

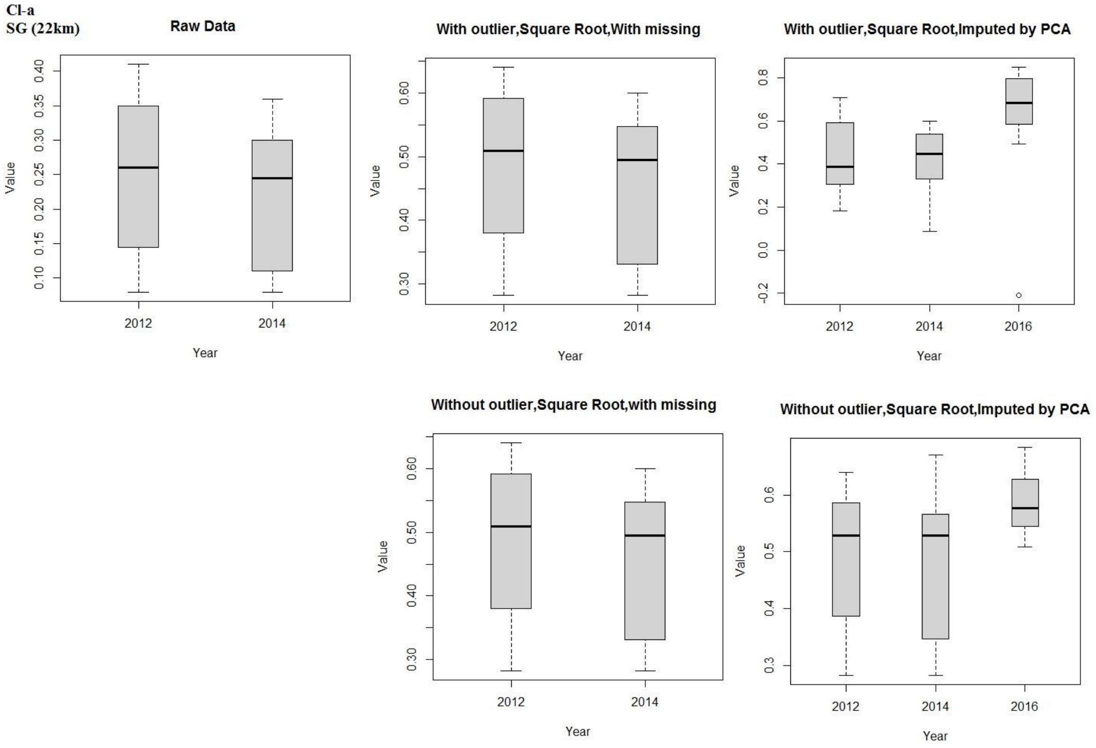


c



Cl-a
PL (13 km)
Raw Data
Value
0.1
0.2
0.3
0.4
2012
2013
2014
Year
With outlier,Square Root,With missing
Value
0.2
0.3
0.4
0.5
0.6
2012
2013
2014
Year
With outlier,Square Root,Imputed by PCA
Value
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
2012
2013
2014
2016
Year
Without outlier,Square Root,with missing
Value
0.2
0.3
0.4
0.5
0.6
2012
2013
2014
Year
Without outlier,Square Root,Imputed by PCA
Value
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
2012
2013
2014
2016
Year

d



Cl-a
SG (22km)
Raw Data
Value
0.10
0.15
0.20
0.25
0.30
0.35
0.40
2012
2014
Year
With outlier,Square Root,With missing
Value
0.30
0.40
0.50
0.60
2012
2014
Year
With outlier,Square Root,Imputed by PCA
Value
-0.2
0.0
0.2
0.4
0.6
0.8
2012
2014
2016
Year
Without outlier,Square Root,with missing
Value
0.30
0.40
0.50
0.60
2012
2014
Year
Without outlier,Square Root,Imputed by PCA
Value
0.3
0.4
0.5
0.6
2012
2014
2016
Year

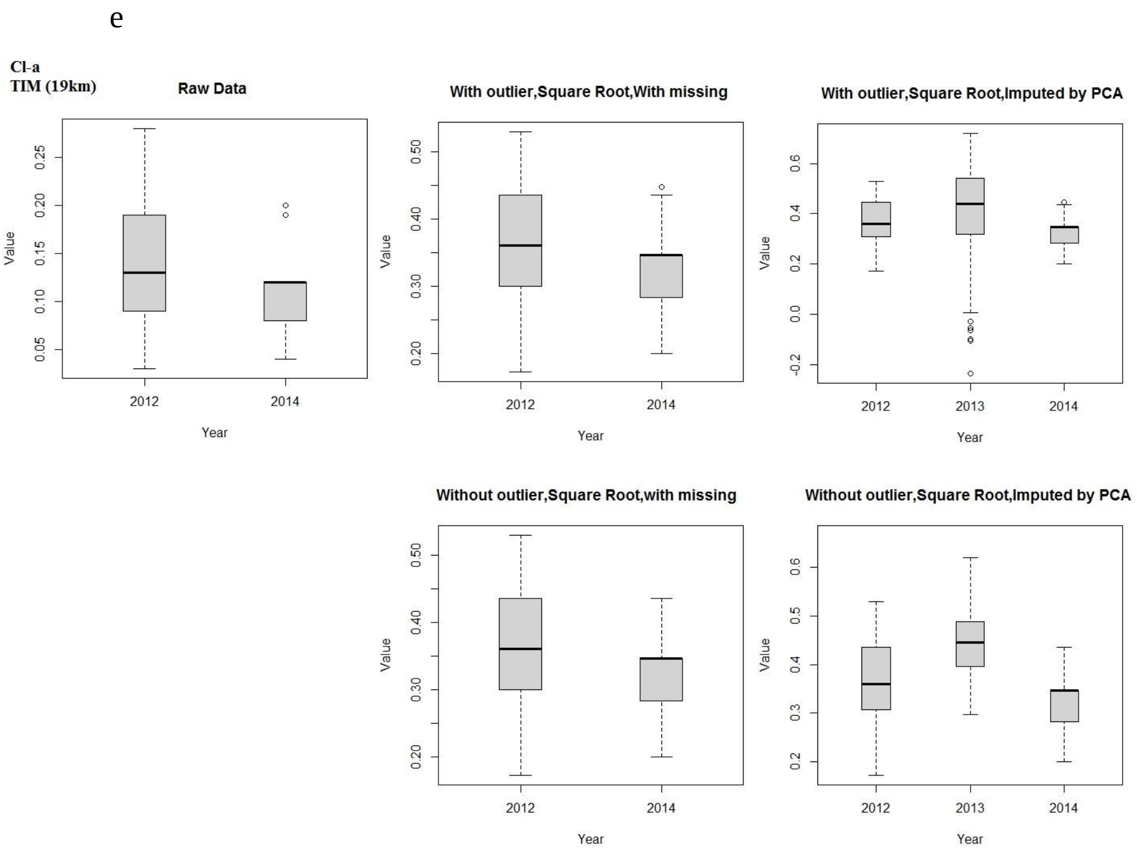
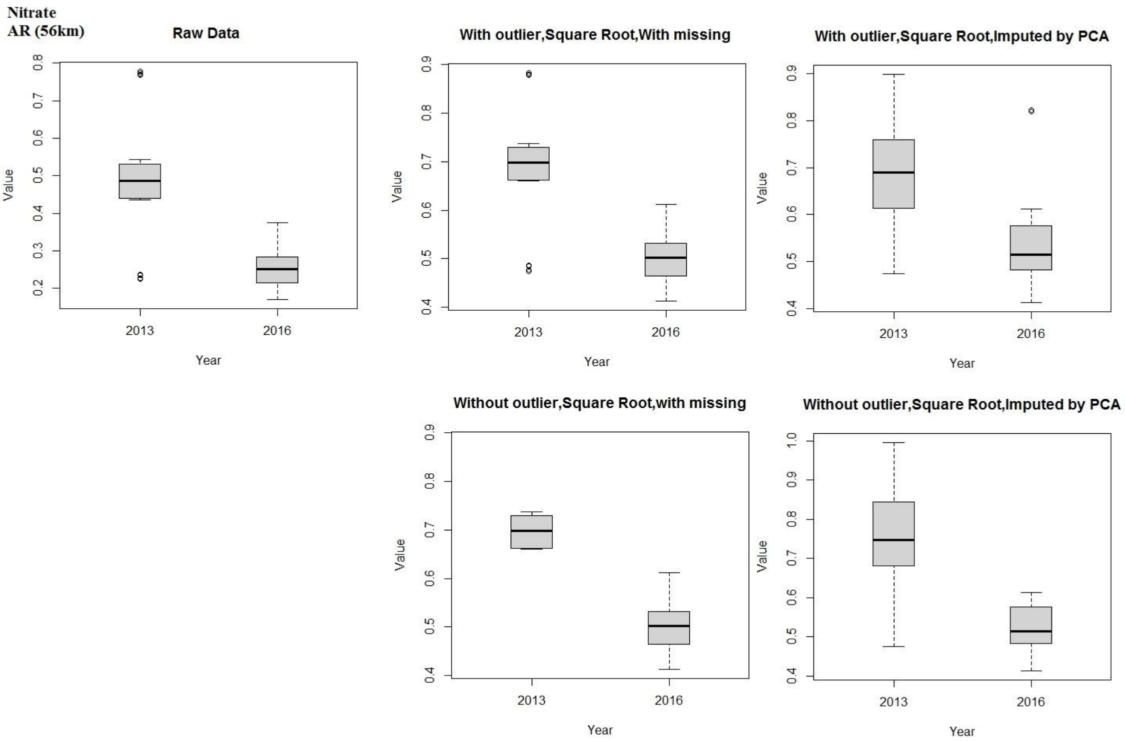


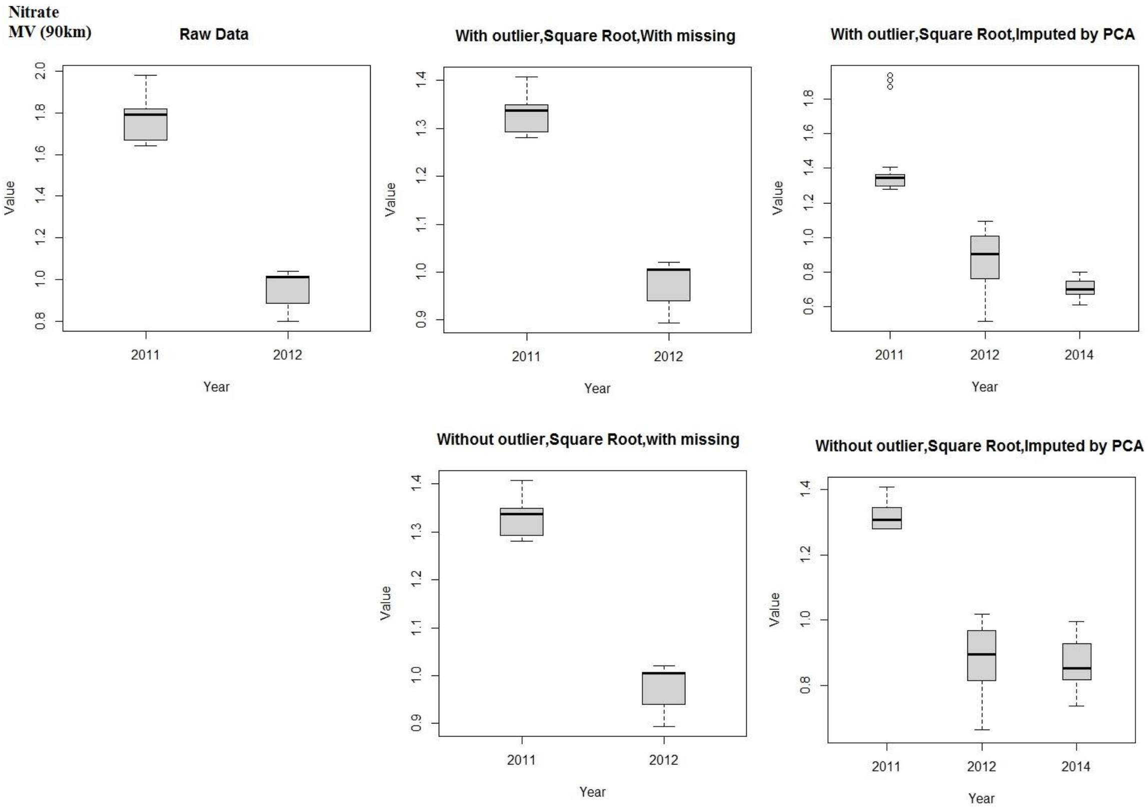
Fig. 37 Distribution of Cla (62.5 % missing) in different years and in a) MV (90km), b) PAB (89km), c) PL (13km), d) SG (22km) and e) TIM (19km).

Fig. 38 shows the distribution of Nitrate in different years and sites. The abundance of Nitrate in AR and 2013 is higher than 2016 (Fig. 38a). .The abundance of Nitrate in MV and 2011 is higher than 2012 (Fig. 38b). The abundance of Nitrate in PAB and 2012 is higher than 2016 (Fig. 38c). The abundance of Nitrate in PL and 2013 is higher than 2016 and 2012 (Fig. 38d). The abundance of Nitrate in SG and 2016 is higher than 2012 (Fig. 38e). The abundance of Nitrate in TIM and 2013 is higher than 2012 (Fig. 38f).

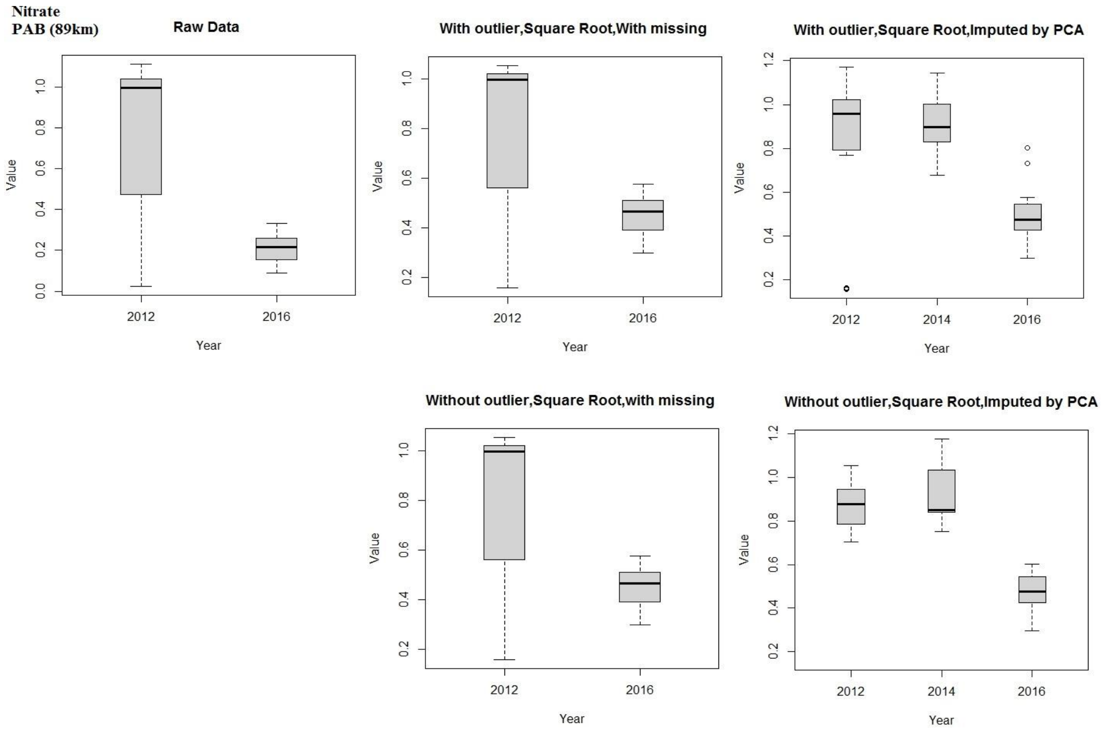
a


Nitrate
AR (56km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2013
2016
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

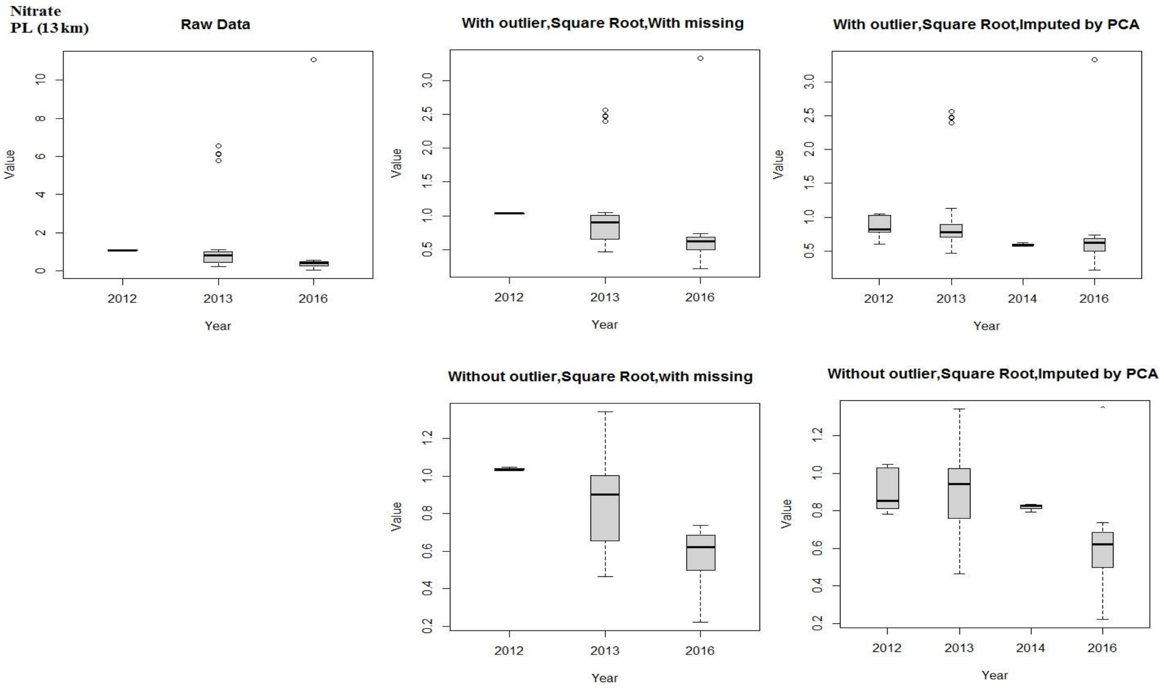
b


Nitrate
MV (90km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2011
2012
2014
0.6
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.6
1.8
2.0

c


Nitrate
PAB (89km)
Raw Data
Value
0.0
0.2
0.4
0.6
0.8
1.0
2012
2016
Year
With outlier,Square Root,With missing
Value
0.2
0.4
0.6
0.8
1.0
2012
2016
Year
With outlier,Square Root,Imputed by PCA
Value
0.2
0.4
0.6
0.8
1.0
1.2
2012
2014
2016
Year
Without outlier,Square Root,with missing
Value
0.2
0.4
0.6
0.8
1.0
2012
2016
Year
Without outlier,Square Root,Imputed by PCA
Value
0.2
0.4
0.6
0.8
1.0
1.2
2012
2014
2016
Year

d


Nitrate
PL (13 km)
Raw Data
Value
0
2
4
6
8
10
2012
2013
2016
Year
With outlier,Square Root,With missing
Value
0.5
1.0
1.5
2.0
2.5
3.0
2012
2013
2016
Year
With outlier,Square Root,Imputed by PCA
Value
0.5
1.0
1.5
2.0
2.5
3.0
2012
2013
2014
2016
Year
Without outlier,Square Root,with missing
Value
0.2
0.4
0.6
0.8
1.0
1.2
2012
2013
2016
Year
Without outlier,Square Root,Imputed by PCA
Value
0.2
0.4
0.6
0.8
1.0
1.2
2012
2013
2014
2016
Year

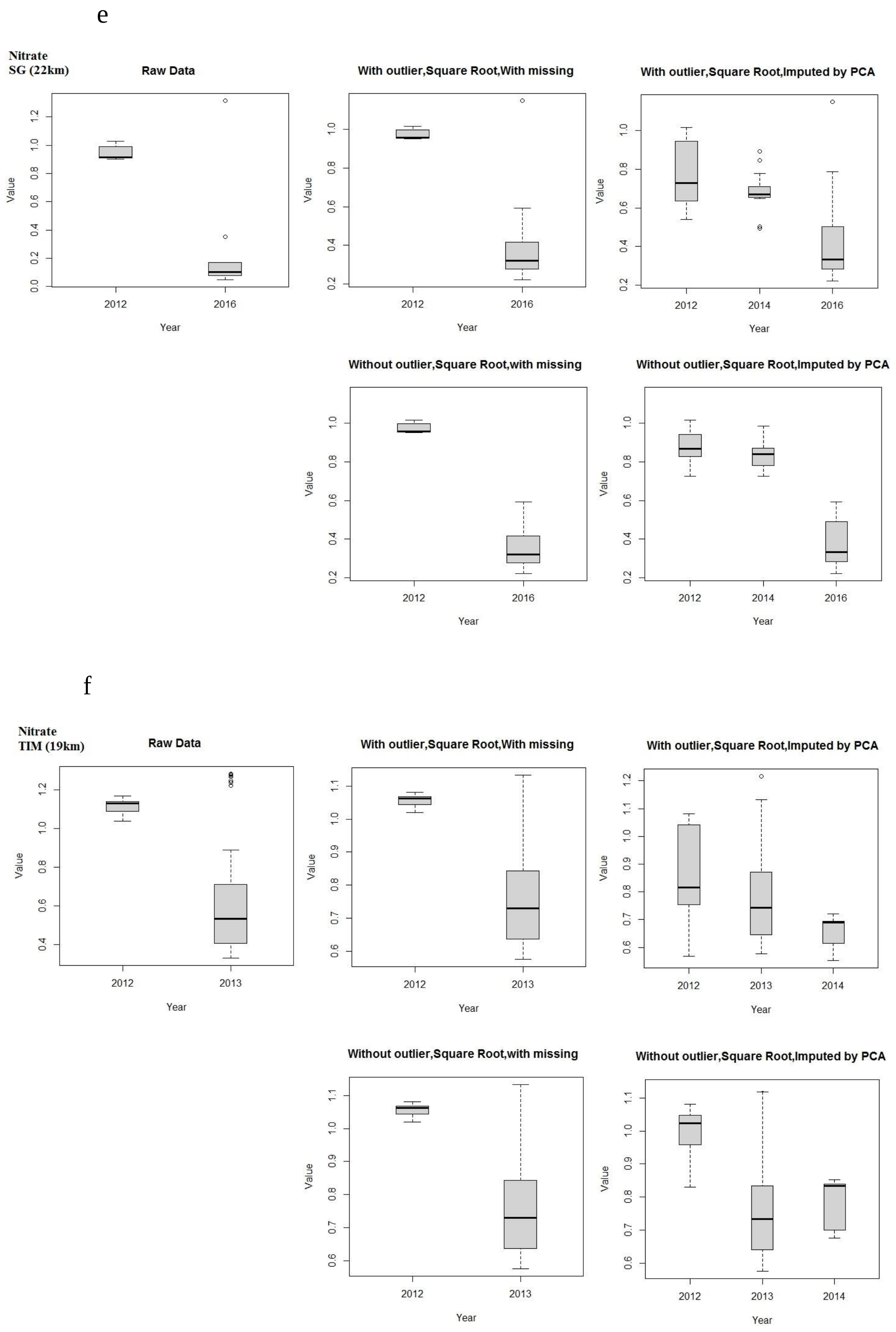
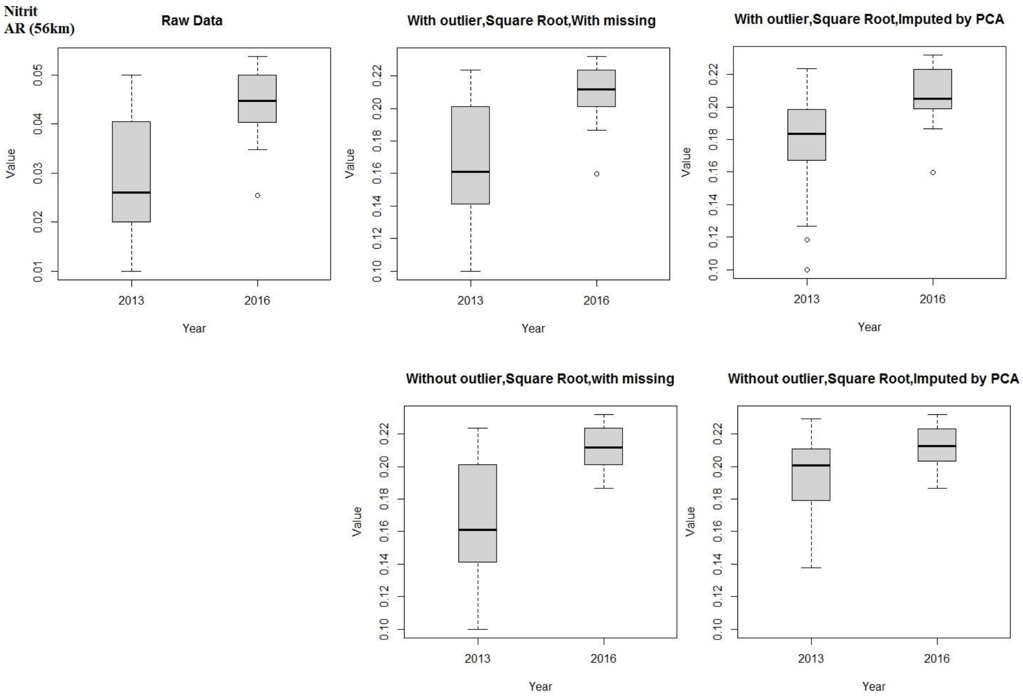


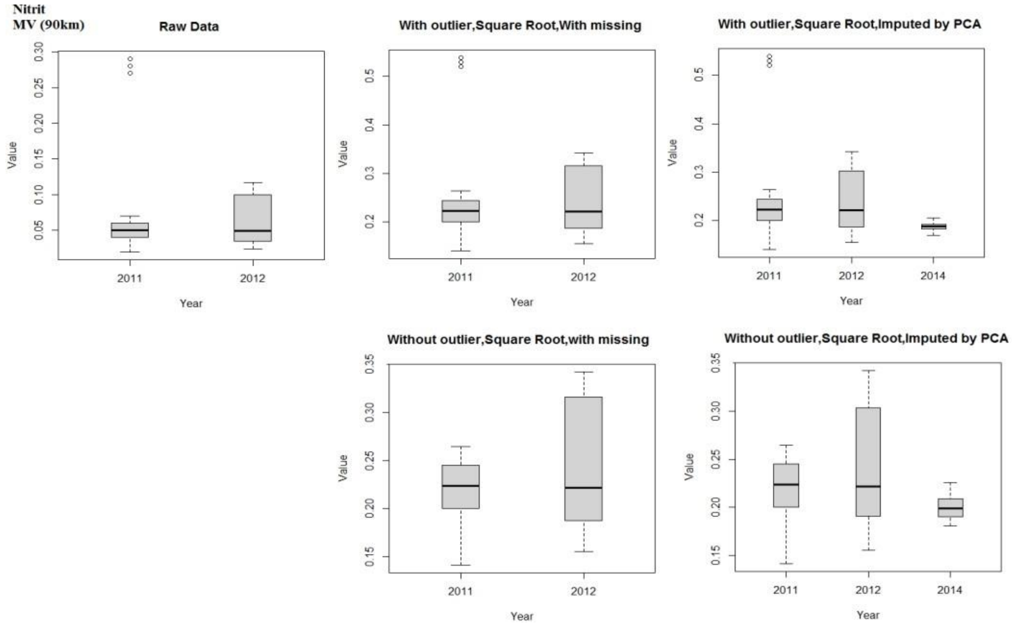
Fig. 38 Distribution of Nitrate (53.31 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

Fig. 39 shows the distribution of Nitrit in different years and sites. The abundance of Nitrit in AR and 2013 is higher than 2016 (Fig. 39a). The abundance of Nitrit in MV and2012 is higher than 2011 (Fig. 39b). The abundance of Nitrit in PAB and 2016 is higher than 2012 (Fig. 39c). The abundance of Nitrit in PL and 2013 is higher than 2016 and 2012 (Fig. 39d). The abundance of Nitrit in SG and 2016 is higher than 2012 (Fig. 39e). The abundance of Nitrit in TIM and 2013 is higher than 2012 (Fig. 39f).

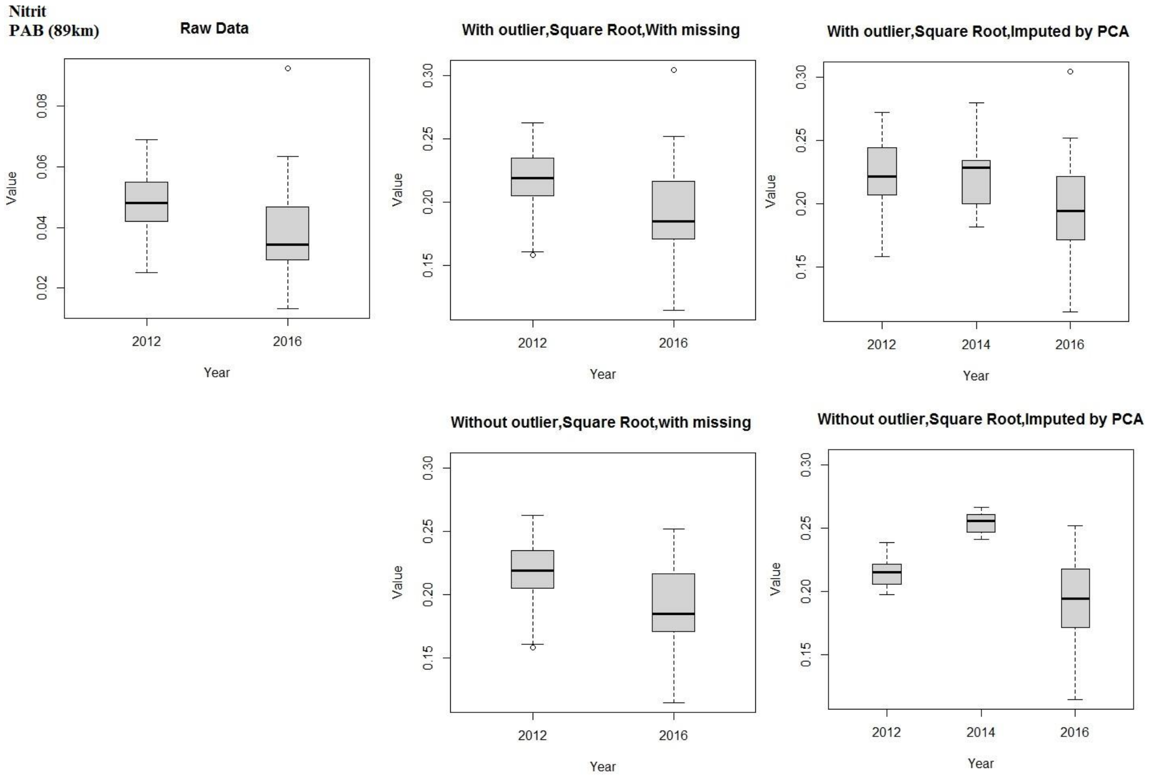
a



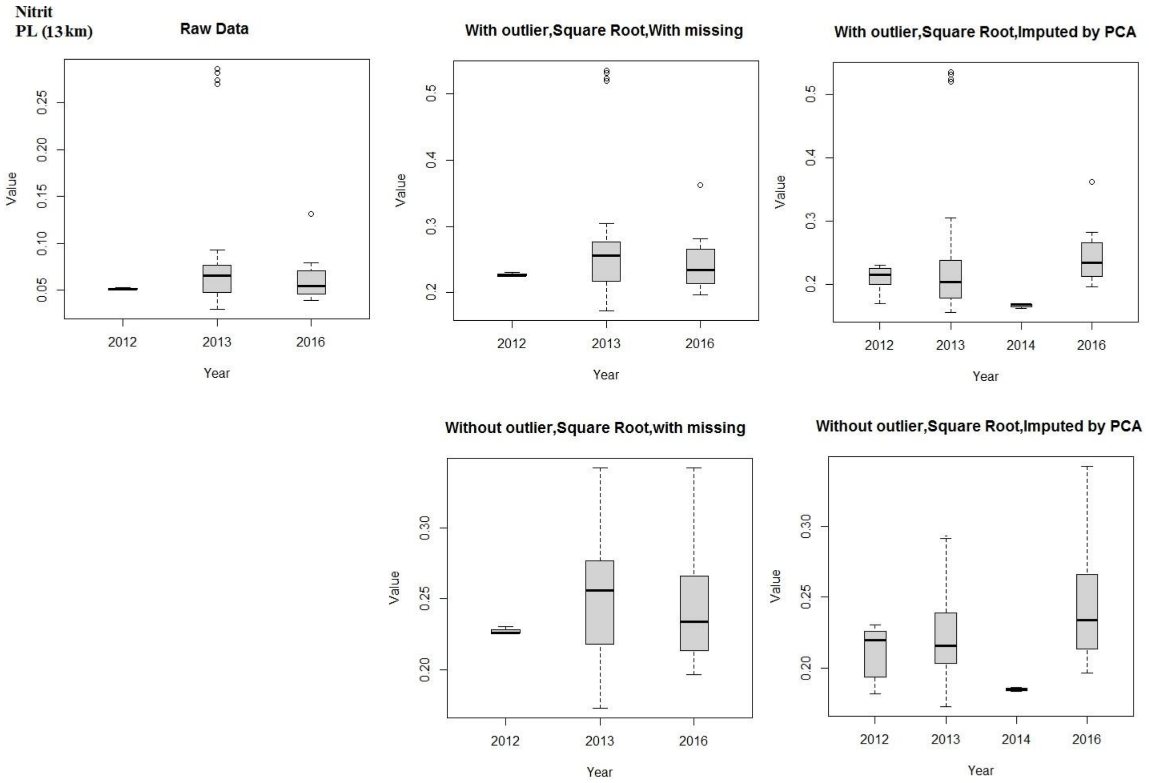
b



c


Nitrit
PAB (89km)
Raw Data
Value
0.02
0.04
0.06
0.08
2012
2016
Year
With outlier,Square Root,With missing
Value
0.15
0.20
0.25
0.30
2012
2016
Year
With outlier,Square Root,Imputed by PCA
Value
0.15
0.20
0.25
0.30
2012
2014
2016
Year
Without outlier,Square Root,with missing
Value
0.15
0.20
0.25
0.30
2012
2016
Year
Without outlier,Square Root,Imputed by PCA
Value
0.15
0.20
0.25
0.30
2012
2014
2016
Year

d


Nitrit
PL (13km)
Raw Data
Value
0.05
0.10
0.15
0.20
0.25
2012
2013
2016
Year
With outlier,Square Root,With missing
Value
0.2
0.3
0.4
0.5
2012
2013
2016
Year
With outlier,Square Root,Imputed by PCA
Value
0.2
0.3
0.4
0.5
2012
2013
2014
2016
Year
Without outlier,Square Root,with missing
Value
0.20
0.25
0.30
2012
2013
2016
Year
Without outlier,Square Root,Imputed by PCA
Value
0.20
0.25
0.30
2012
2013
2014
2016
Year

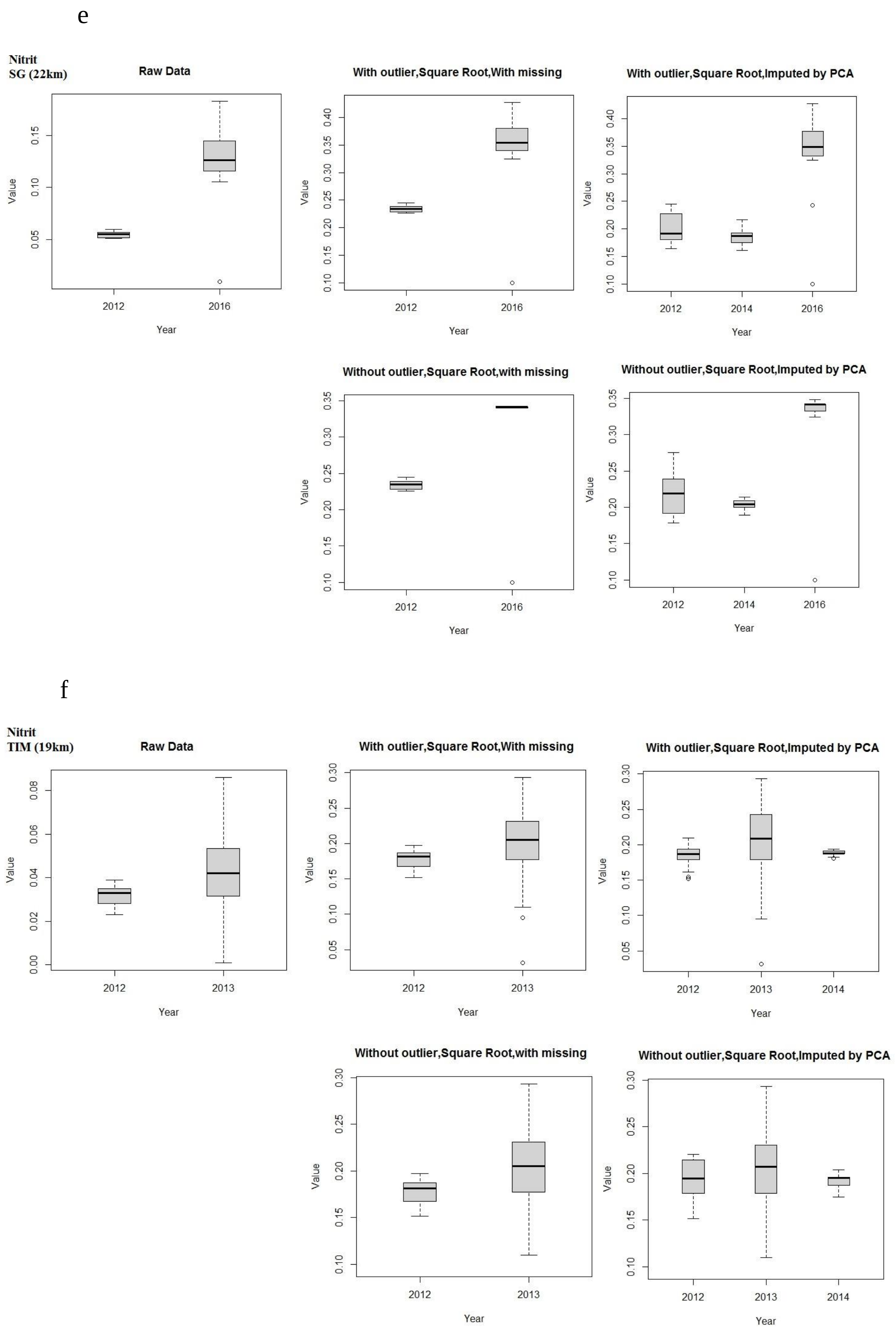
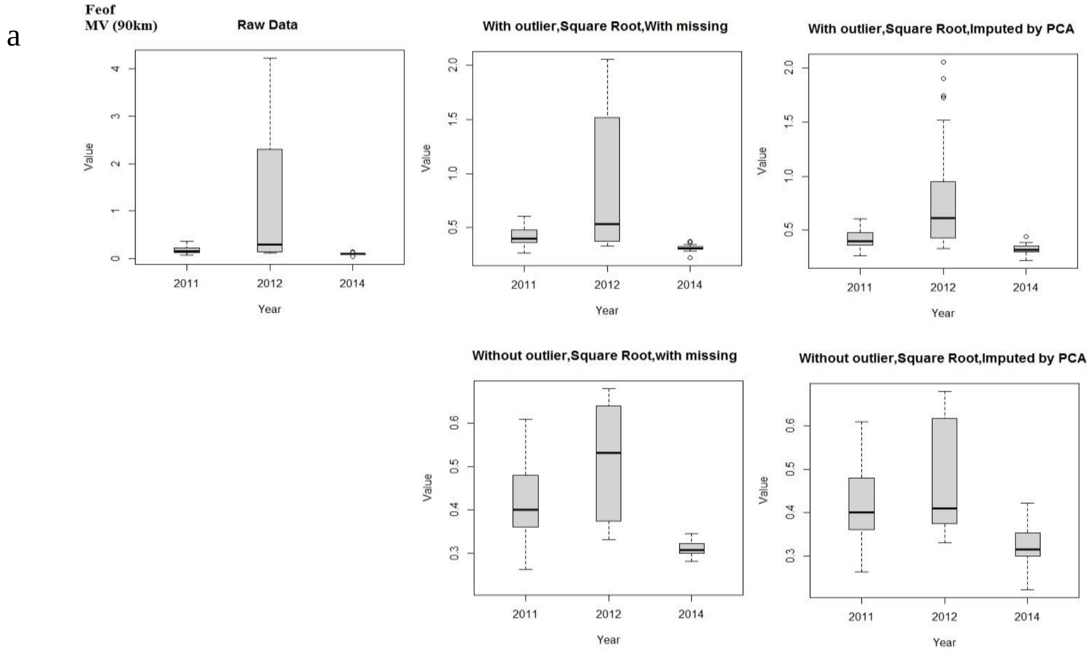
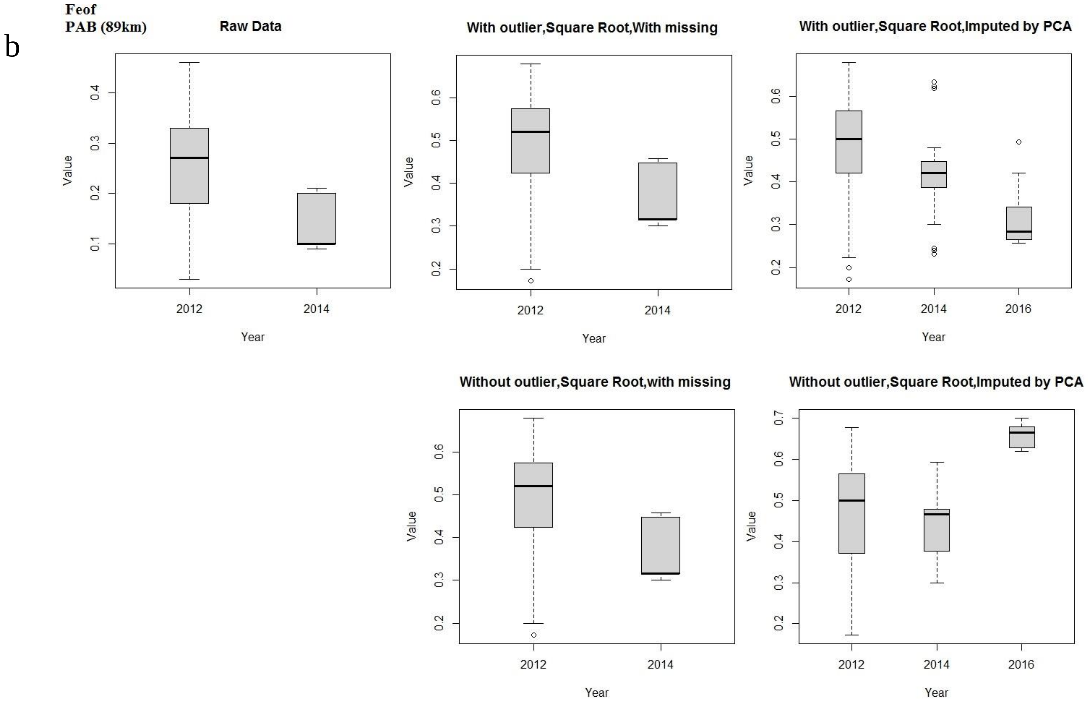


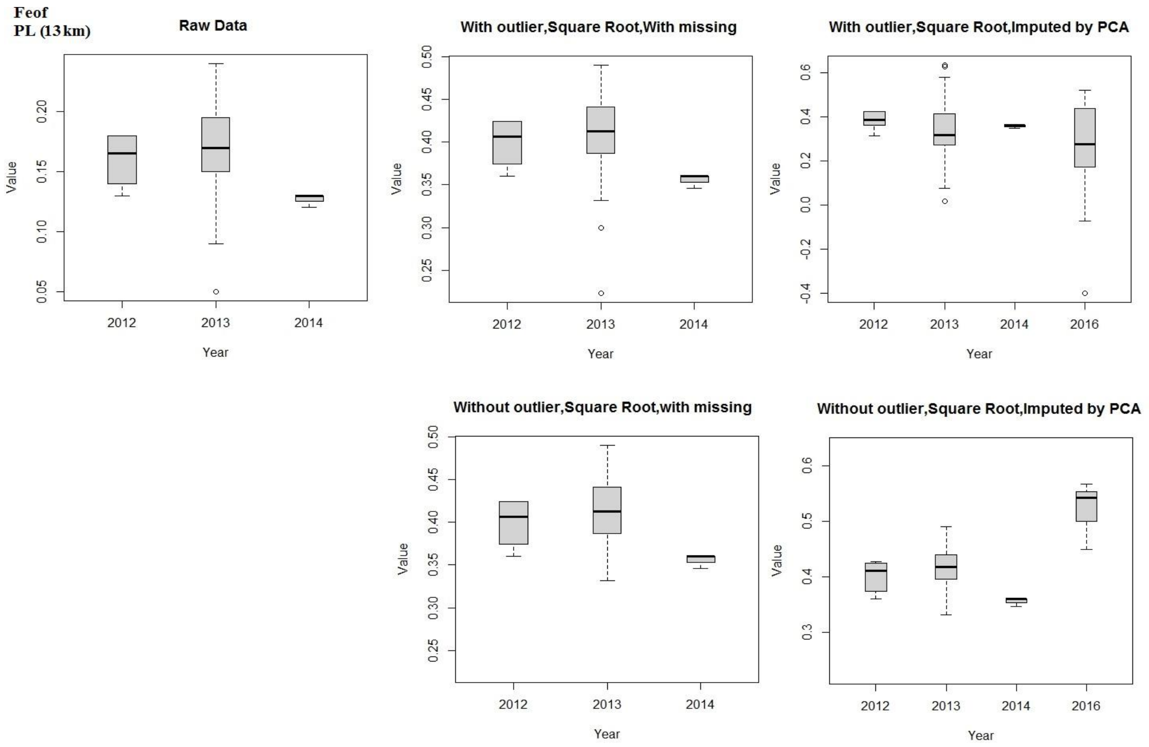
Fig. 39 Distribution of Nitrate (47.79 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

Fig. 40 shows the distribution of Feof in different years and sites. The abundance of Feof in MV and 2012 is higher than 2011 and 2014 (Fig. 40a). The abundance of Feof in PAB and 2012 is higher than 2014 (Fig. 40b). The abundance of Feof in PL and 2013 is higher than 2012 and 2014 (Fig. 40c). The abundance of Feof in SG and 2012 is higher than 2014 (Fig. 40d). The abundance of Feof in TIM and 2012 is higher than 2014 (Fig. 40e).

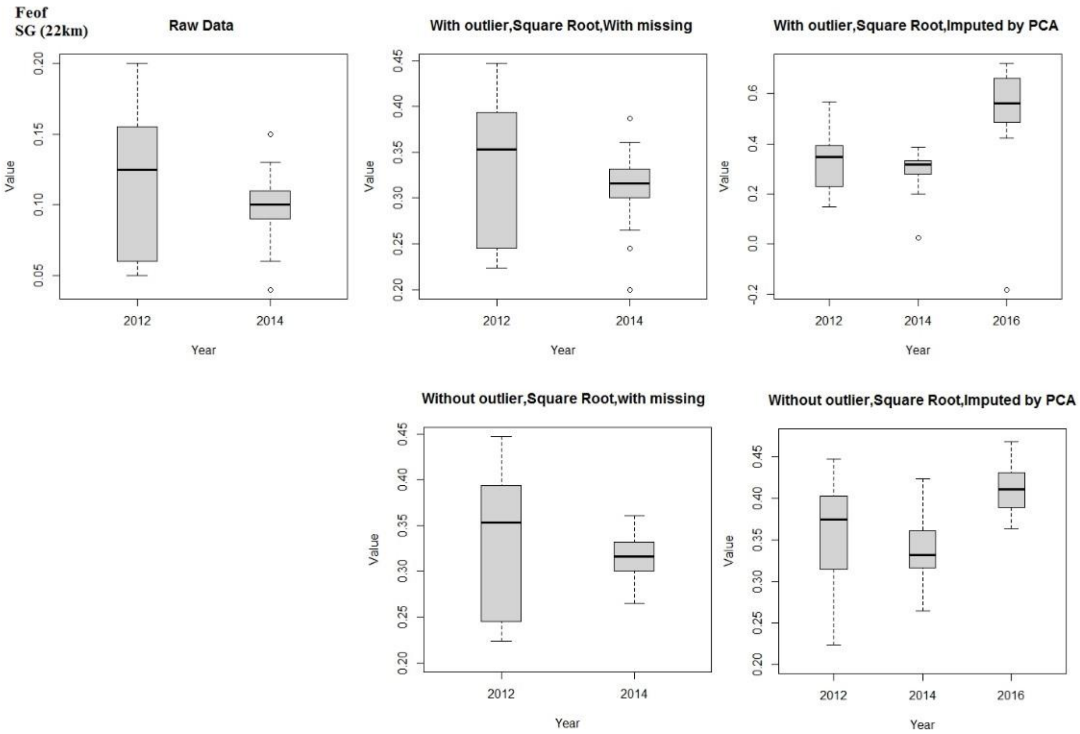




c


Feof
PL (13km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2013
2014
2016
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
-0.4
-0.2
0.0
0.2
0.4
0.6
0.3
0.5

d


Feof
SG (22km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2014
2016
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
-0.2
0.0
0.2
0.4
0.6

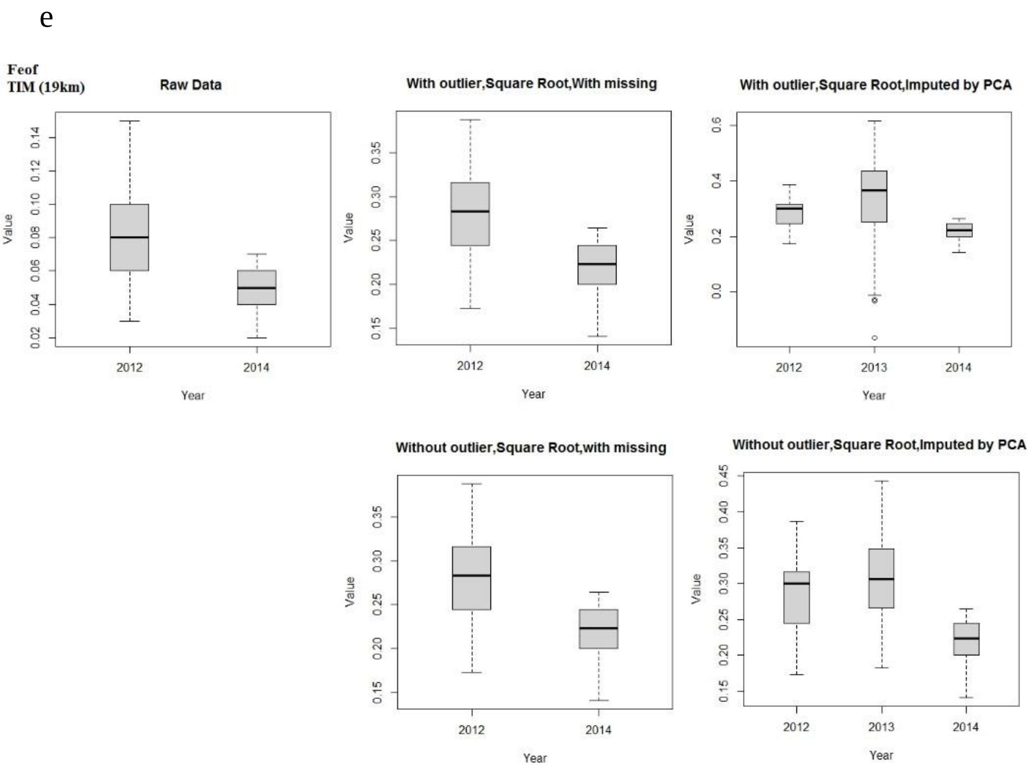
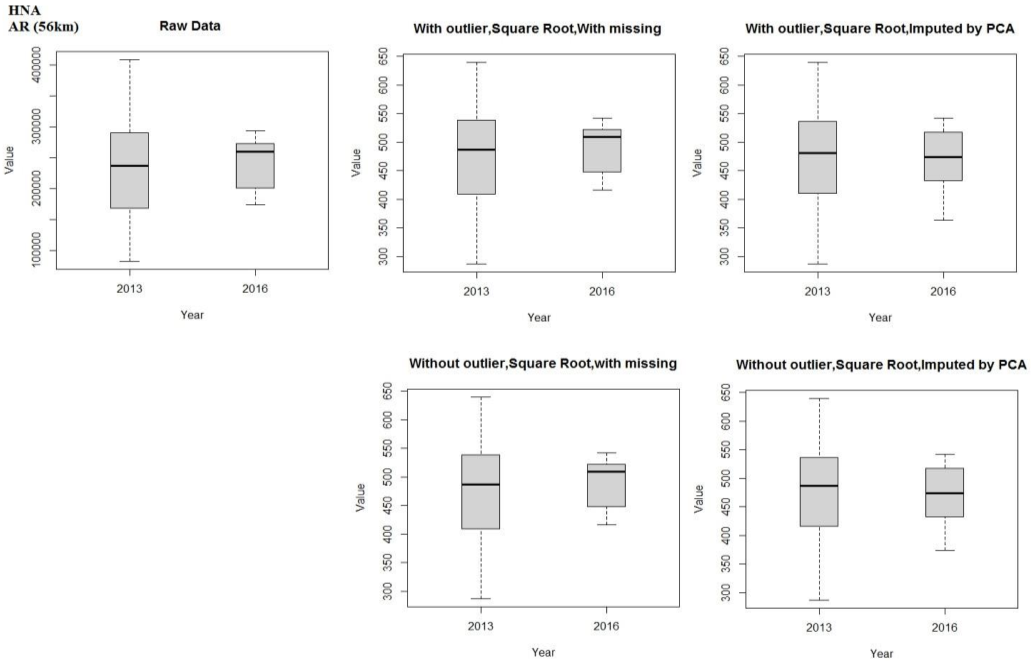


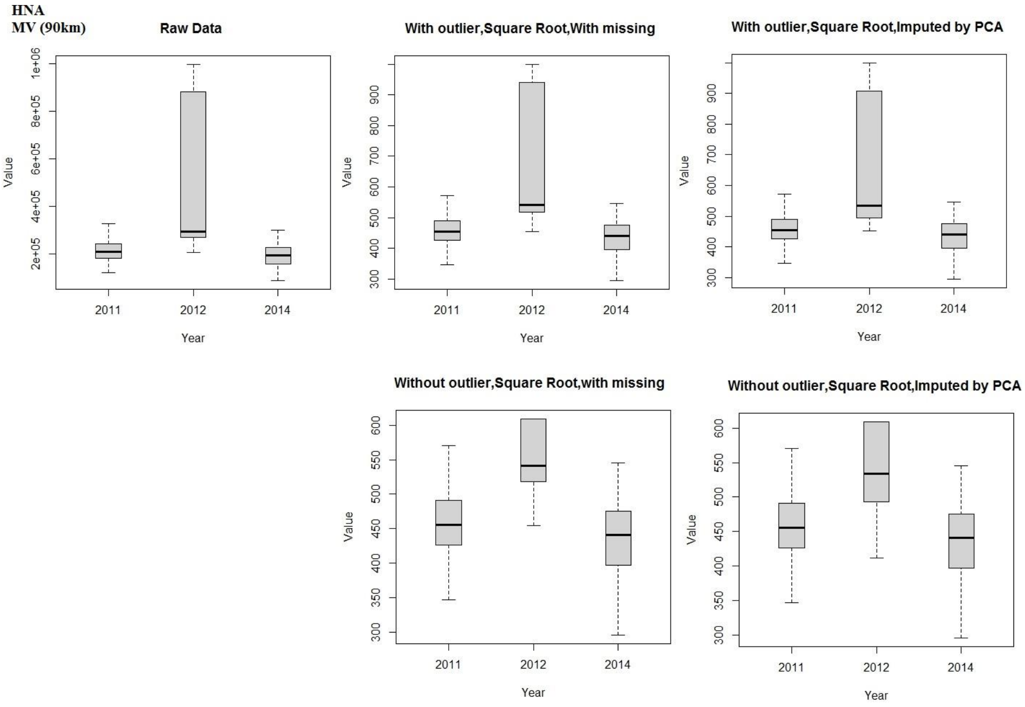
Fig. 40 Distribution of Feof (62.5 % missing) in different years and in a) MV (90km),b) PAB (89km), c) PL (13km), d) SG (22km) and e) TIM (19km).

Fig. 41 shows the distribution of HNA in different years and sites. The abundance of HNA in AR and 2013 is higher than 2016 (Fig. 41a).The abundance of HNA in MV and2012 is higher than 2014 and 2011 (Fig. 41b). The abundance of HNA in PAB and 2012is higher than 2014 and 2016 (Fig. 41c). The abundance of HNA in PL and 2013 is higher than 2012 and 2014 and 2016 (Fig. 41d). The abundance of HNA in SG and 2012 is higher than 2014 and 2016 (Fig. 41e). The abundance of HNA in TIM and 2013 is higherthan 2012 and 2014 (Fig. 41f).

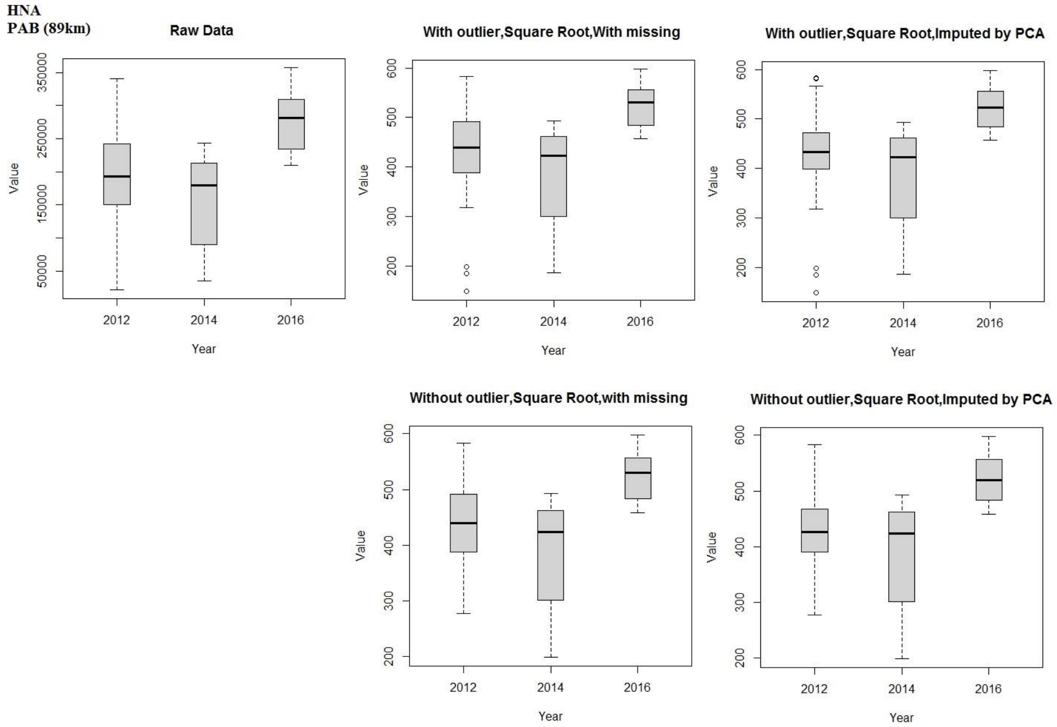
a


HNA
AR (56km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2013
2016
100000
200000
300000
400000
300
350
400
450
500
550
600
650

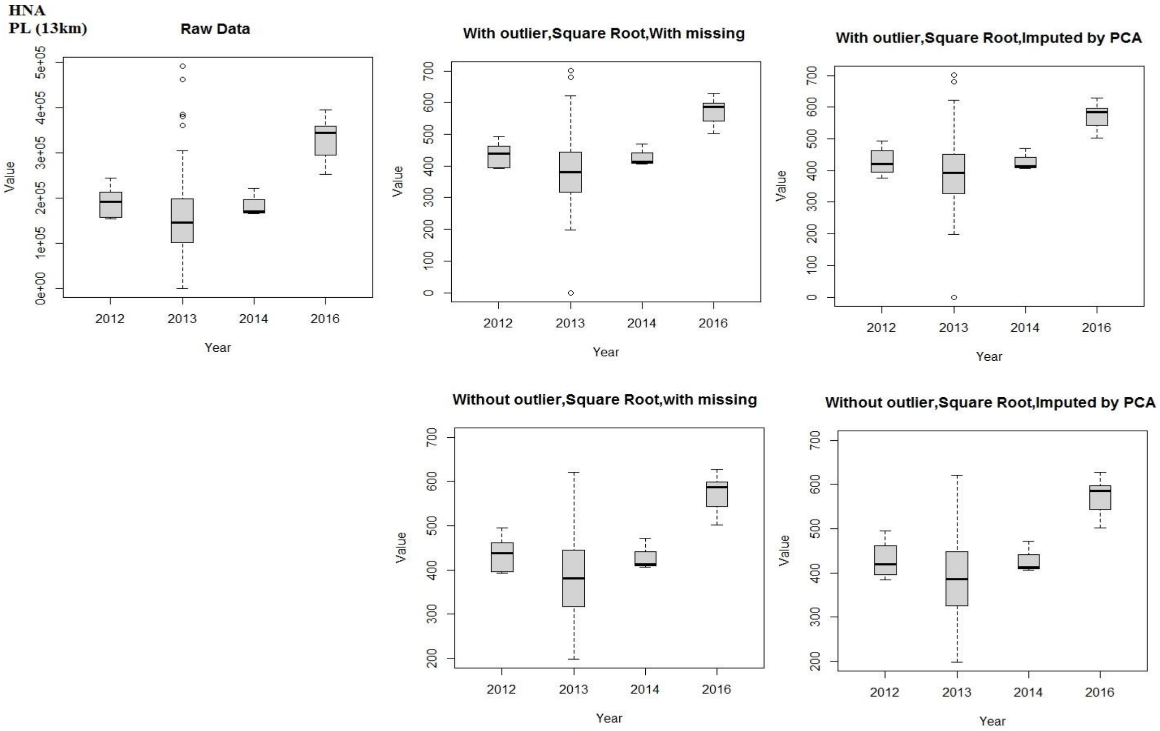
b


HNA
MV (90km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2011
2012
2014
2e+05
4e+05
6e+05
8e+05
1e+06
300
400
500
600
700
800
900
350
450
550

c


HNA
PAB (89km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2014
2016
50000
150000
250000
350000
200
300
400
500
600

d


HNA
PL (13km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2013
2014
2016
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
0
100
200
300
400
500
600
700

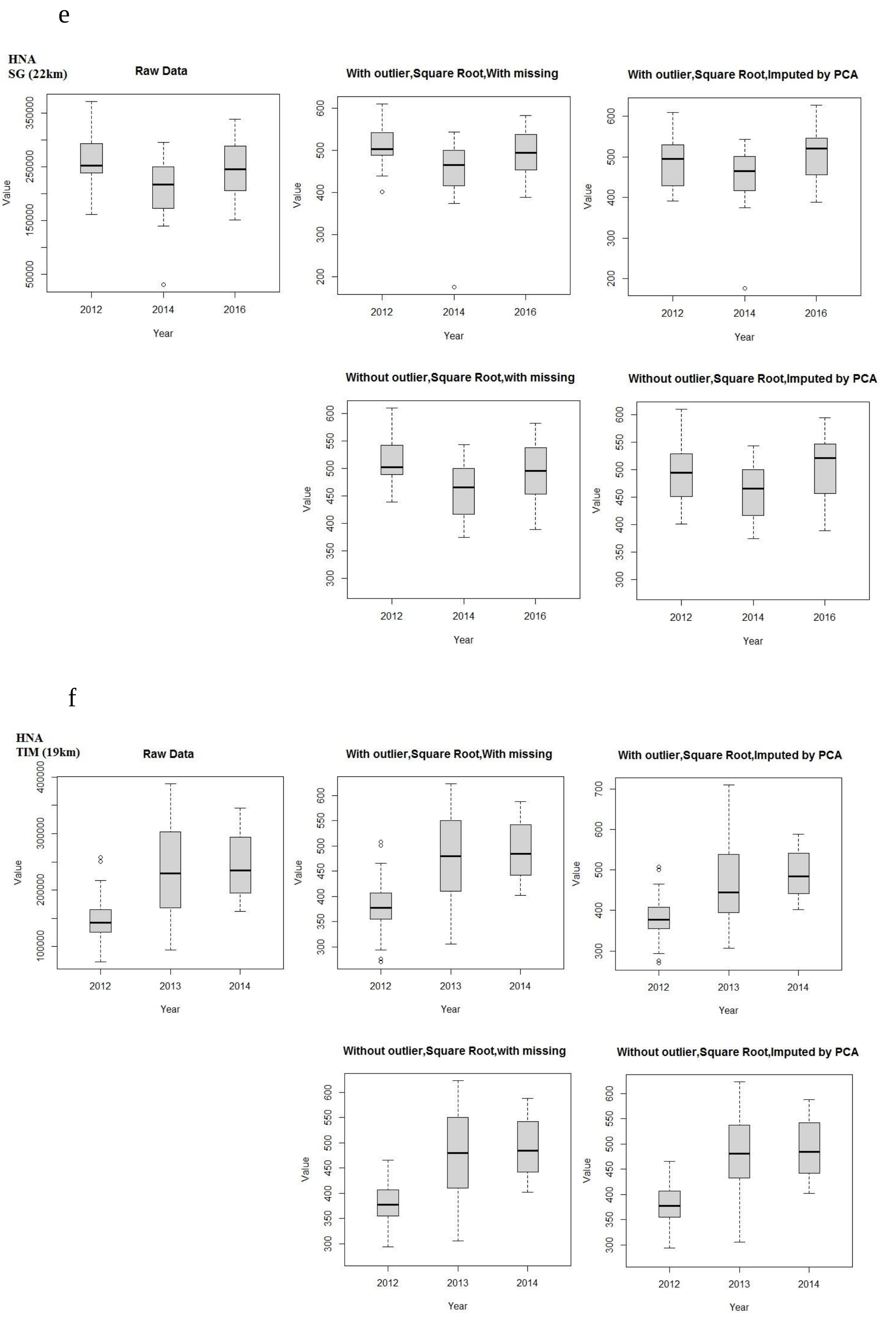
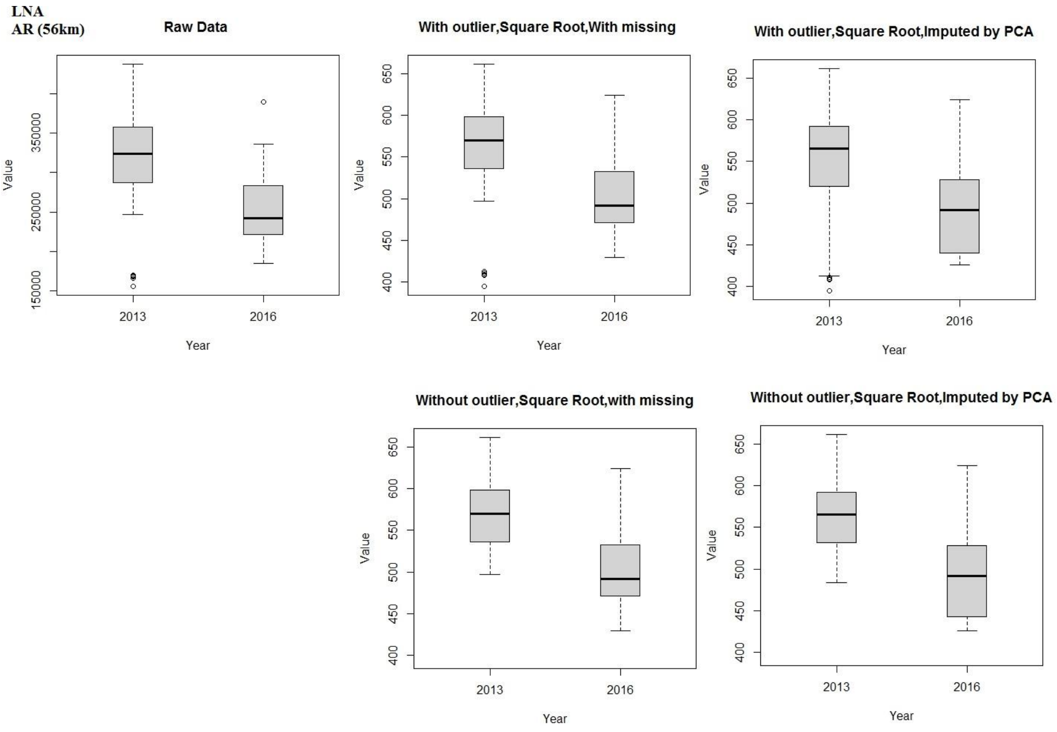


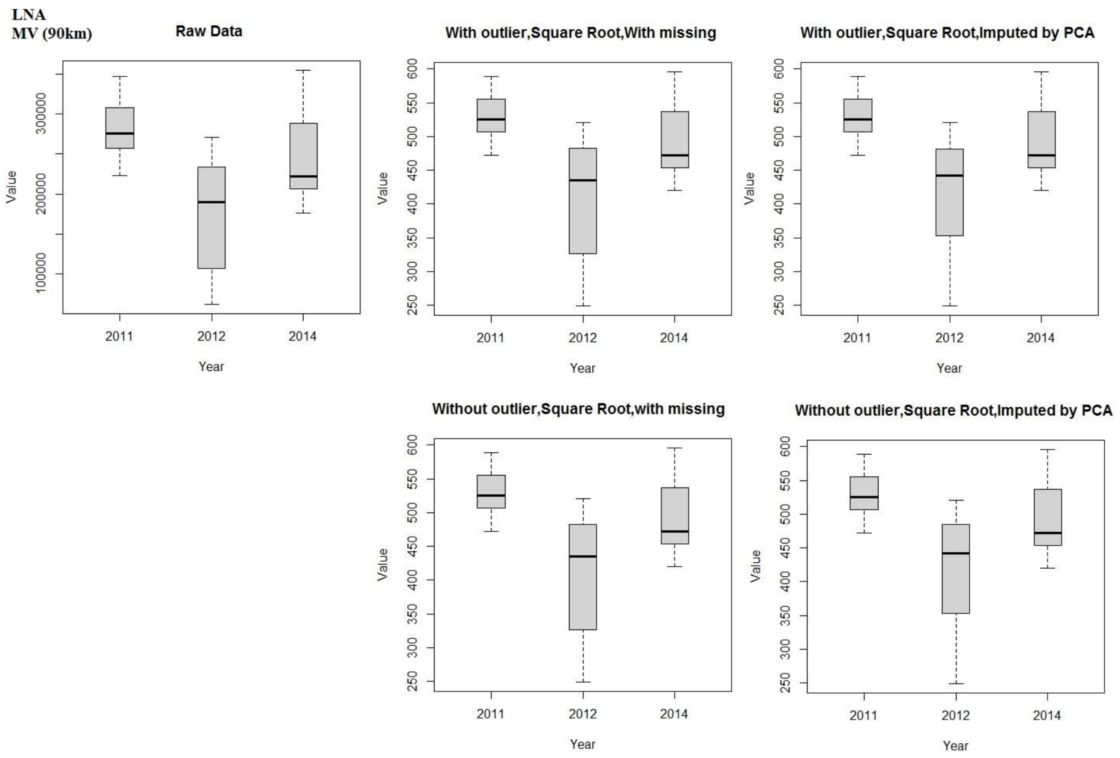
Fig. 41 Distribution of HNA (17.46 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

Fig. 42 shows the distribution of LNA in different years and sites. The abundance of LNA in AR and 2013 is higher than 2016 (Fig. 42a).The abundance of LNA in MV and 2012 is higher than 2012 and 2011 and 2014 (Fig. 42b). The abundance of LNA in PAB and 2016 is higher than 2012 and 2014 (Fig. 42c). The abundance of LNA in PL and 2013 is higher than 2012 and 2014 and 2016 (Fig. 42d). The abundance of LNA in SG and 2014 is higher than 2012 and 2016 (Fig. 42e). The abundance of LNA in TIM and 2013 is higher than 2012 and 2014 (Fig. 42f).

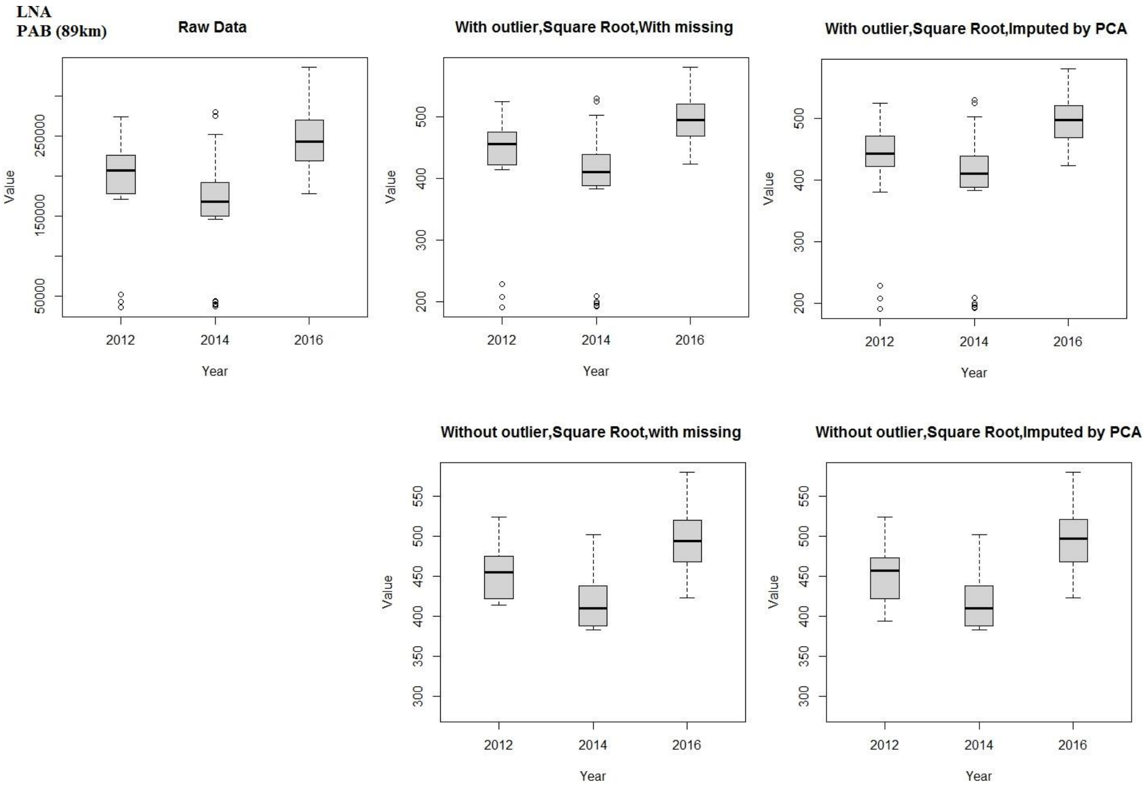
a



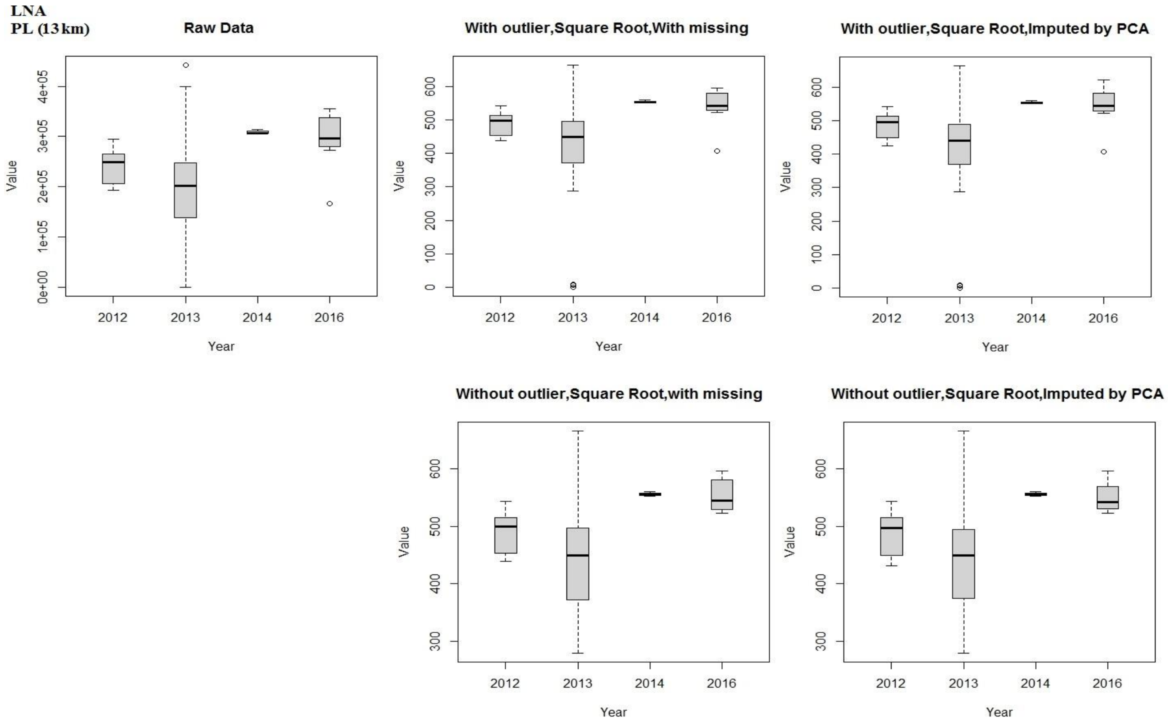
b


LNA
MV (90km)
Raw Data
Value
100000
200000
300000
2011
2012
2014
Year
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
250
300
350
400
450
500
550
600

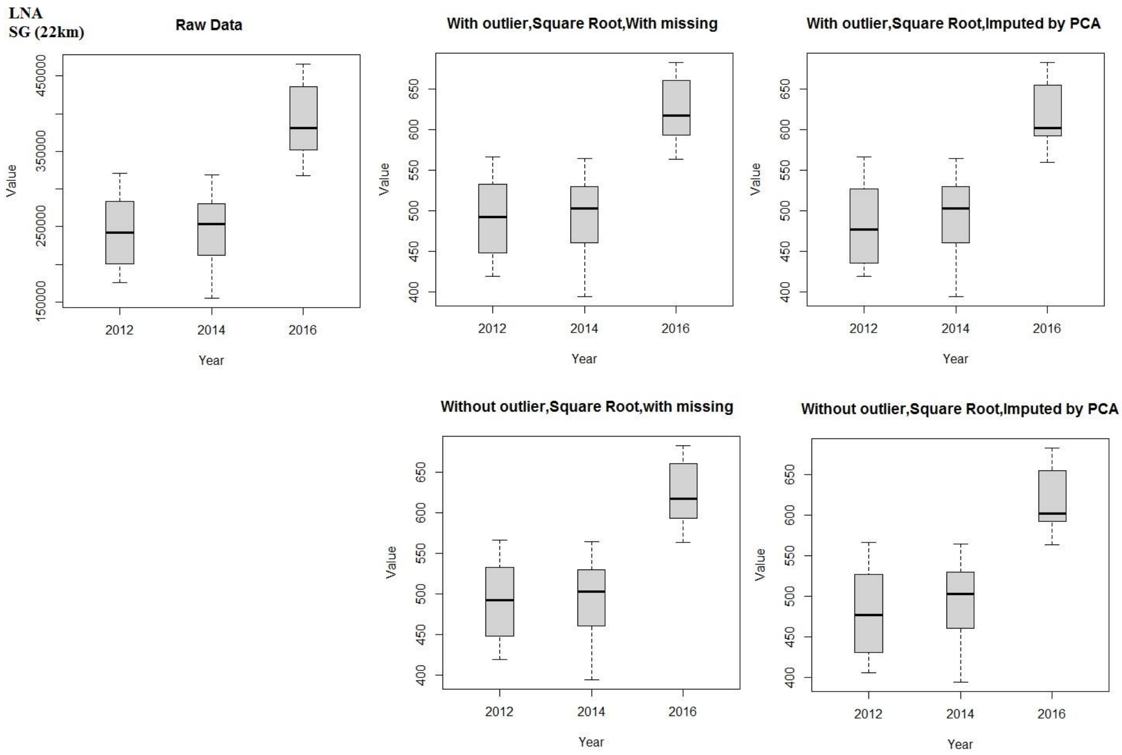
c


LNA
PAB (89km)
Raw Data
Value
50000
150000
250000
2012
2014
2016
Year
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
200
300
400
500
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
300
350
400
450
500
550

d


LNA
PL (13km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
0e+00
1e+05
2e+05
3e+05
4e+05
0
100
200
300
400
500
600
2012
2013
2014
2016

e


LNA
SG (22km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
150000
250000
350000
450000
400
450
500
550
600
650
2012
2014
2016

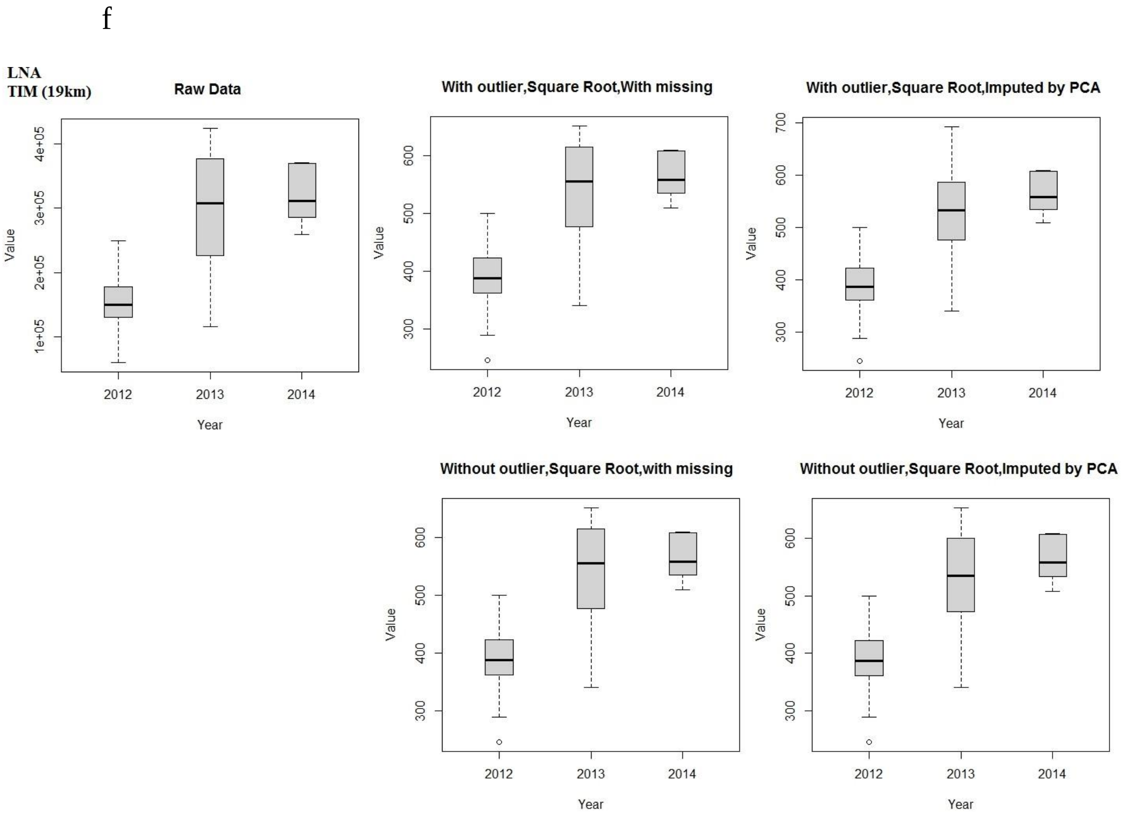
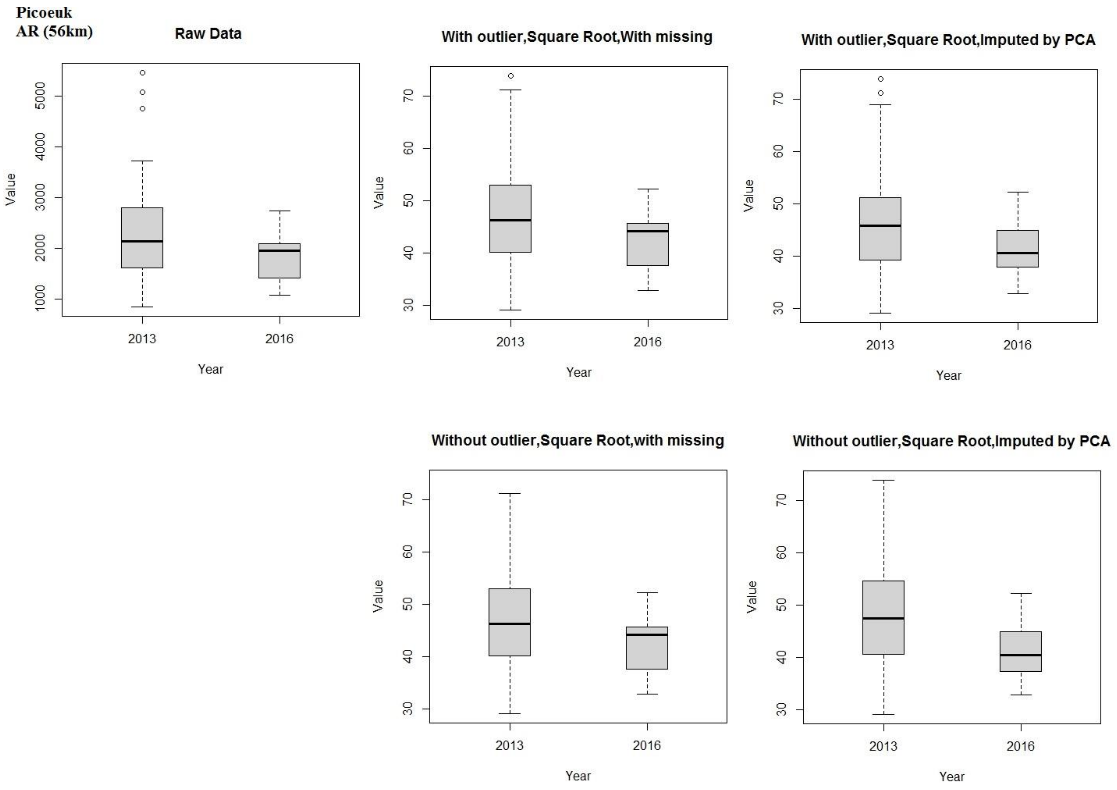


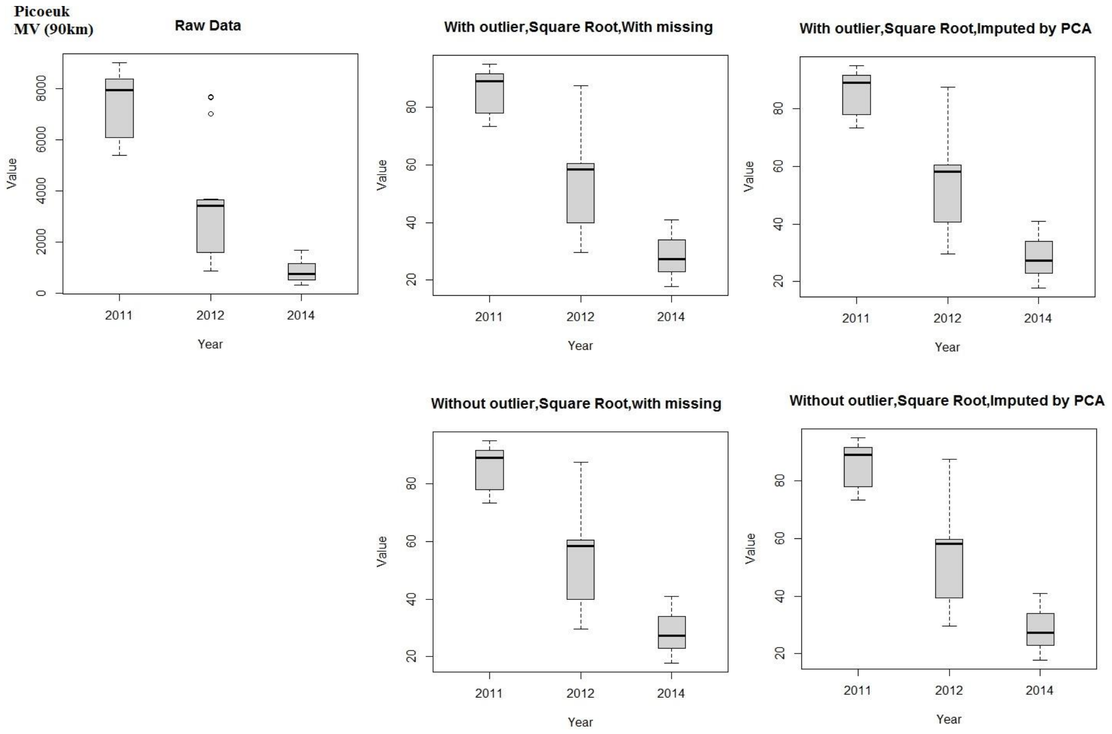
Fig. 42 Distribution of LNA (17.46 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

Fig. 43 shows the distribution of Picoeuk in different years and sites. The abundance of Picoeuk in AR and 2013 is higher than 2016 (Fig. 43a).The abundance of Picoeuk in MV and 2011 is higher than 2012 and 2014 (Fig. 43b). The abundance of Picoeuk in PAB and 2016 is higher than 2012 and 2014 and 2016 (Fig. 43c). The abundance of Picoeuk in PL and 2013 is higher than 2012 and 2014 and 2016 (Fig. 43d). The abundance of Picoeuk in SG and 2016 is higher than 2012 and 2014 (Fig. 43e). The abundance of Picoeuk in TIM and 2013 is higher than 2012 and 2014 (Fig. 43f).

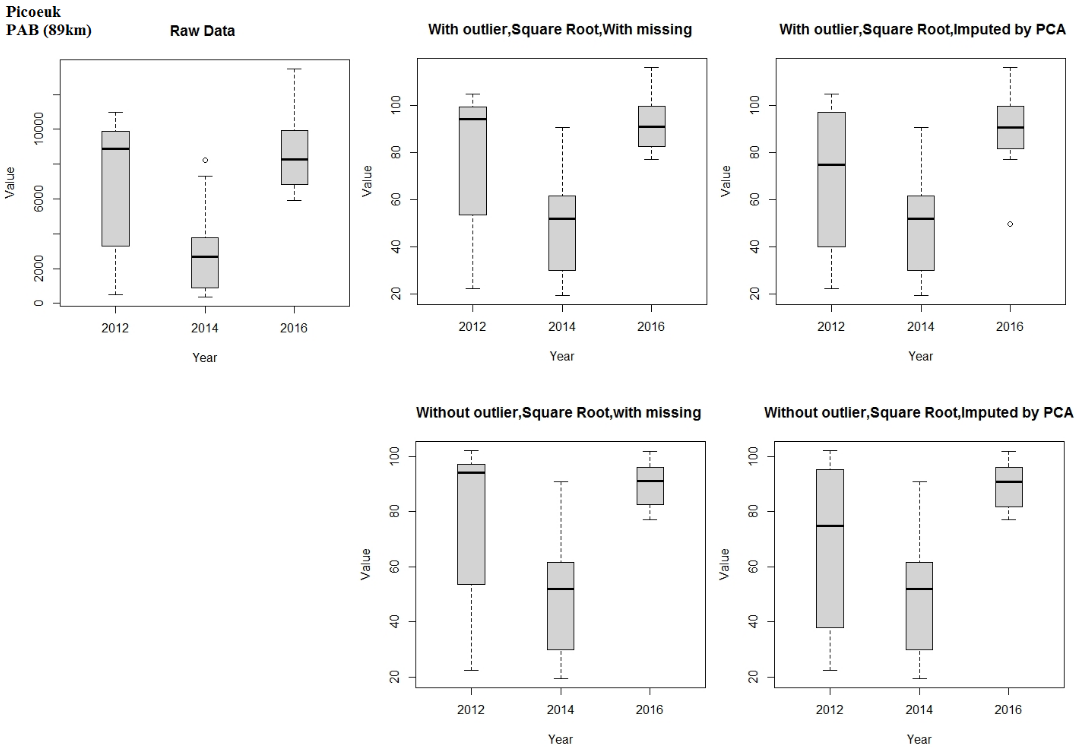
a


Picoeuk
AR (56km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2013
2016
1000
2000
3000
4000
5000
30
40
50
60
70

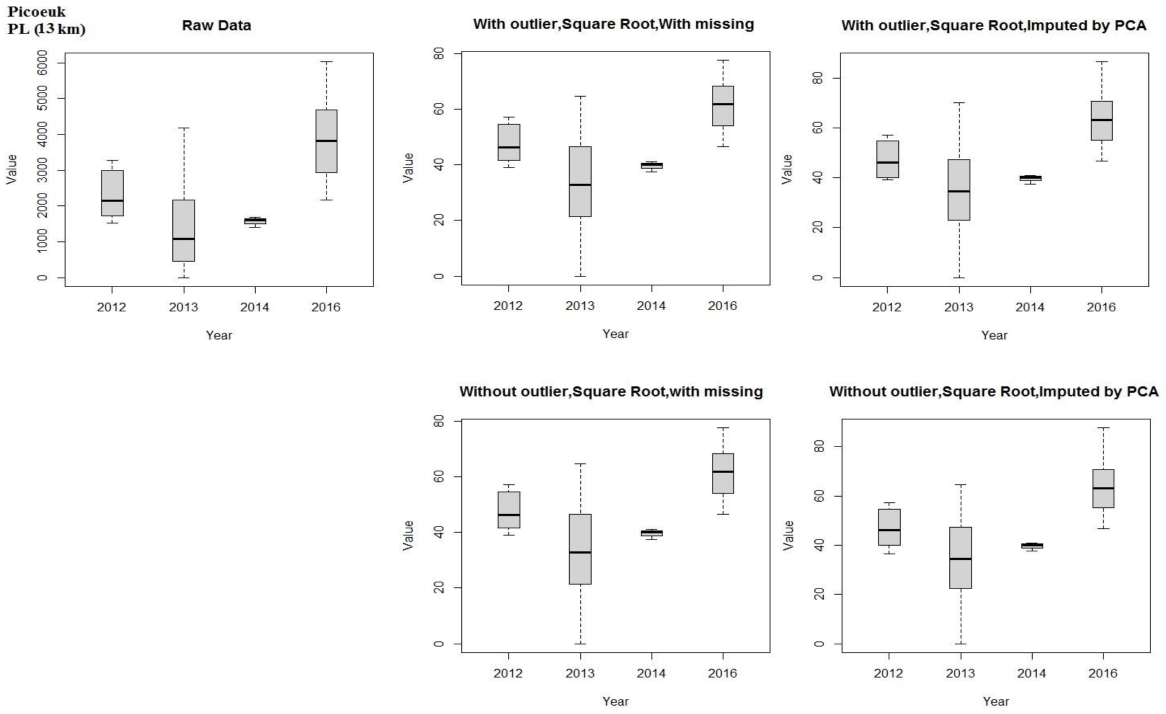
b


Picoeuk
MV (90km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2011
2012
2014
0
2000
4000
6000
8000
20
40
60
80

c


Picoeuk
PAB (89km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2014
2016
0
2000
6000
10000
20
40
60
80
100

d


Picoeuk
PL (13 km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2013
2014
2016
0
1000
2000
3000
4000
5000
6000
20
40
60
80

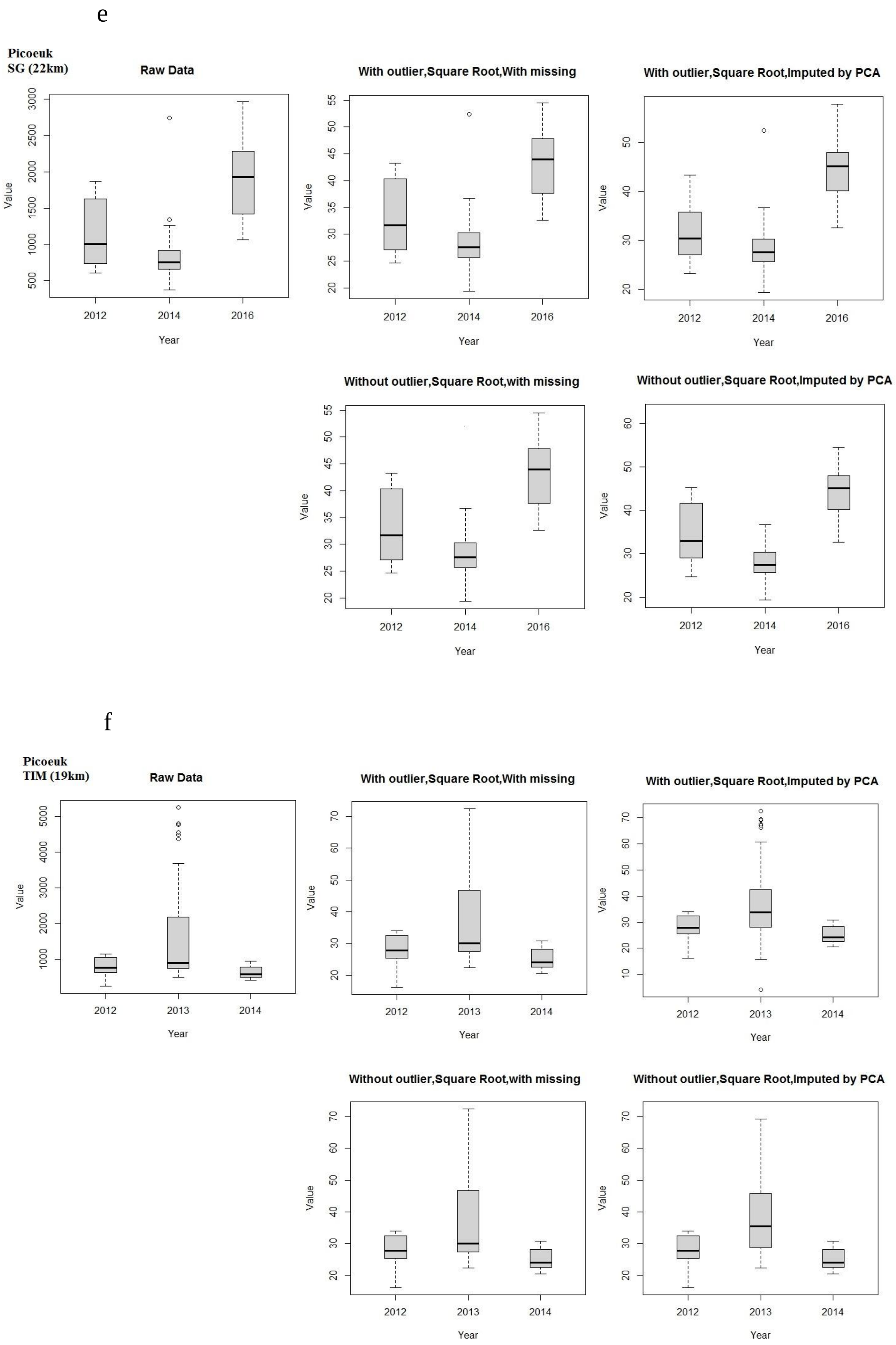
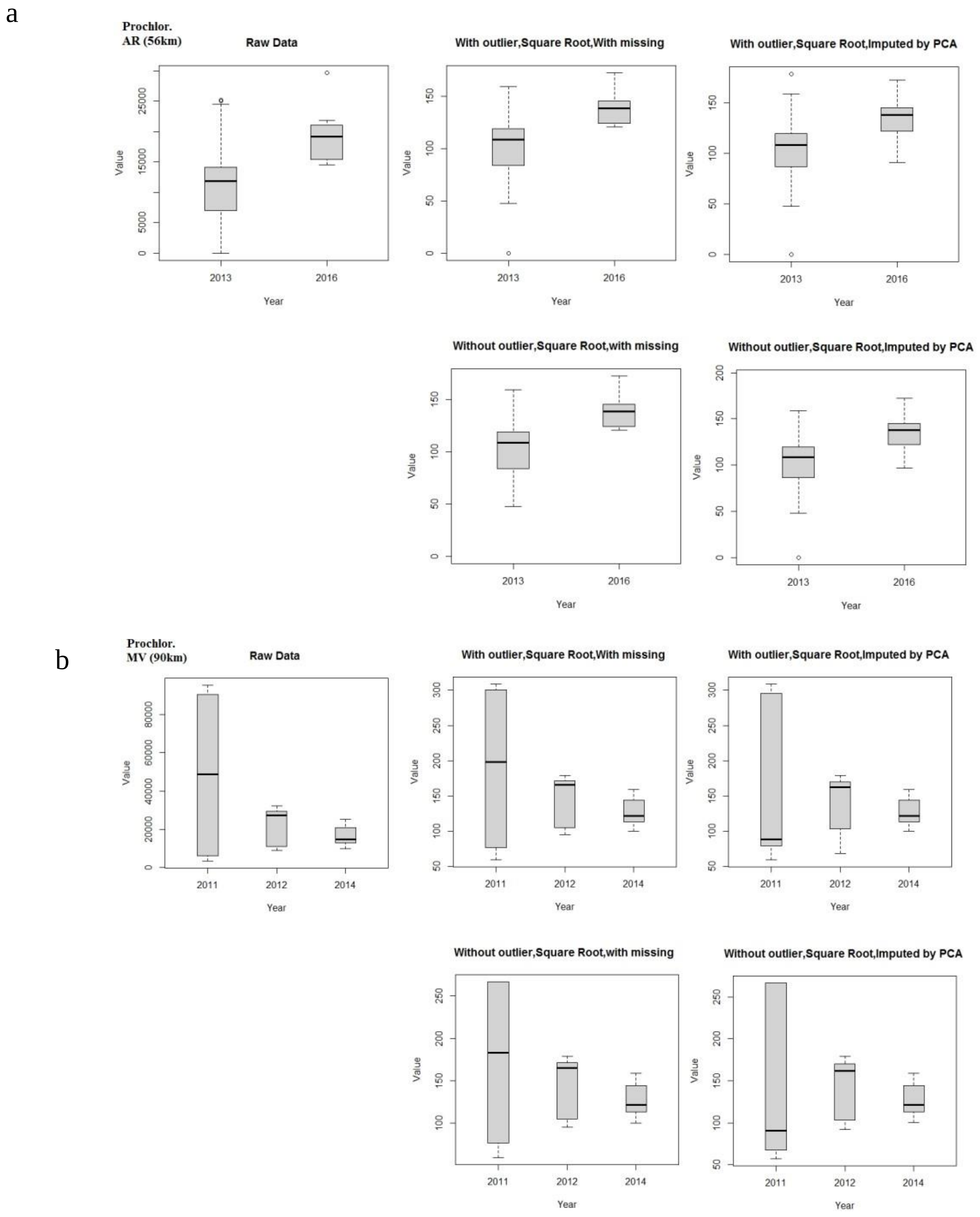
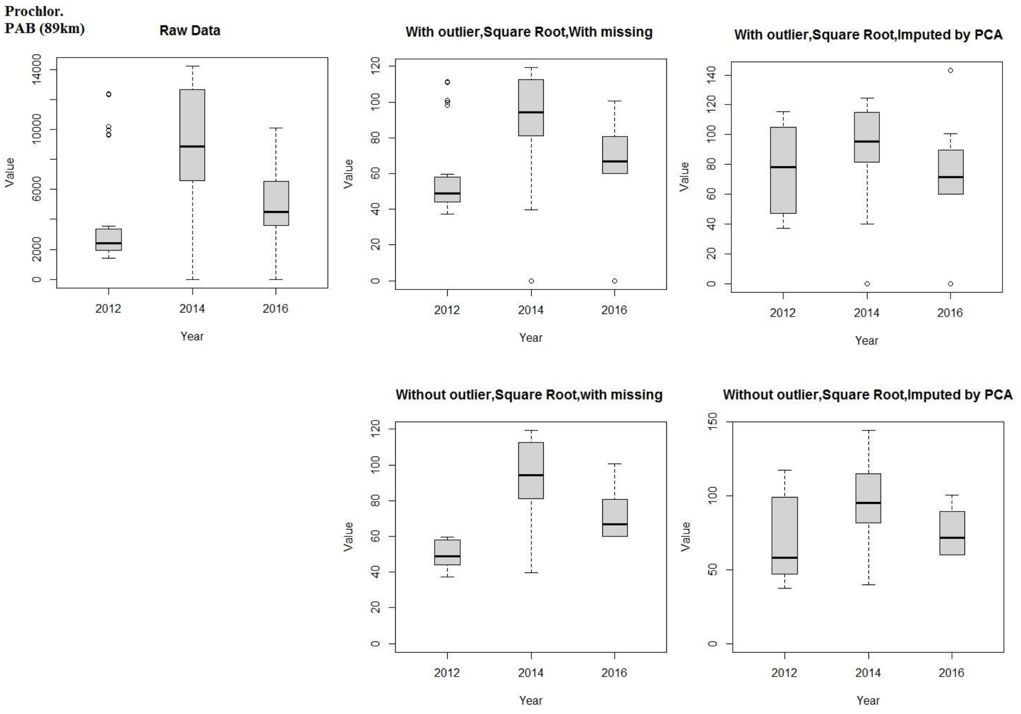


Fig. 43 Distribution of Picoeuk (15.81 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

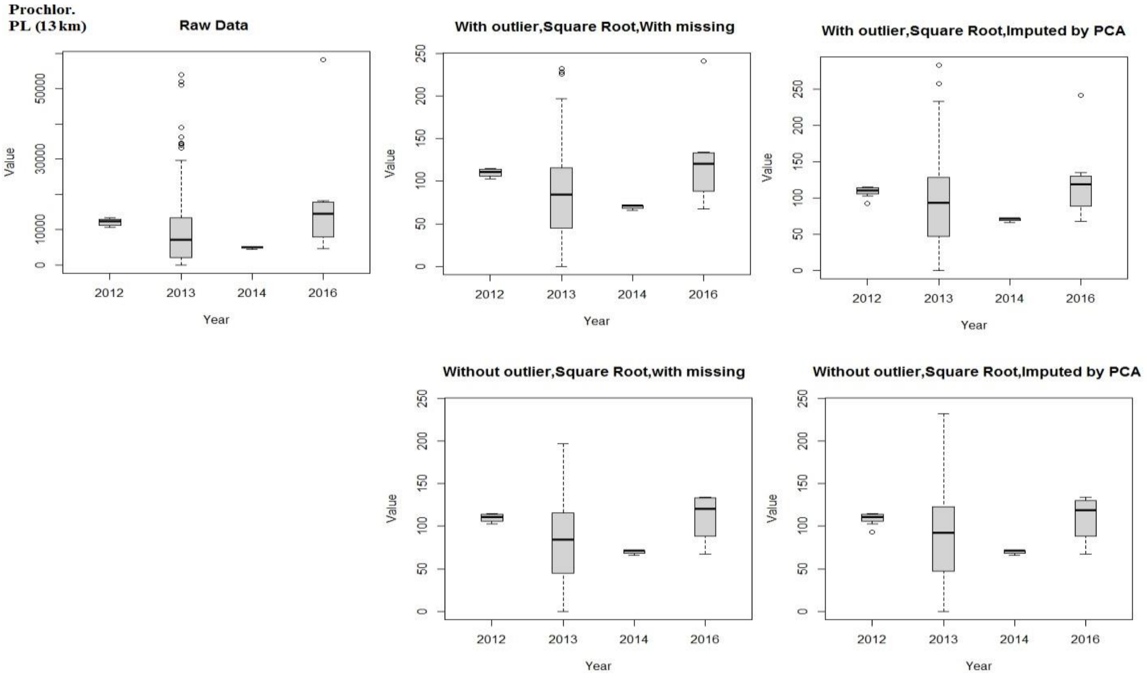
Fig. 44 shows the distribution of Prochlor in different years and sites. The abundance of Prochlor in AR and 2013 is higher than 2016 (Fig. 44a).The abundance of Prochlor in MV and 2011 is higher than 2012 and 2014 (Fig. 44b). The abundance of Prochlor in PAB and 2014 is higher than 2012 and 2016 (Fig. 44c). The abundance of Prochlor in PL and 2013 is higher than 2012 and 2014 and 2016 (Fig. 44d). The abundance of Prochlor in SG and 2014 is higher than 2012 and 2016 (Fig. 44e). The abundance of Prochlor in TIM and 2013 is higher than 2012 and 2014 (Fig. 44f).



c


Prochlor.
PAB (89km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2014
2016
0
2000
6000
10000
14000
0
20
40
60
80
100
120
140
150
50

d


Prochlor.
PL (13km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2013
2014
2016
0
10000
30000
50000
50
100
150
200
250

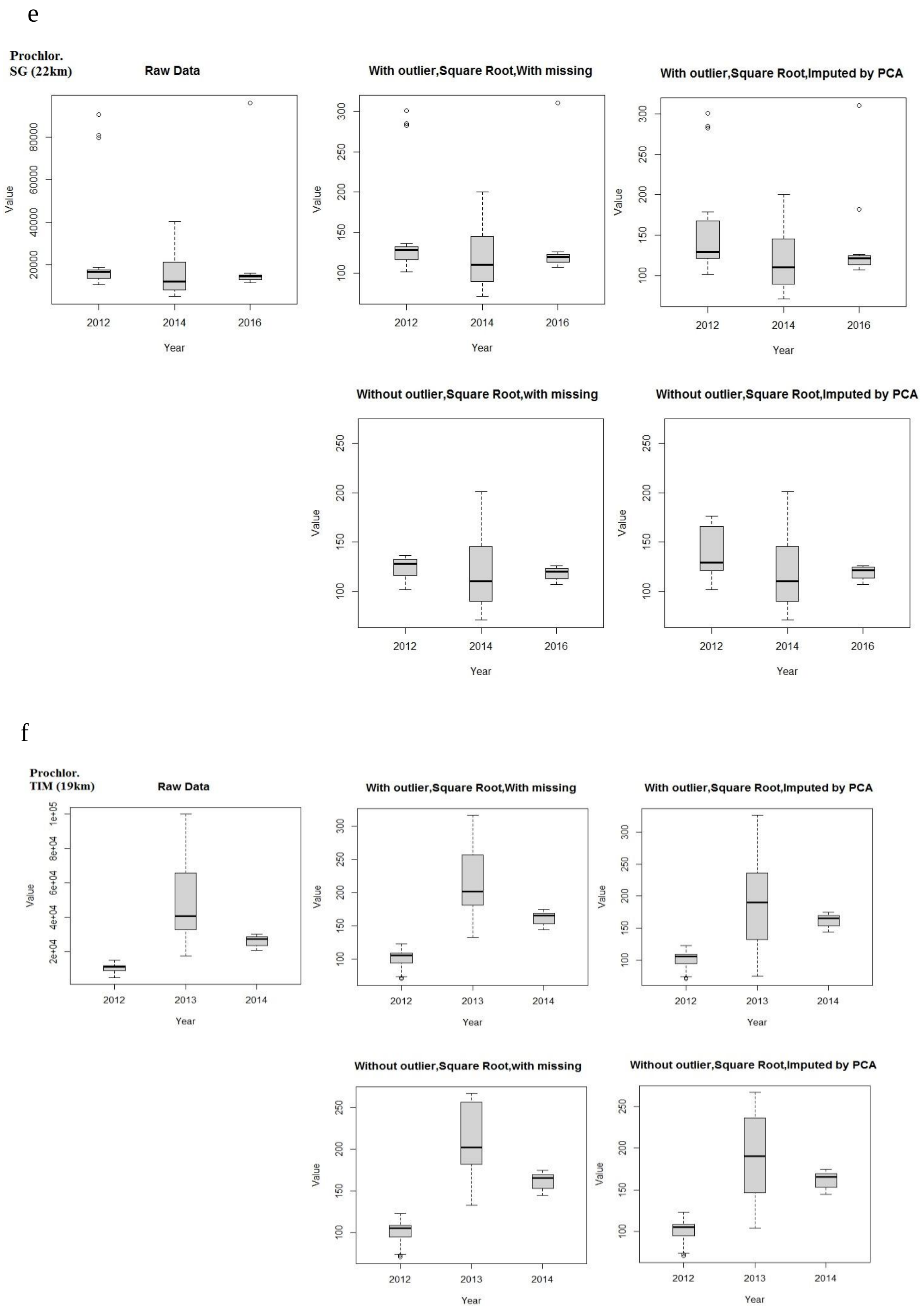
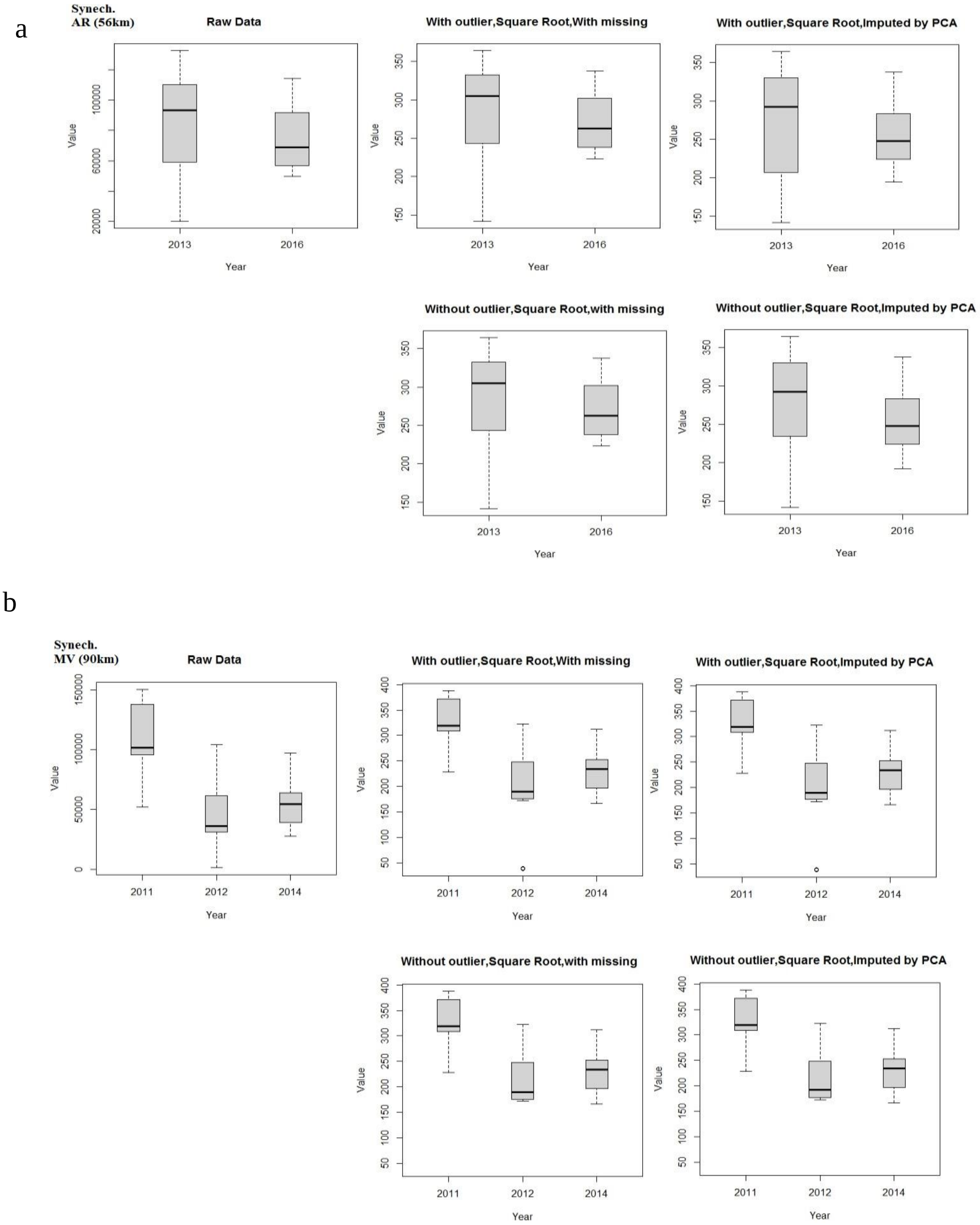
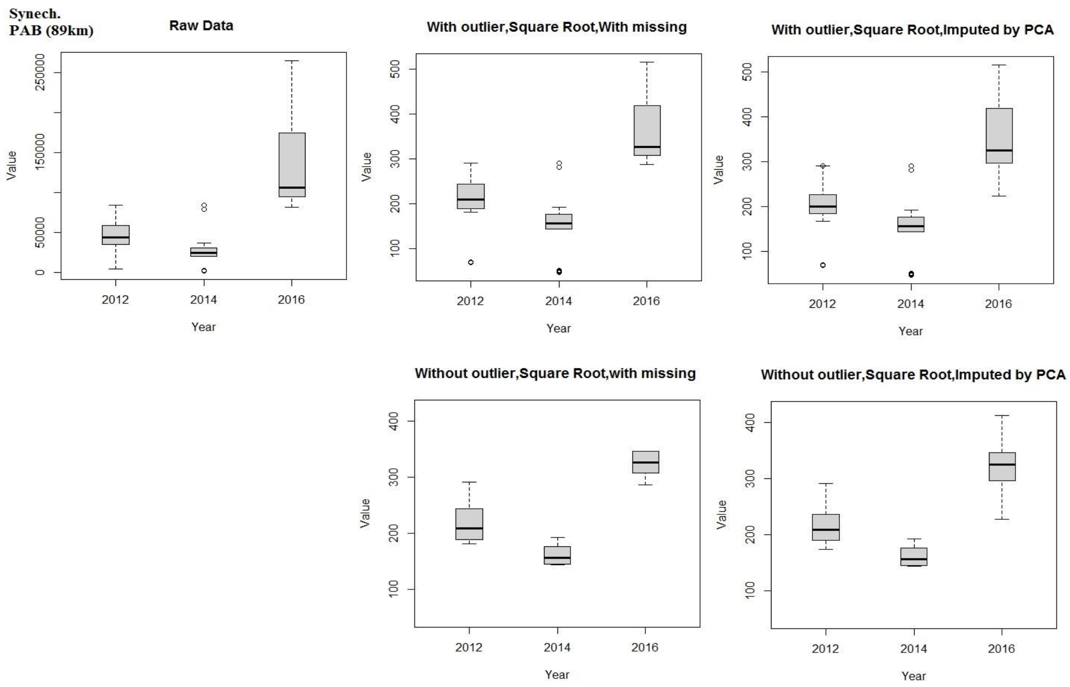


Fig. 44 Distribution of Prochlor (19.21 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

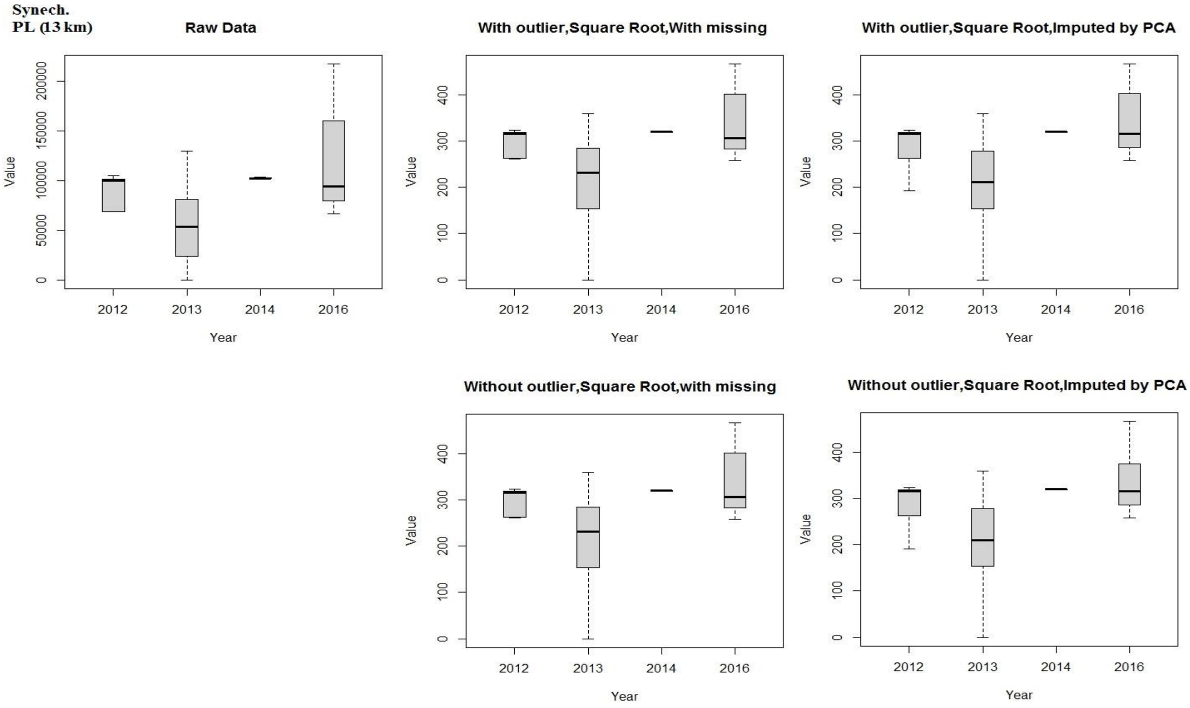
Finally, Fig. 45 shows the distribution of Synech in different years and sites. The abundance of Synech in AR and 2013 is higher than 2016 (Fig. 45a).The abundance of Synech in MV and 2012 is higher than 2011 and 2014 (Fig. 45b). The abundance of Synech in PAB and2016 is higher than 2012 and 2014 (Fig. 45c). The abundance of Synech in PL and 2013is higher than 2012 and 2014 and 2016 (Fig. 45d). The abundance of Synech in SG and 2012 is higher than 2014 and 2016 (Fig. 45e). The abundance of Synech in TIM and 2013is higher than 2012 and 2014 (Fig. 45f).



c


Synech.
PAB (89km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2014
2016
0
50000
150000
250000
100
200
300
400
500

d


Synech.
PL (13 km)
Raw Data
With outlier,Square Root,With missing
With outlier,Square Root,Imputed by PCA
Without outlier,Square Root,with missing
Without outlier,Square Root,Imputed by PCA
Value
Year
2012
2013
2014
2016
0
50000
100000
150000
200000
100
200
300
400

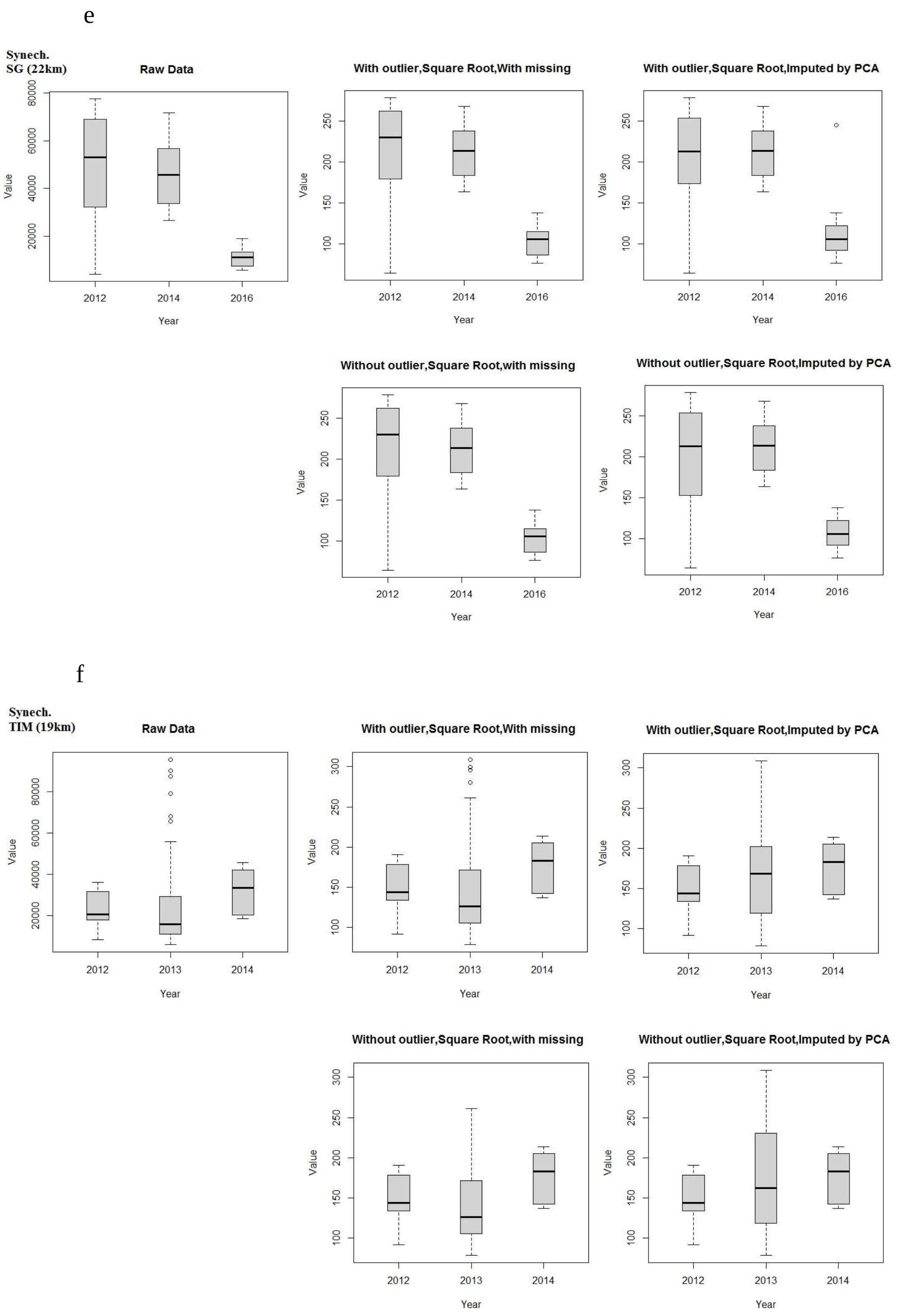


Fig. 45 Distribution of Synech (15.81 % missing) in different years and in a) AR (56km), b) MV (90km), c) PAB (89km), d) PL (13km), e) SG (22km) and f) TIM (19km).

3.3. Modelling results

In this Section we discuss the modelling performance, the relative importance of environmental variables and the difference of biophysical and biochemical parametrs and microbial abundance in the internal and external arcs of Abrolhos.

3.3.1. Model performance

Table 5 shows the predictive performance of the BRT and RF models based on the 10-fold cross-validation, including the MAE, RMSE and R^2 values. The results suggest that model RF was highly predictive based on the range of R^2 values from 0.66 to 0.84 in compare to BRT with R^2 values from 0.55 to 0.78 (Table 5). In general, the descriptive statistics indicate approximately similar levels of prediction accuracy of the BRT and RF models, based on the R^2 (0.65 vs. 0.75, respectively).

The coefficient of determination (R^2) is a statistical measure of how close the data are to the fitted regression line. It is also known as the R-squared. RMSE and MAE have the same unit as the dependent variables (DV). It means that there is no absolute good or bad threshold and can be defined based on DV. The R^2 values suggest that both models can explain approximately 70% of the total microbial abundance variability and both approaches accurately predicted the microbial abundance, based on R^2 values of approximately 0.7.

Table 5. Summary statistics of the predictive quality of boosted regression trees (BRT) and random forest (RF) models for microbial abundance with 100 runs; the mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2) are used to evaluate accuracy.

Model	Index	Bacterial	Cl-a	Feof	HNA	LNA	Nanouk.	Picoeuk.	Prochlor.	Synech.	Virus
RF	RMSE	73	0.13	0.10	56	47	14	12	37	46	305
	MAE	48	0.07	0.06	41	35	7.5	8.6	24.6	32	230
	R^2	**0.80**	**0.78**	**0.75**	**0.81**	**0.84**	**0.73**	**0.73**	**0.66**	**0.70**	**0.78**
BRT	RMSE	71	0.13	0.12	66	51	16	14	37	54	345
	MAE	51	0.08	0.07	46	37	9.33	10	26	38	262
	R^2	**0.77**	**0.67**	**0.63**	**0.65**	**0.78**	**0.55**	**0.57**	**0.58**	**0.60**	**0.63**

3.3.2. Relative importance of environmental variables

The relative importance of each predictor, which has been determined from 100 runs of the BRT and RF models, shown in Figs. 46-55. We normalized the importance of variables in the RF model to 100% to provide a simple basis for comparison with the BRT model. The relative importance of predictors almost differed between the two models, but

DOC was the most influential factor in the BRT and RF models (Figs. 46-55). In both, the three most important predictors, based on their mean values, were DOC, TN, and silicate (Figs. 46-55). The Heatmap of the relative importance of each predictor, in both BRT and RF models,is presented in Fig. 56.

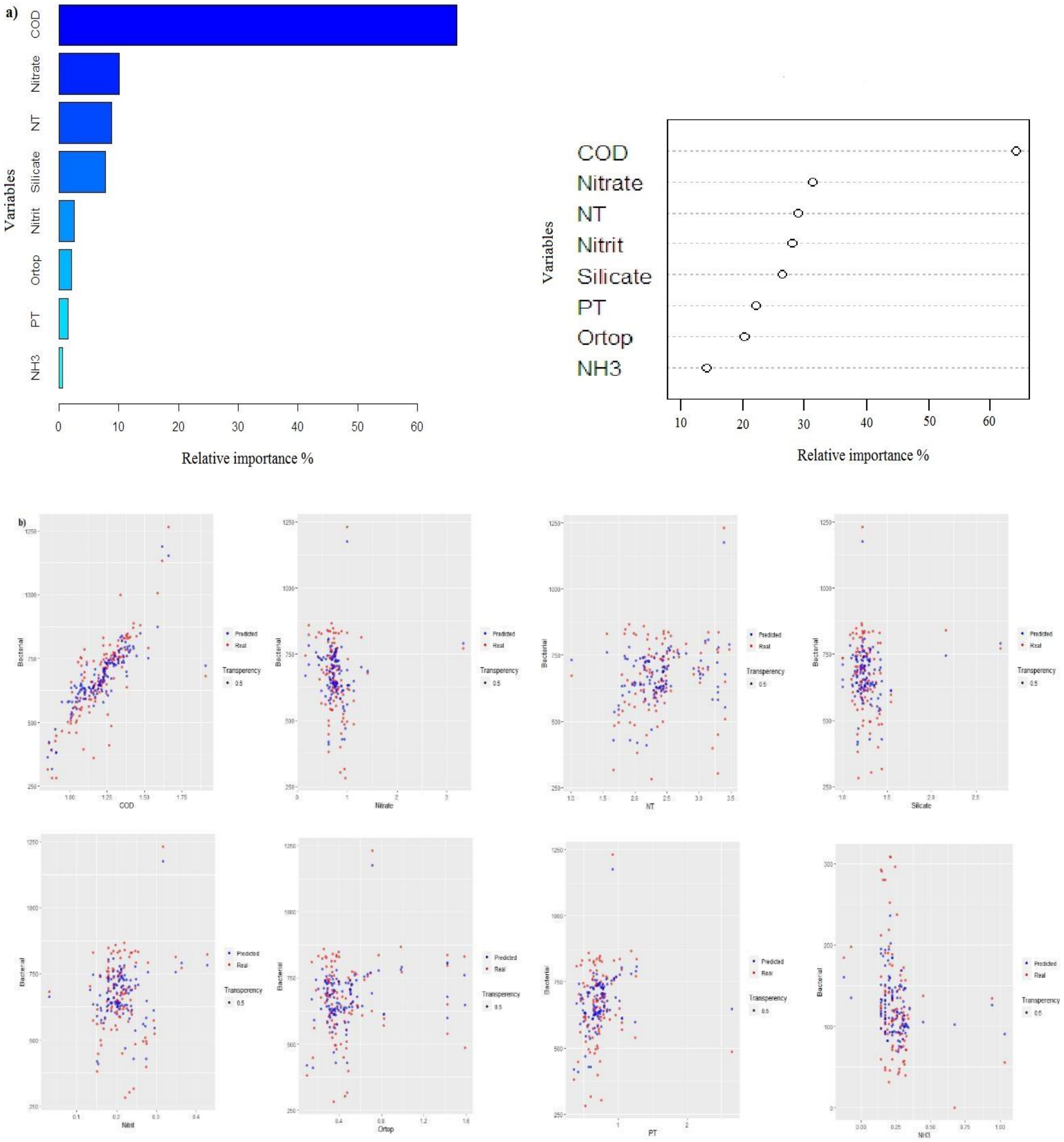


Fig. 46. a) Relative importance of each variable on bacterial abundance as determined from 100 runs of the boosted regression trees (BRT, left) and random forest (RF, right) models, which are shown in decreasing order and normalized to 100%. b) Relationships between predicted and real value s of bacterial abundance and all predictors.

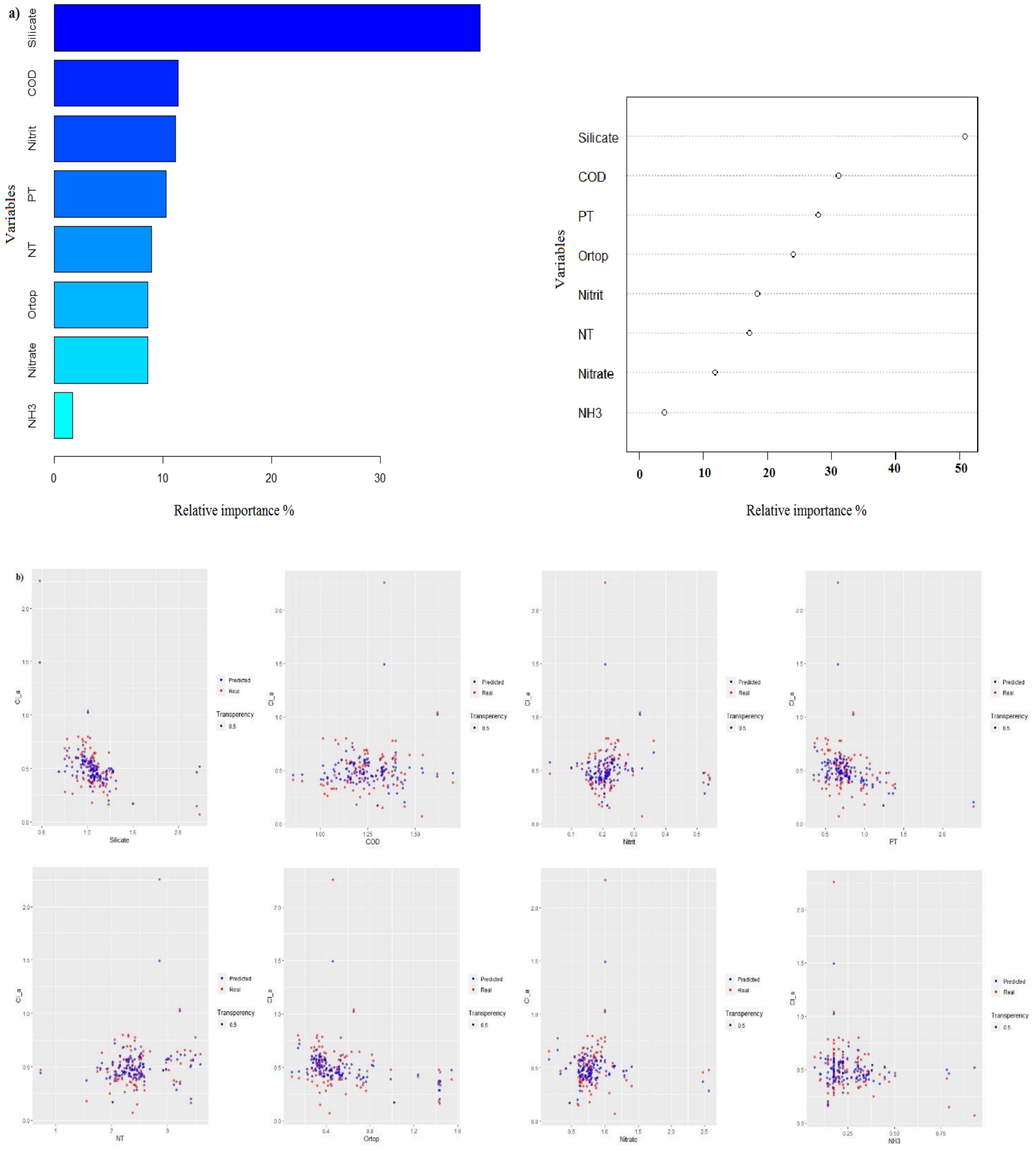


Fig. 47. a) Relative importance of each variable on Cl-a abundance as determined from 100 runs of the boosted regression trees (BRT, left) and random forest (RF, right) models, which are shown in decreasing order and normalized to 100%. b) Relationships between predicted and real value s of Cl-a abundance and all predictors.

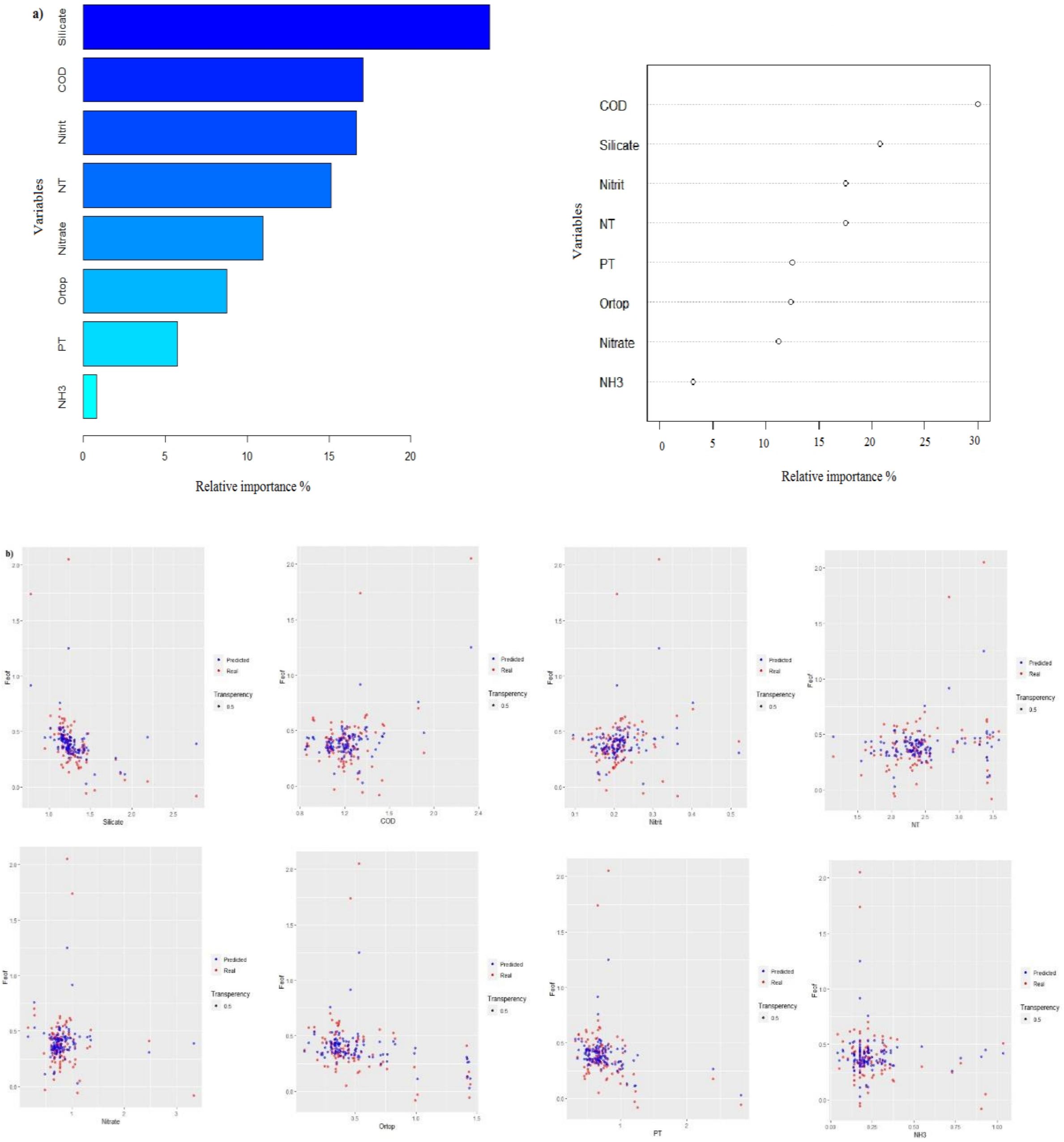


Fig. 48. a) Relative importance of each variable on Feof abundance as determined from 100 runs of the boosted regression trees (BRT, left) and random forest (RF, right) models, which are shown in decreasing order and normalized to 100%. b) Relationships between predicted and real value s of Feof abundance and all predictors.

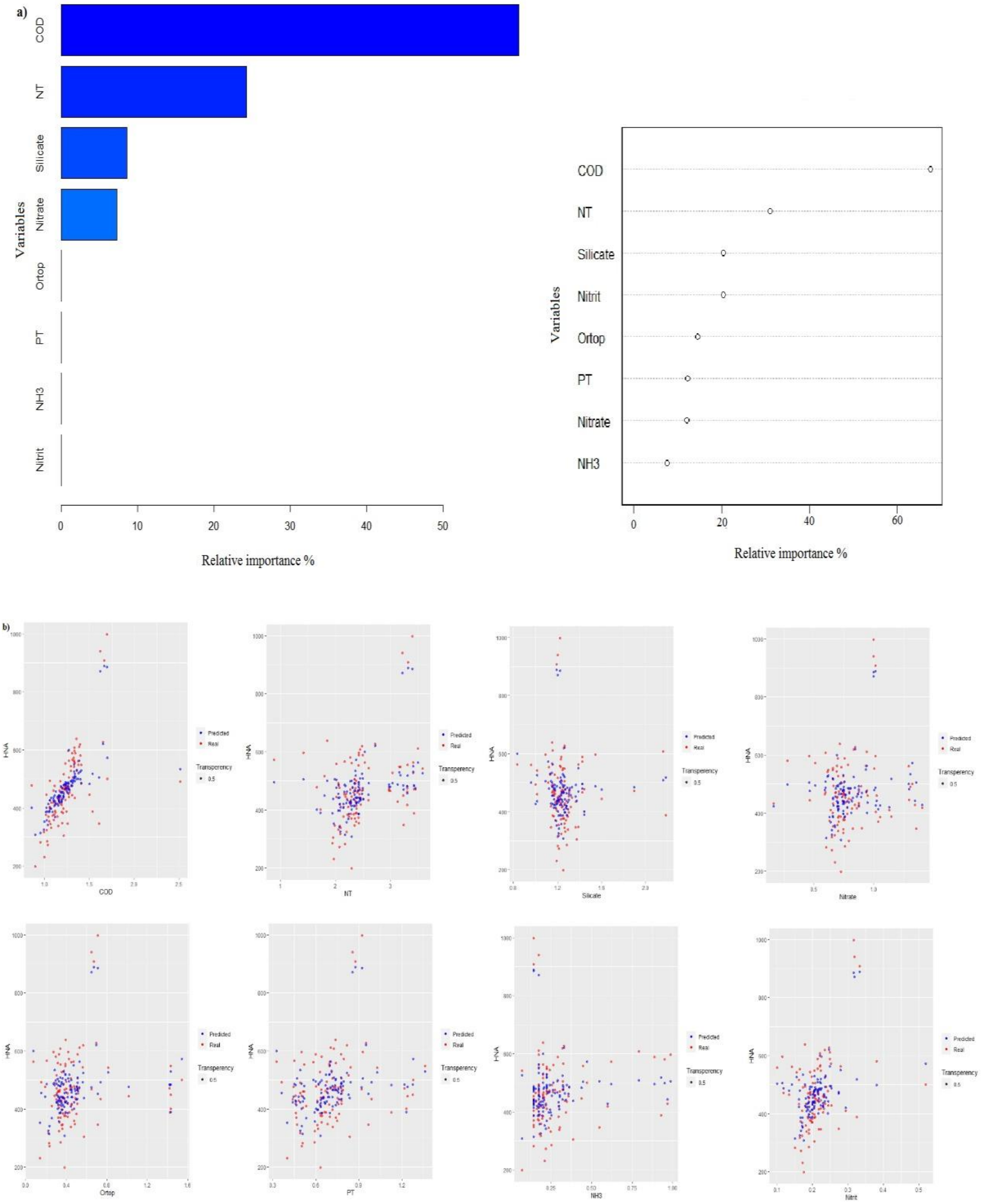


Fig. 49. a) Relative importance of each variable on HNA as determined from 100 runs of the boosted regression trees (BRT, left) and random forest (RF, right) models, which are shown in decreasing order and normalized to 100%. b) Relationships between predicted and real value s of HNA and all predictors.

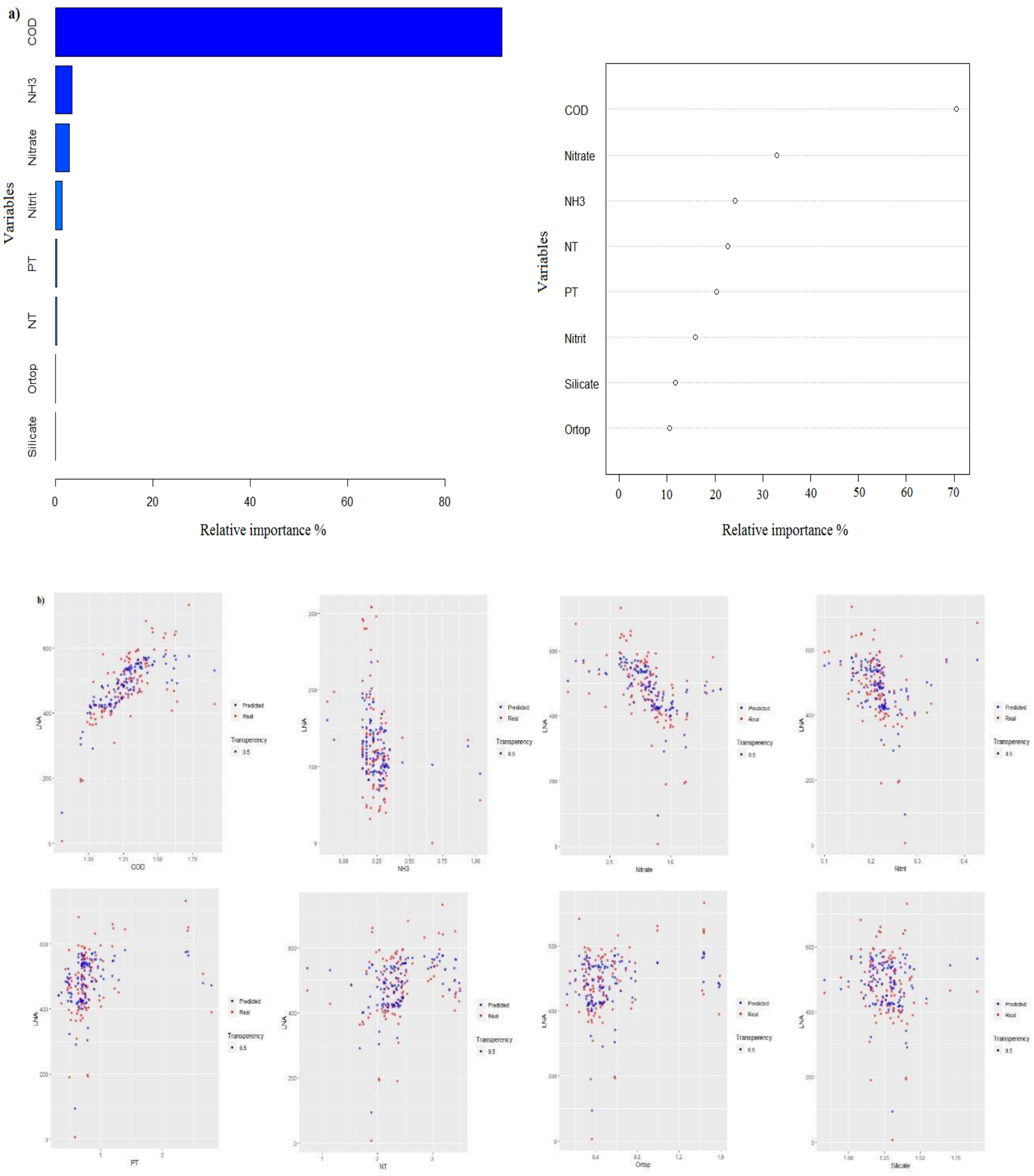


Fig. 50. a) Relative importance of each variable on LNA as determined from 100 runs of the boosted regression trees (BRT, left) and random forest (RF, right) models, which are shown in decreasing order and normalized to 100%. b) Relationships between predicted and real value s of LNA and all predictors.

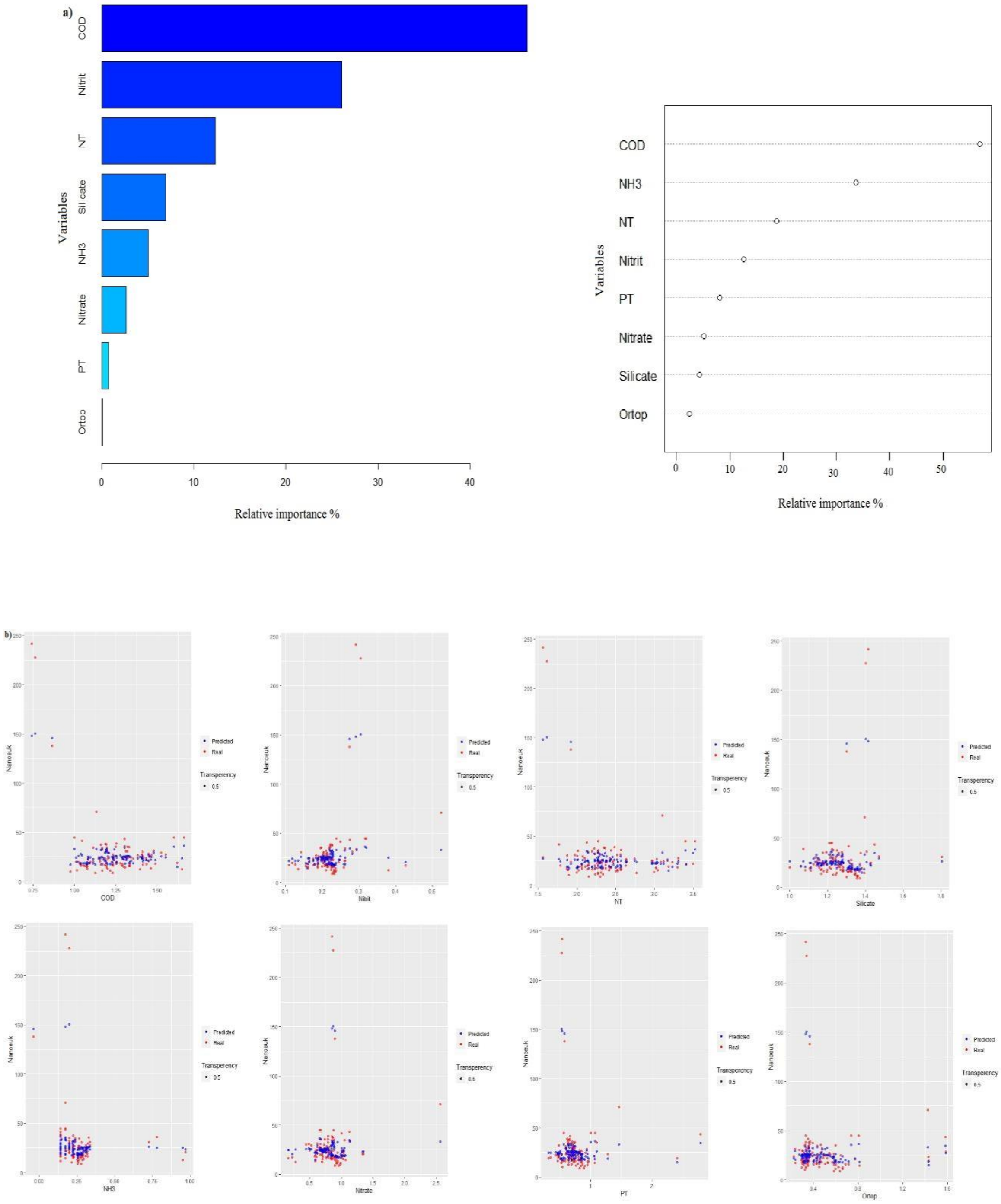


Fig. 51. a) Relative importance of each variable on Nanoeuk abundance as determined from 100 runs of the boosted regression trees (BRT, left) and random forest (RF, right) models, which are shown in decreasing order and normalized to 100%. b) Relationships between predicted and real value s of Nanoeukabundance and all predictors.

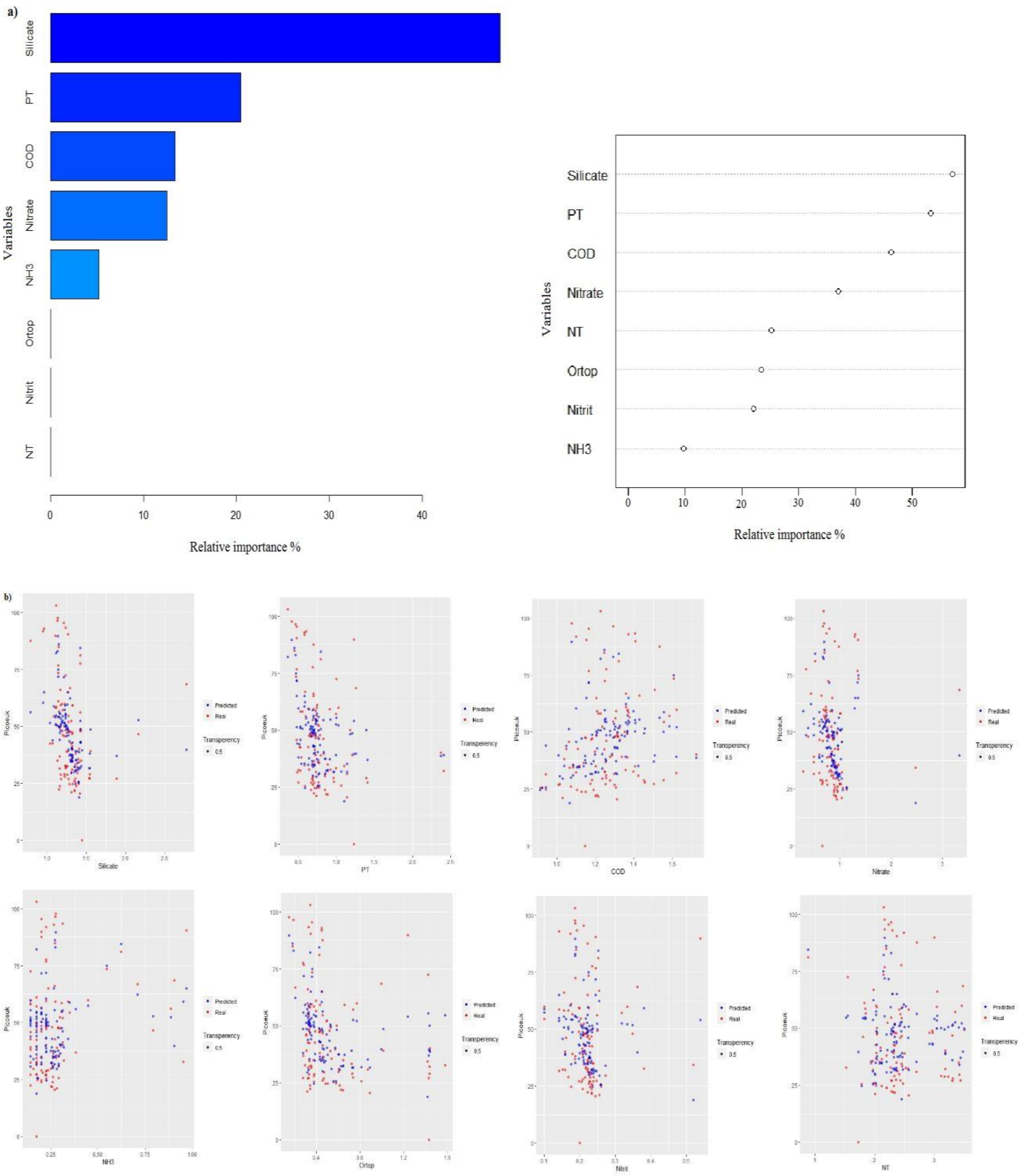


Fig. 52. a) Relative importance of each variable on Picoeuk abundance as determined from 100 runs of the boosted regression trees (BRT, left) and random forest (RF, right) models, which are shown in decreasing order and normalized to 100%. b) Relationships between predicted and real value s of Picoeuk abundance and all predictors.

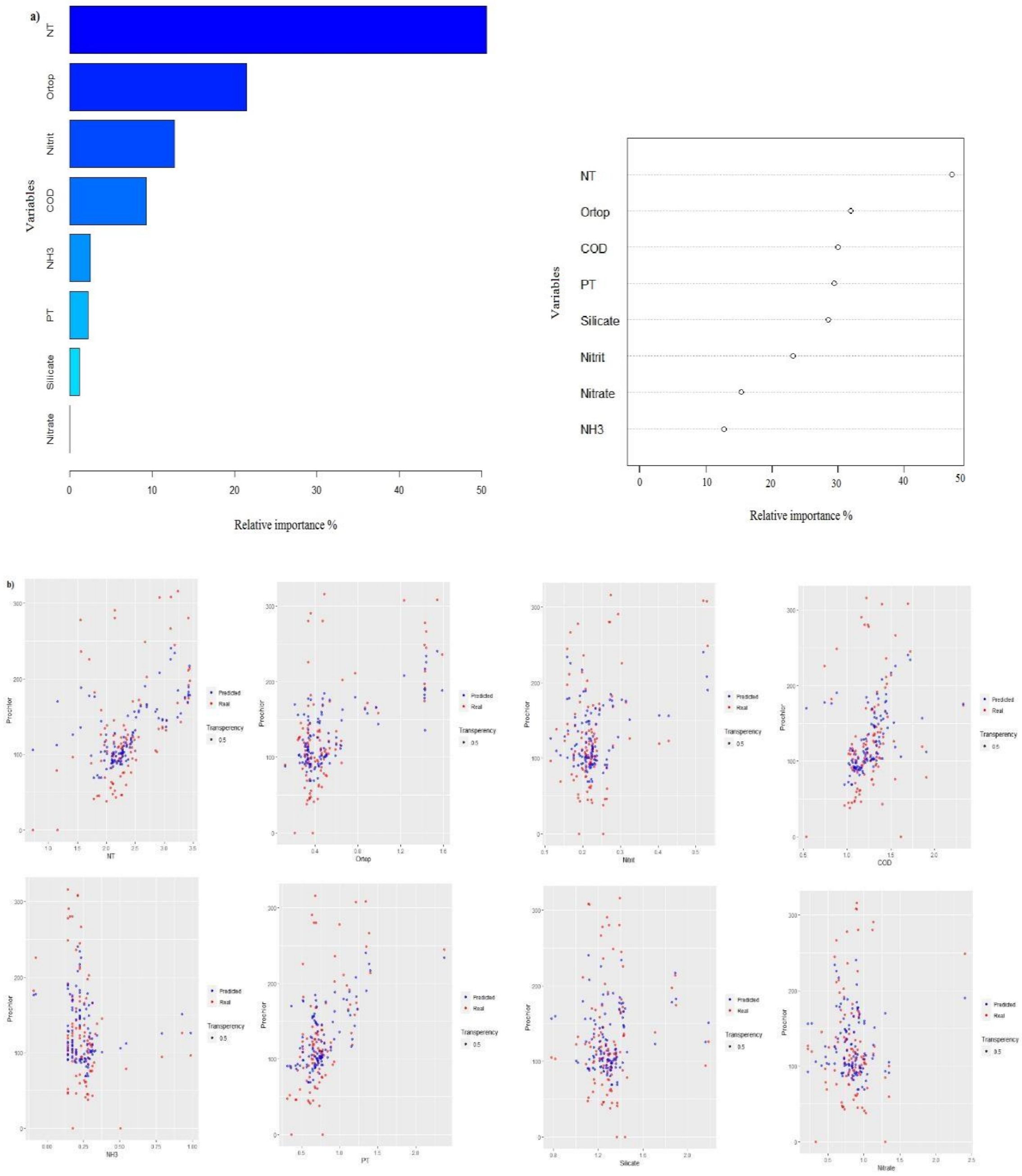


Fig. 53. a) Relative importance of each variable on Prochlor abundance as determined from 100 runs of the boosted regression trees (BRT, left) and random forest (RF, right) models, which are shown in decreasing order and normalized to 100%. b) Relationships between predicted and real value s of Prochlorabundance and all predictors.

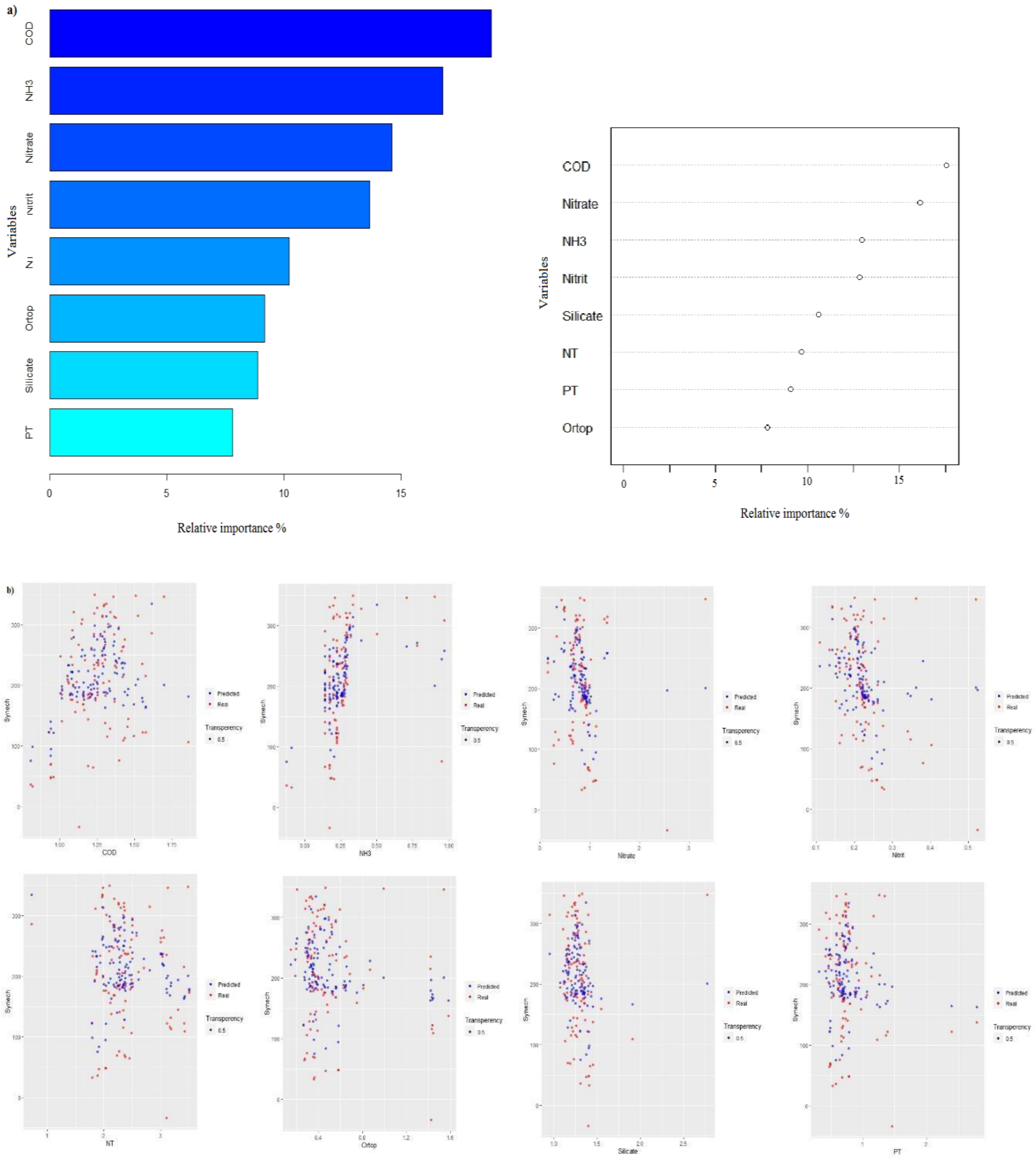


Fig. 54. a) Relative importance of each variable on Synech abundance as determined from 100 runs of the boosted regression trees (BRT, left) and random forest (RF, right) models, which are shown in decreasing order and normalized to 100%. b) Relationships between predicted and real value s of Synech abundance and all predictors.

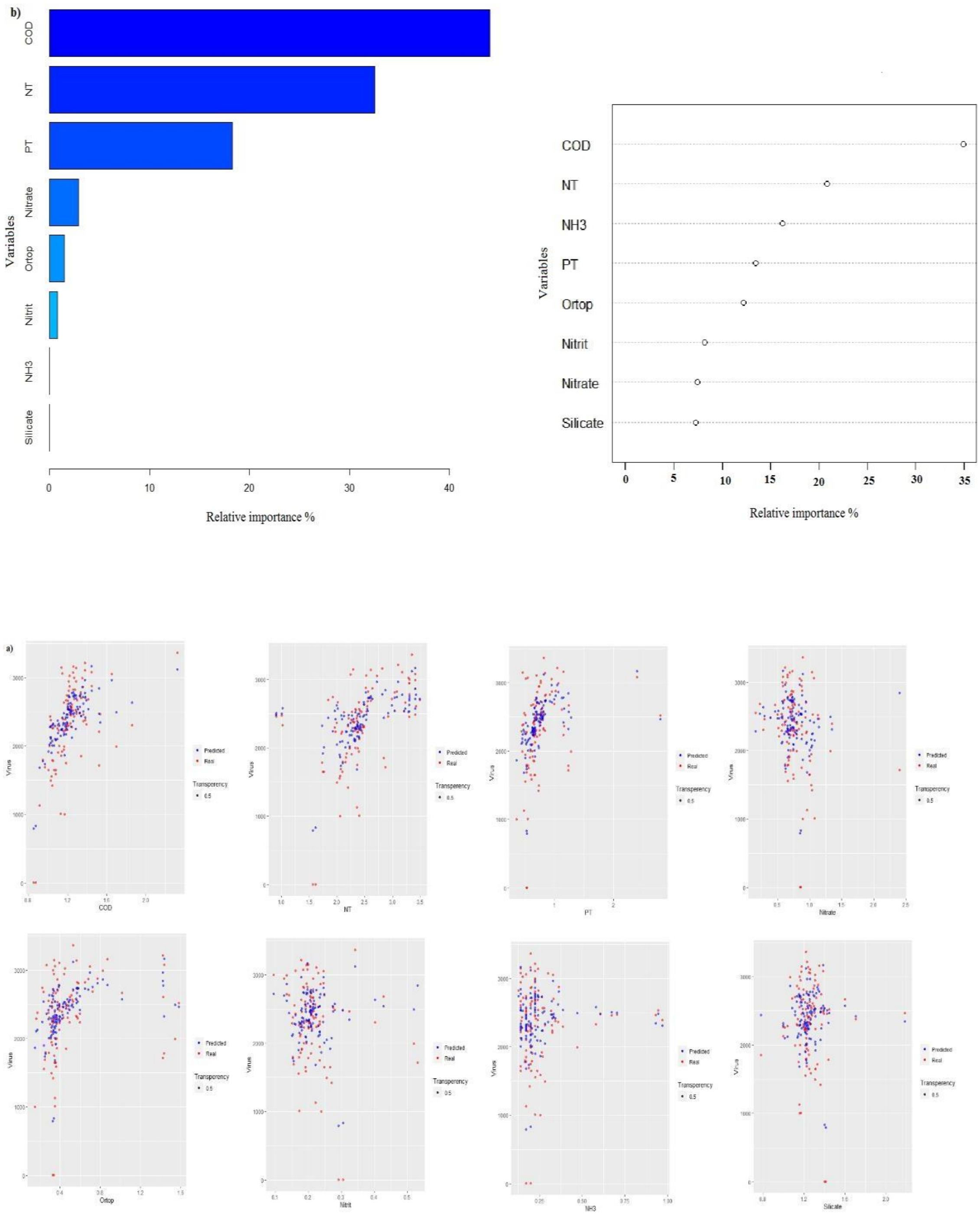


Fig. 55. a) Relative importance of each variable on Virus abundance as determined from 100 runs of the boosted regression trees (BRT, left) and random forest (RF, right) models, which are shown in decreasing order and normalized to 100%. b) Relationships between predicted and real value s of Virus abundance and all predictors.

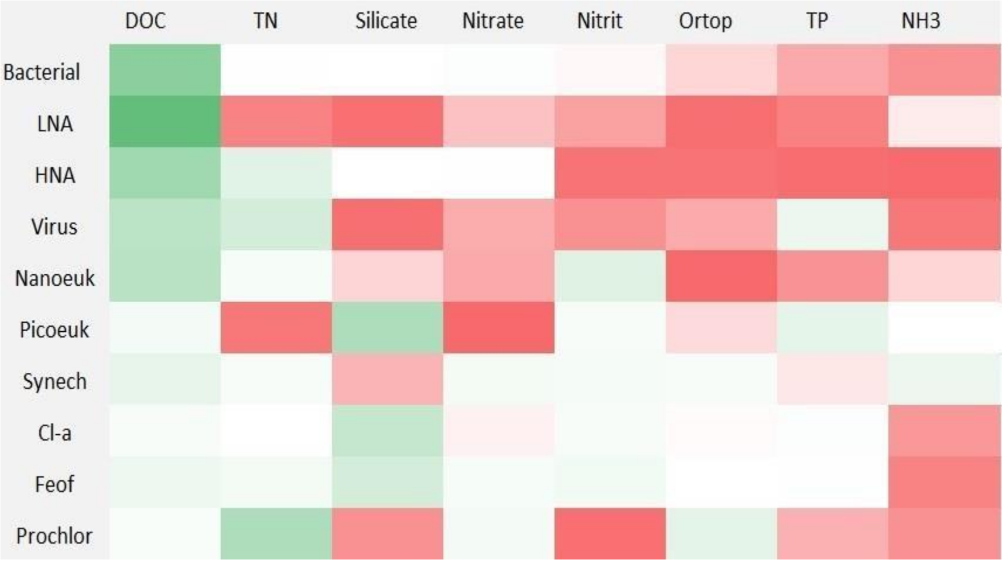


Fig. 56. The Heatmap of the relative importance of each predictor, in both BRT and RF models. The green and red colors represent the highest and lowest importance of each independent variable, respectively. The white color represent the midpoint of values.

3.3.3. The Difference of Biophysical and Biochemical Parameters and Microbial Abundance in Internal and External ArcS of Abrolhos

The water temperature showed an annual profile of variation. The difference of temperature values between Internal and External arcs and in years 2011 to 2016 is presented in Fig. 57. We can see that the external arc of Abrolhos (MV and PAB) presented the lowest temperature values along the years.

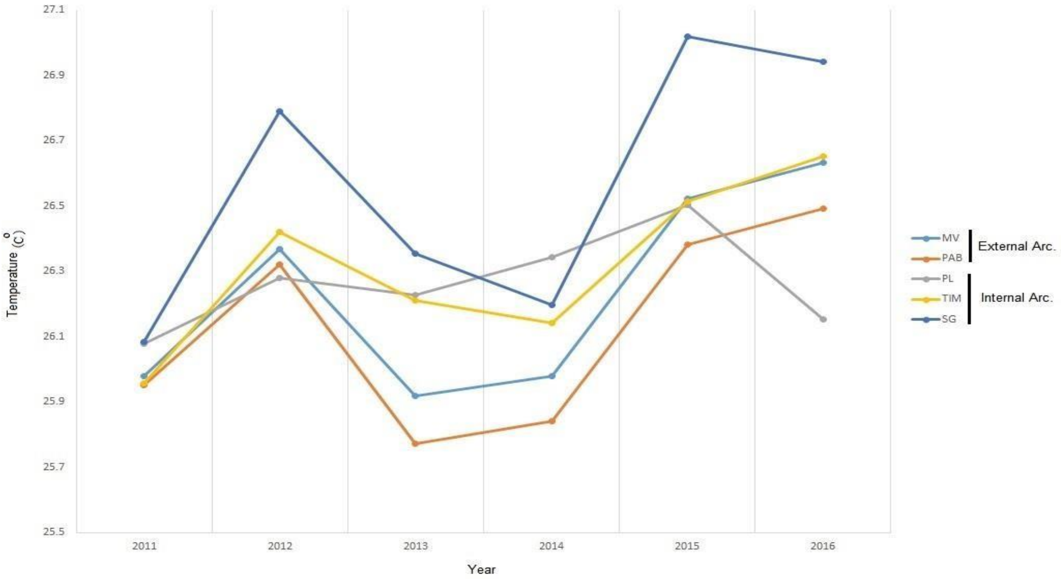


Fig. 57. Temperature variation years 2011 to 2016 – Internal and External arcs of Abrolhos.

The difference of hydrodynamic velocity between Internal and External arcs in 2010 is presented in Fig. 58. We can see that the AR external arc presents the highest hydrodynamic velocity, and the internal arc PL presents the lowest hydrodynamic velocity in the first, medium, and third quartile of data. There is no significant differences among other sites of inner and outer reefs.

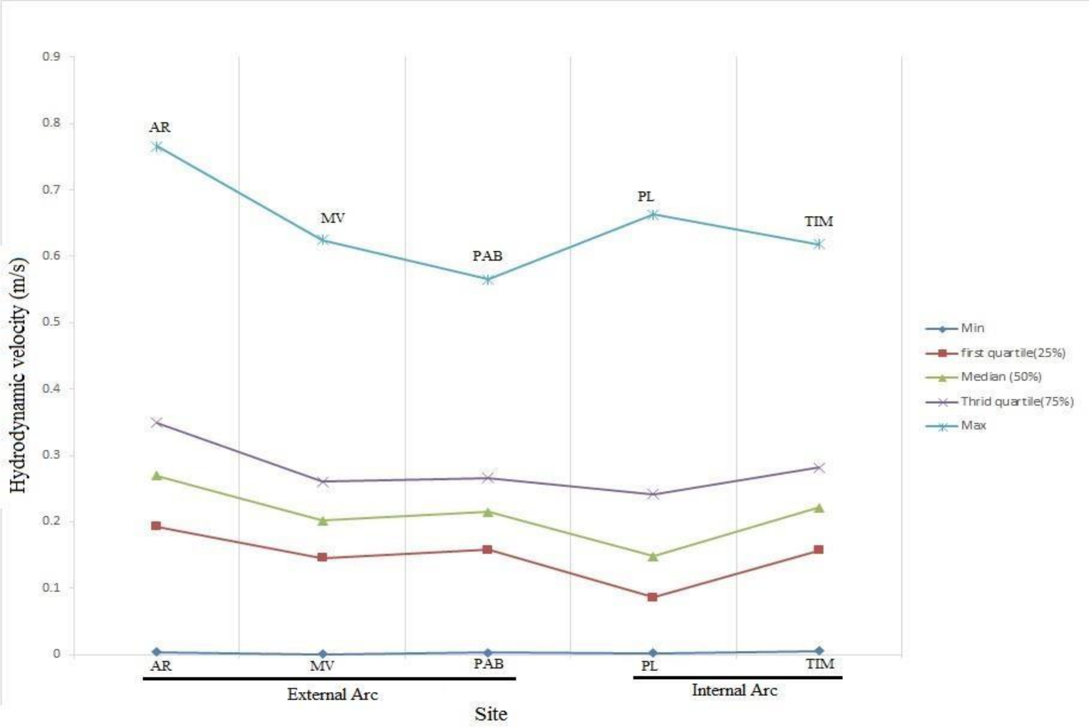


Fig. 58. The hydrodynamic velocity among Internal and External arcs in 2010.

The microbial abundance and biochemical parameters are presented in Figs. 59-76 for the six studied areas and grouped in the Internal and External arcs. In the Internal arcs (TIM, PL and SG), the total amount of microbial abundance and biochemical variables, such as nutrients, are higher than in the External arcs (AR, PAB and MV). The average value of each variable per site is also presented in Figs. 59-76.

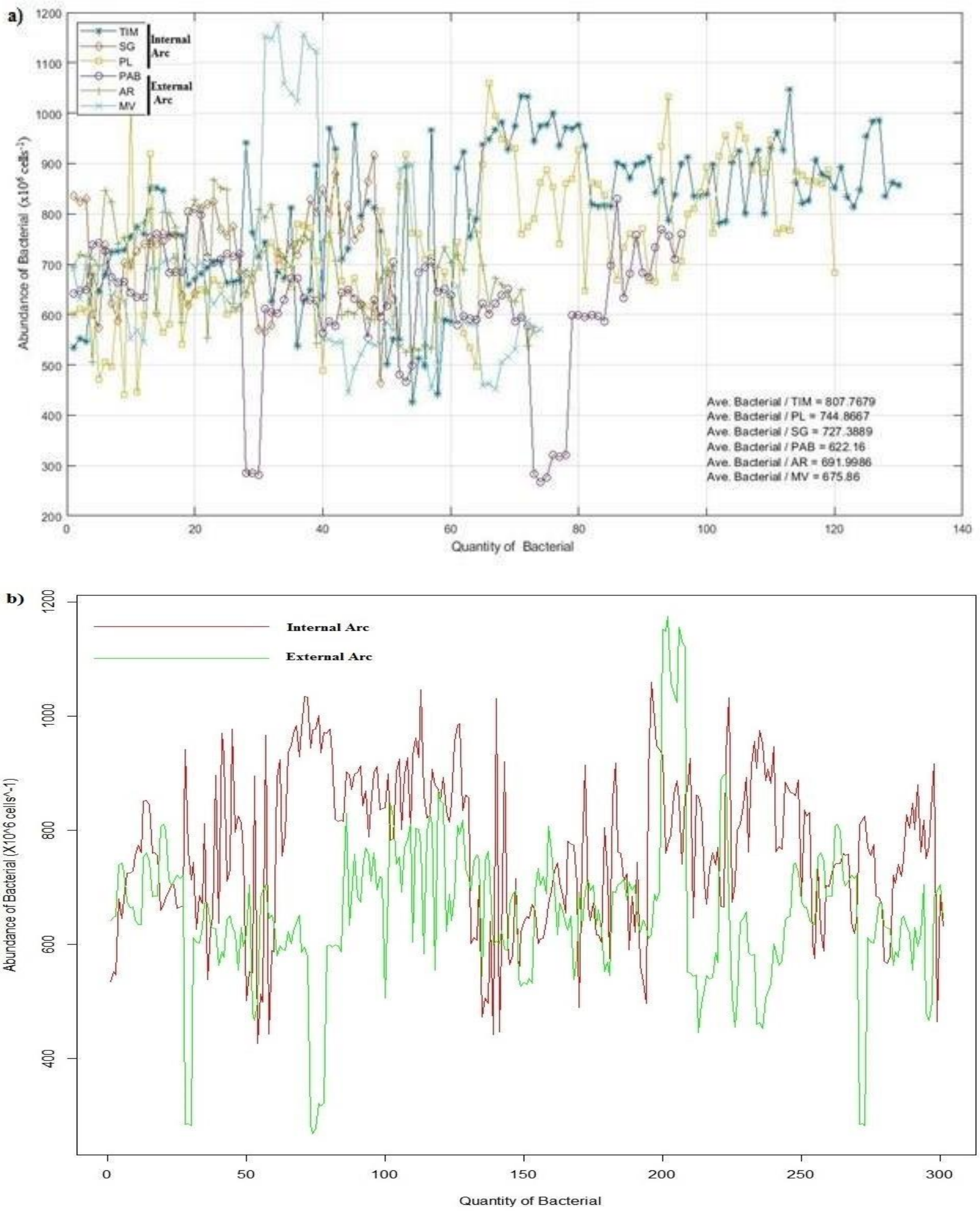


Fig. 59. Bacterial abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

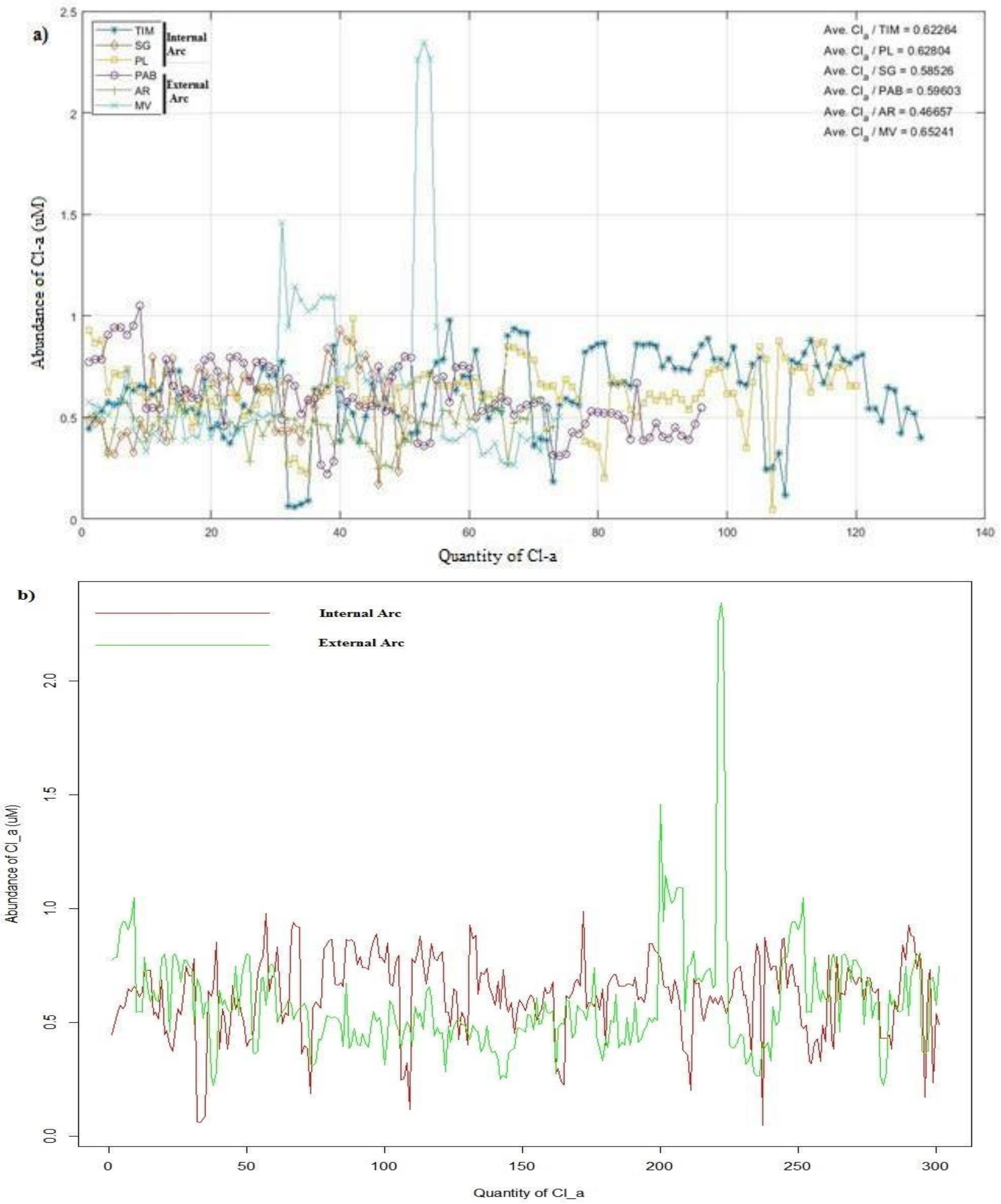


Fig. 60. Cl-a abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

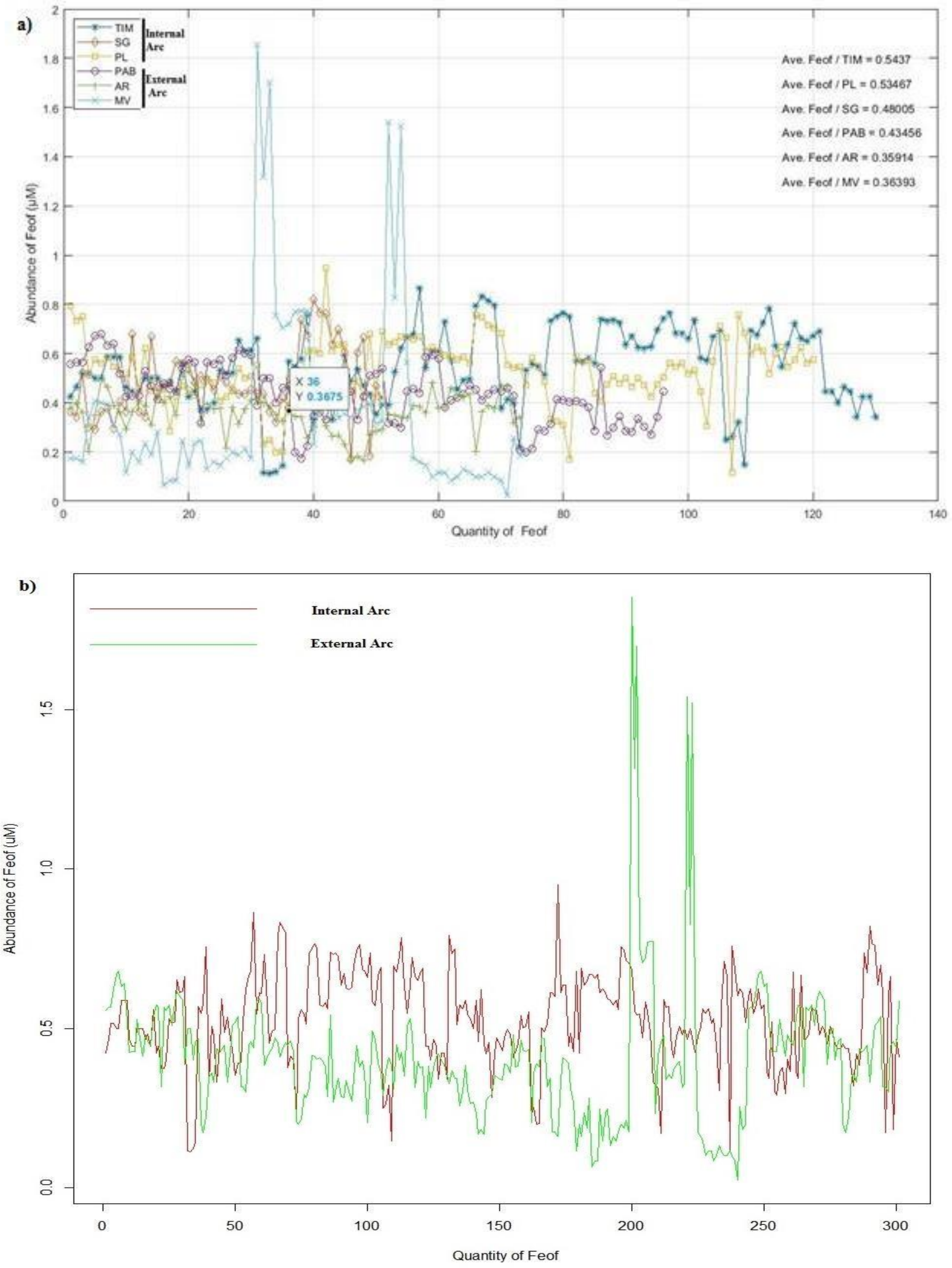


Fig. 61. Feof abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

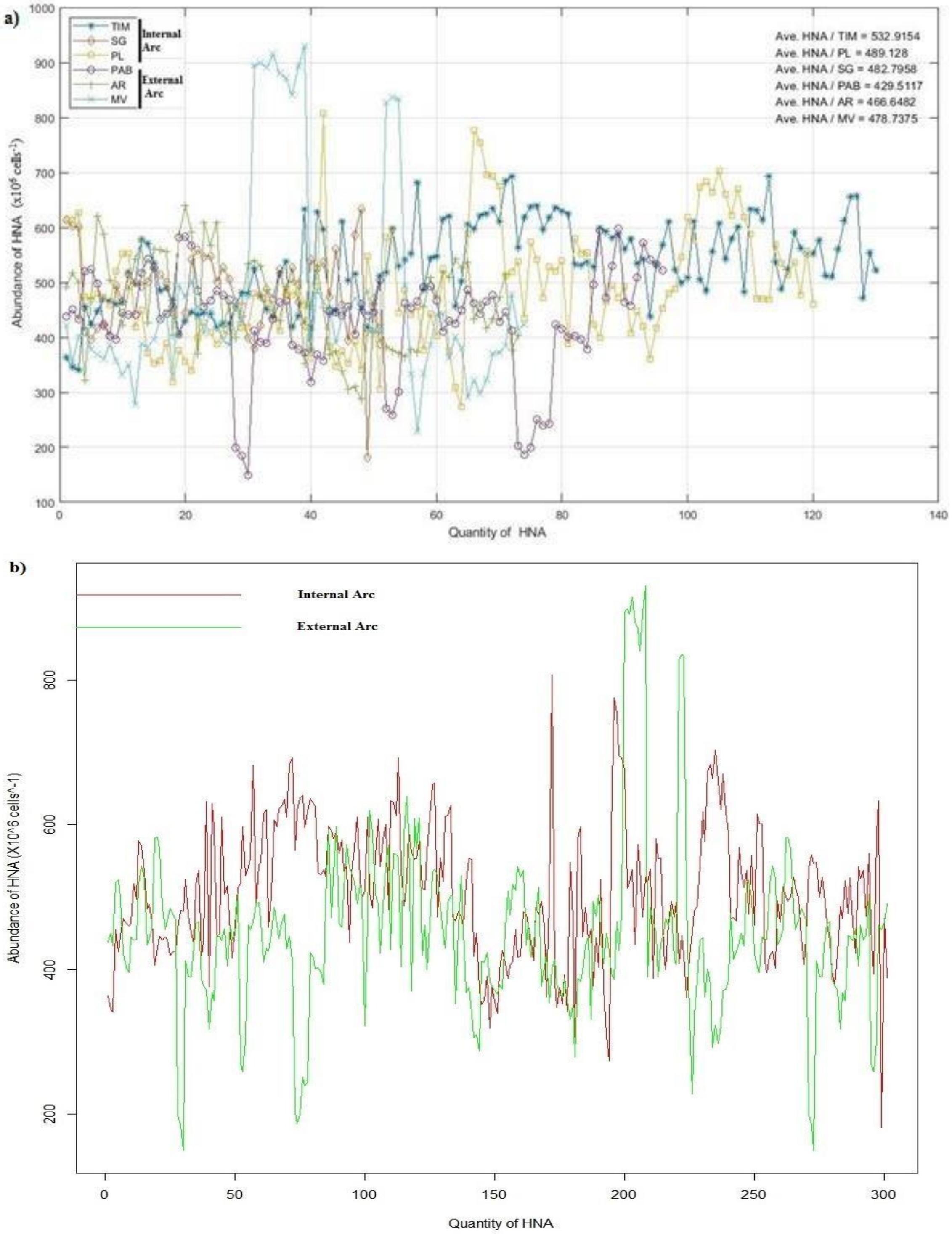


Fig. 62. HNA abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

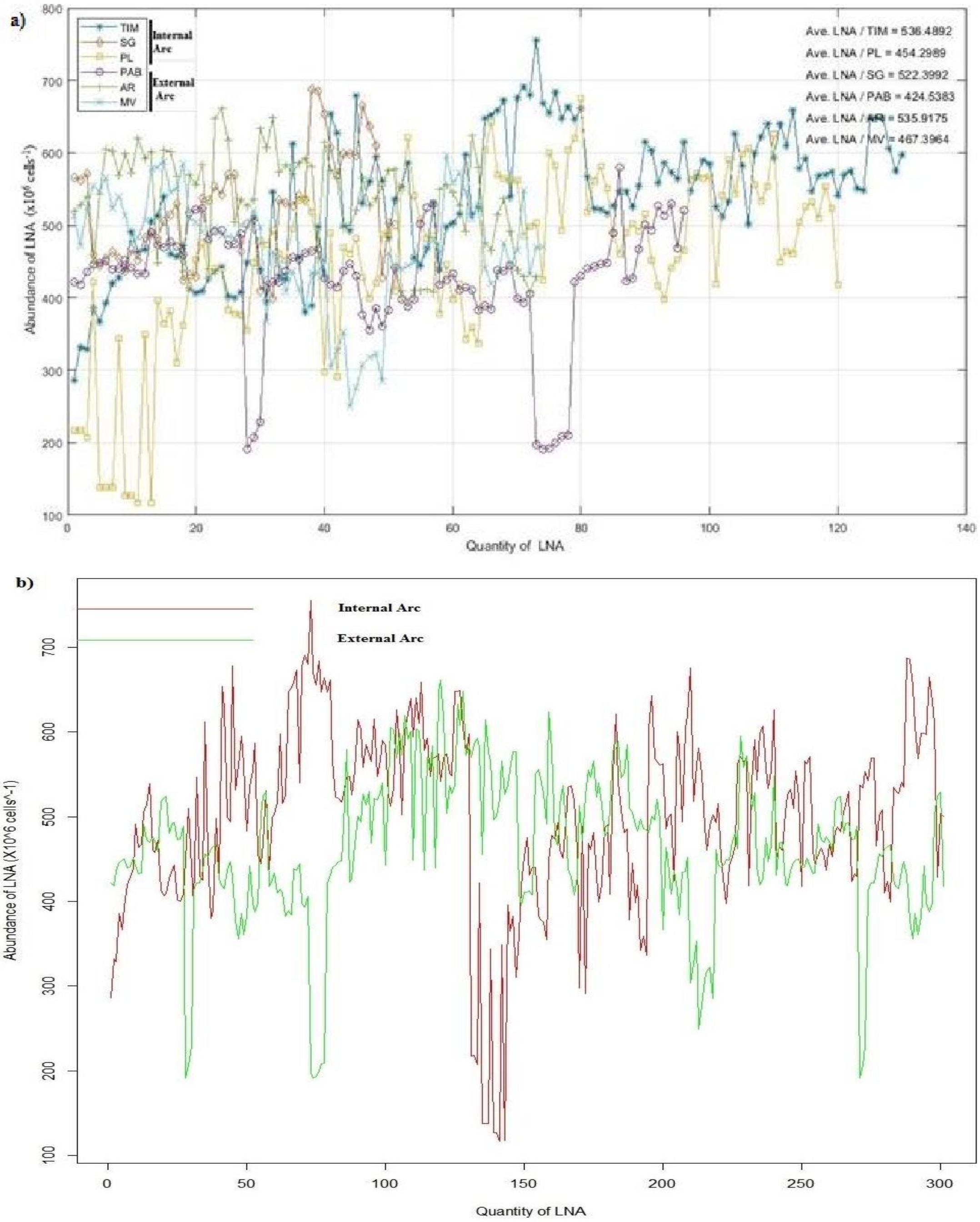


Fig. 63. LNA abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

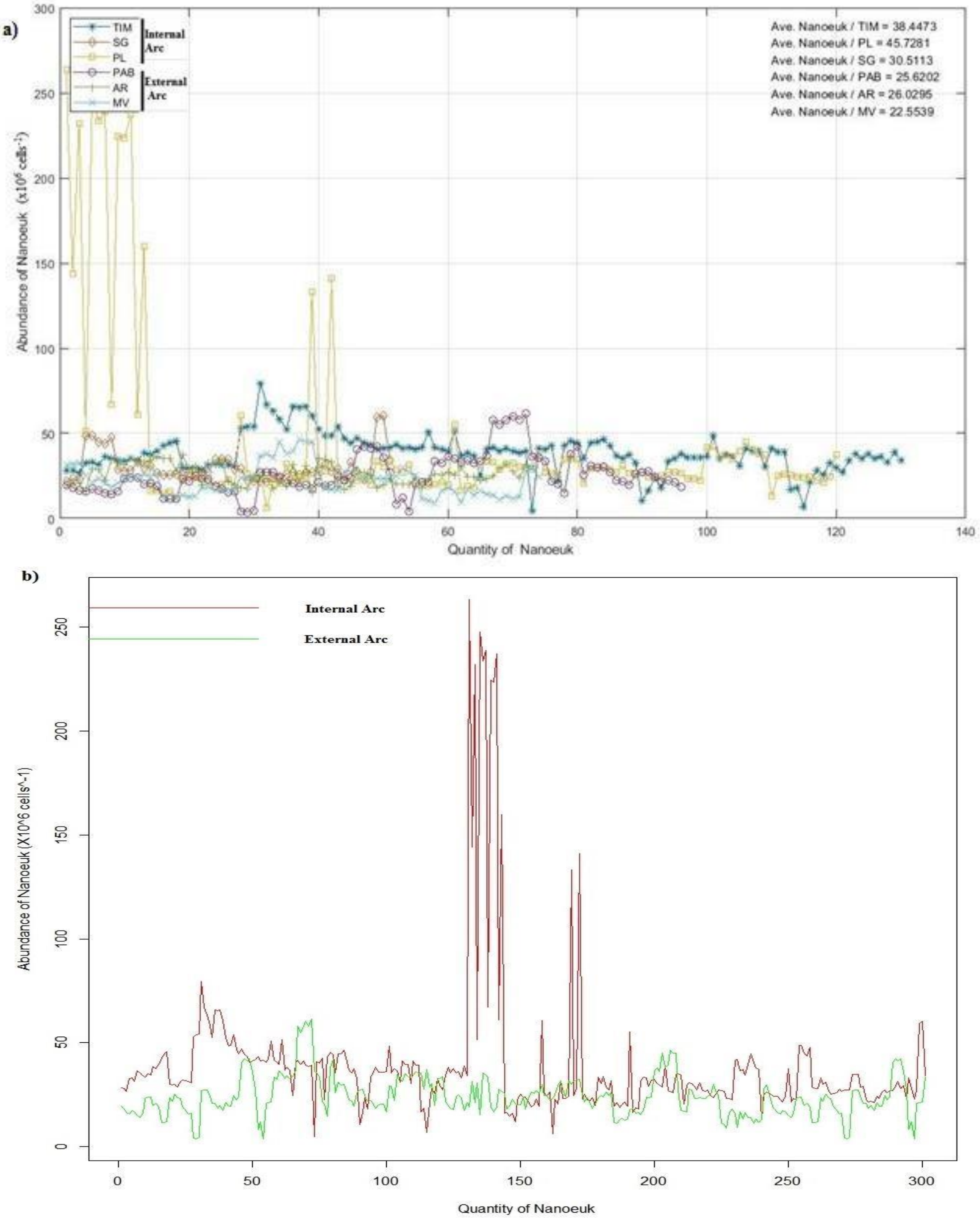


Fig. 64. Nanoeuk abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

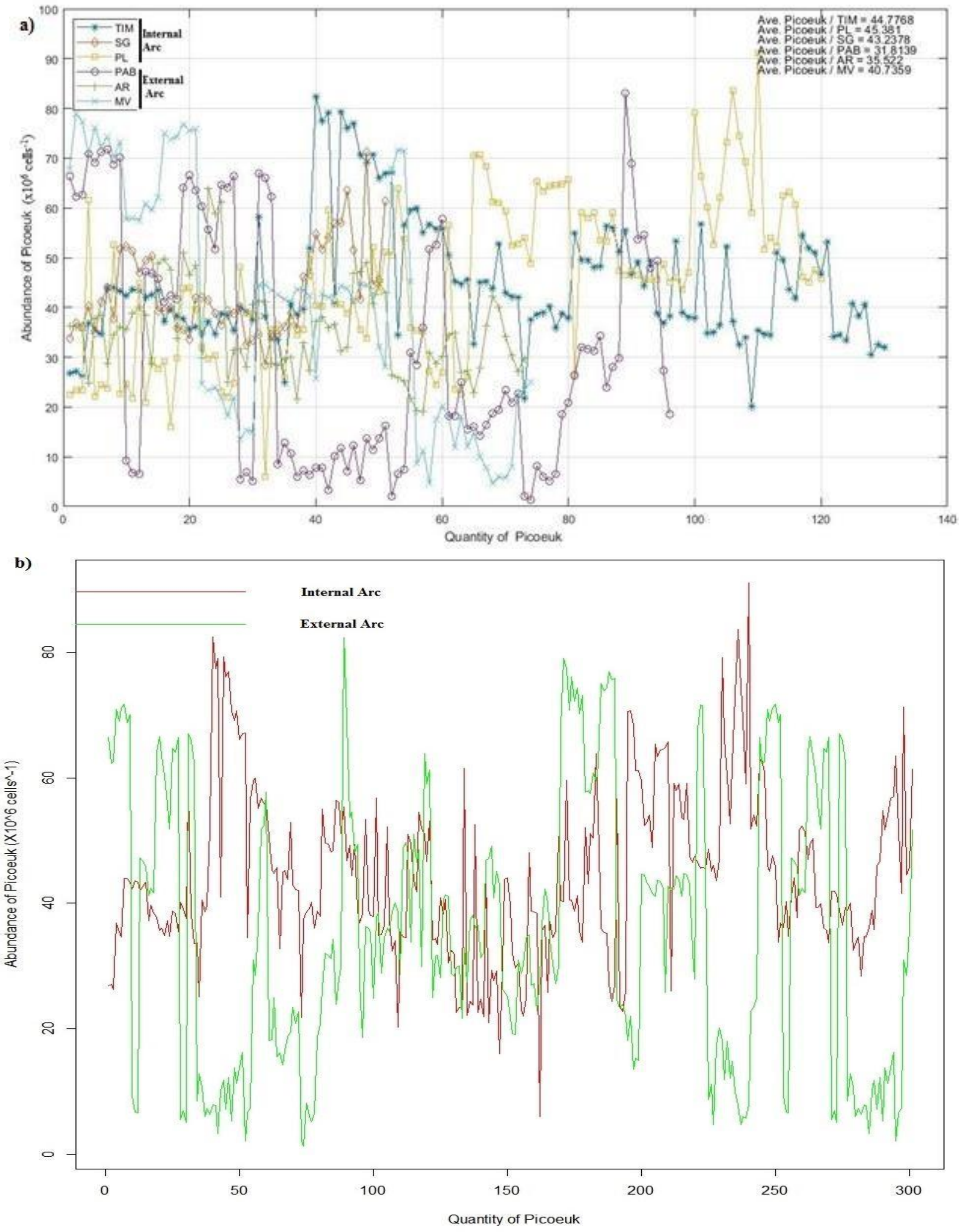


Fig. 65. Picoeuk abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

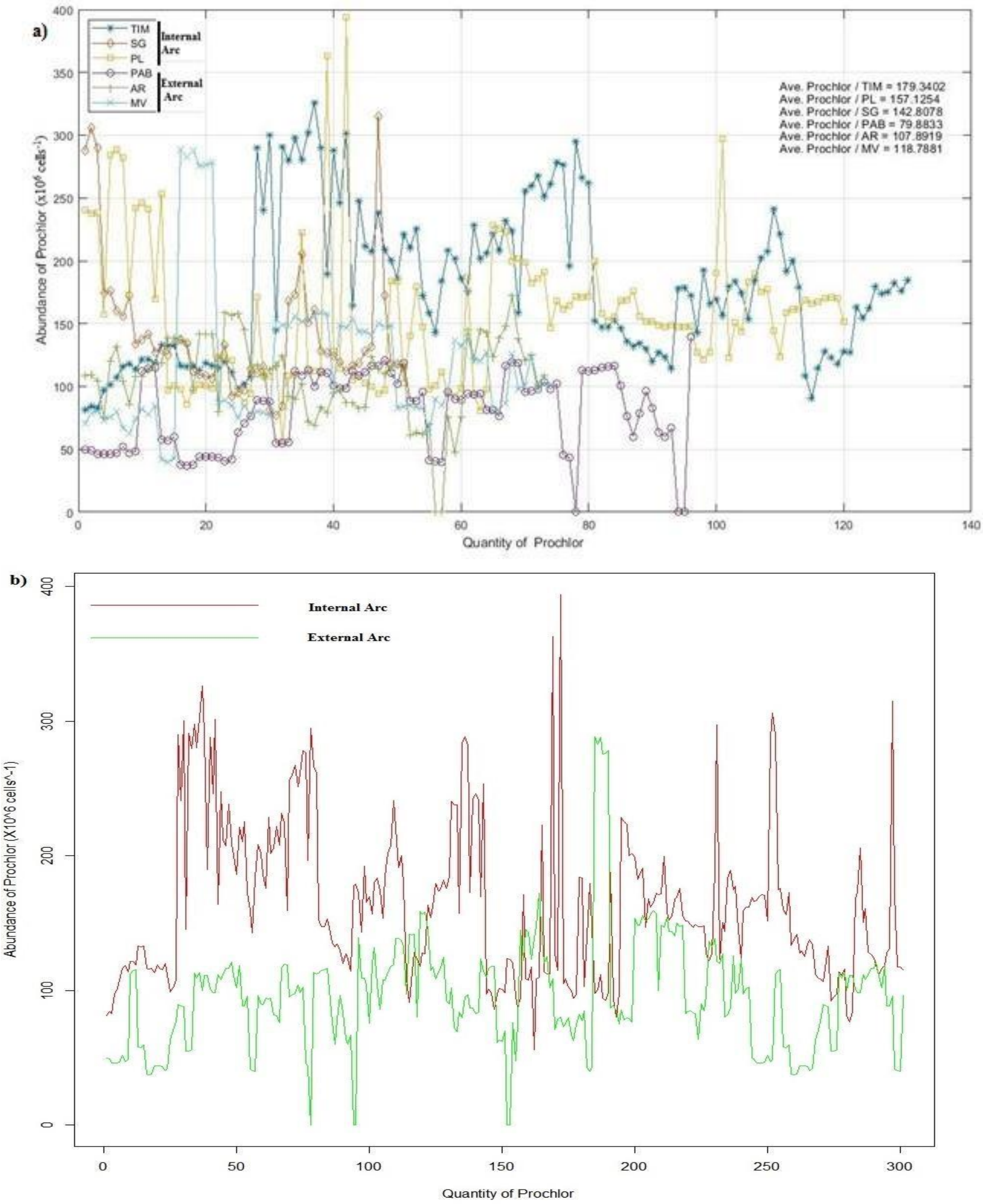


Fig. 66. Prochlor abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

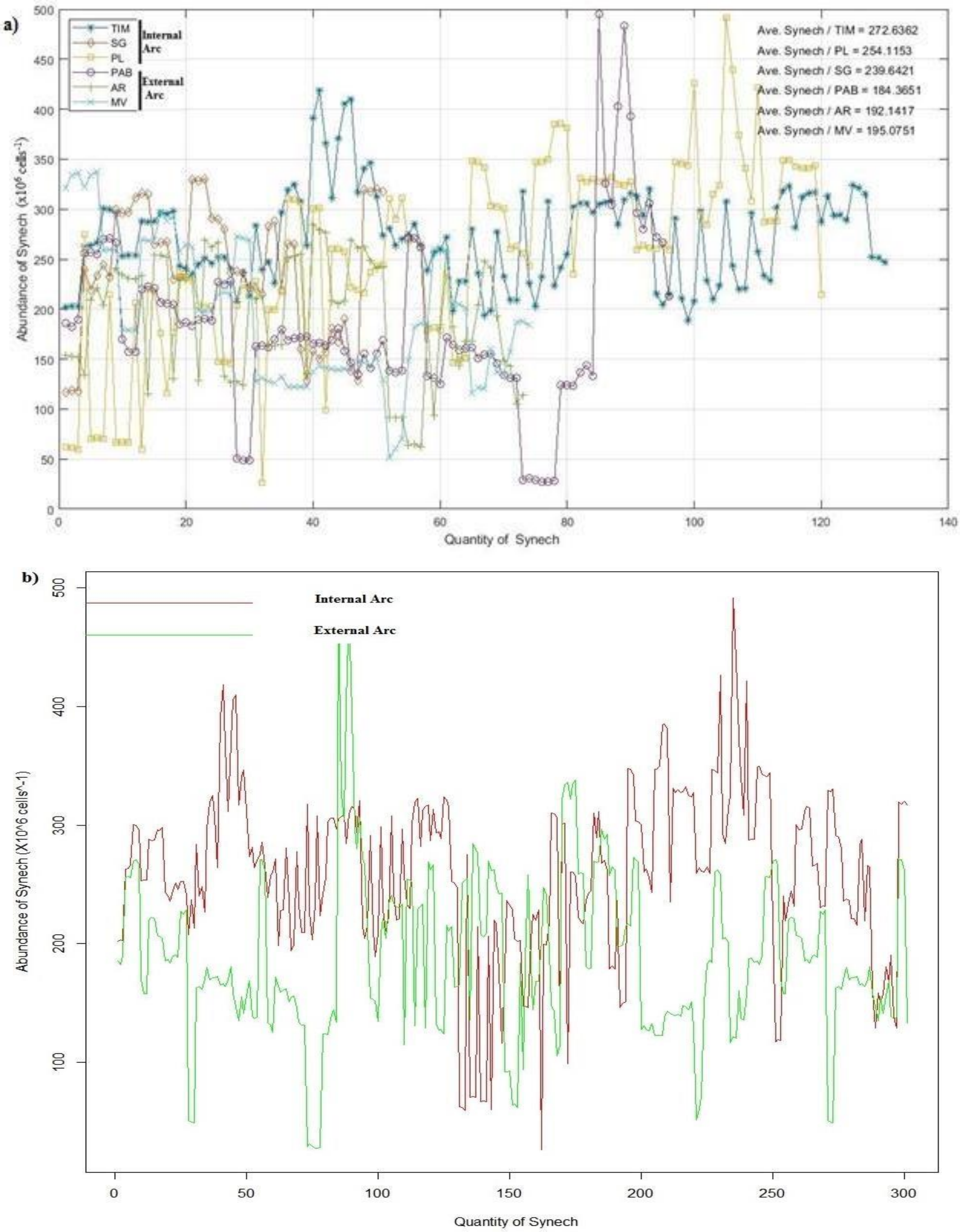


Fig. 67. Synech abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

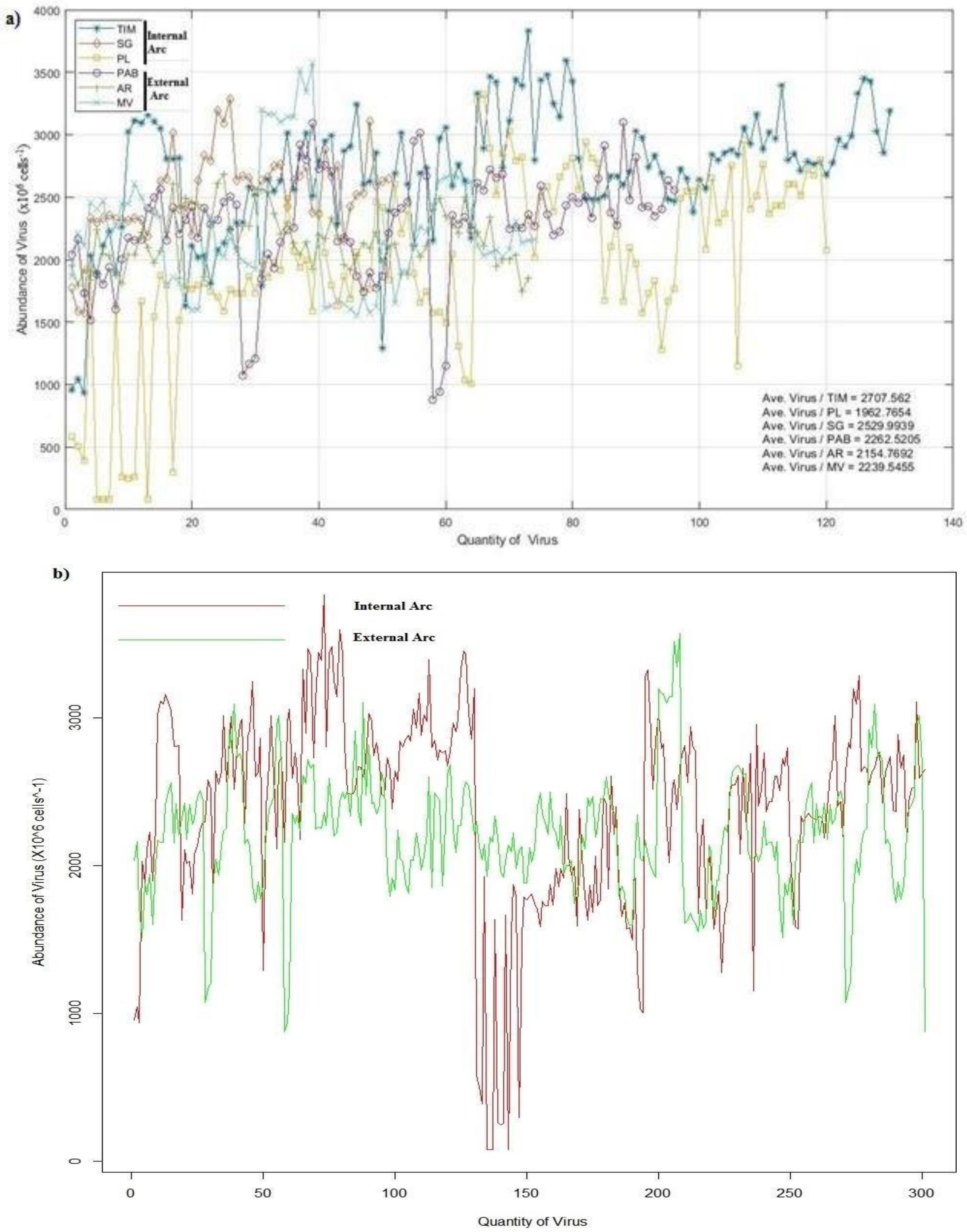


Fig. 68. Virus abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

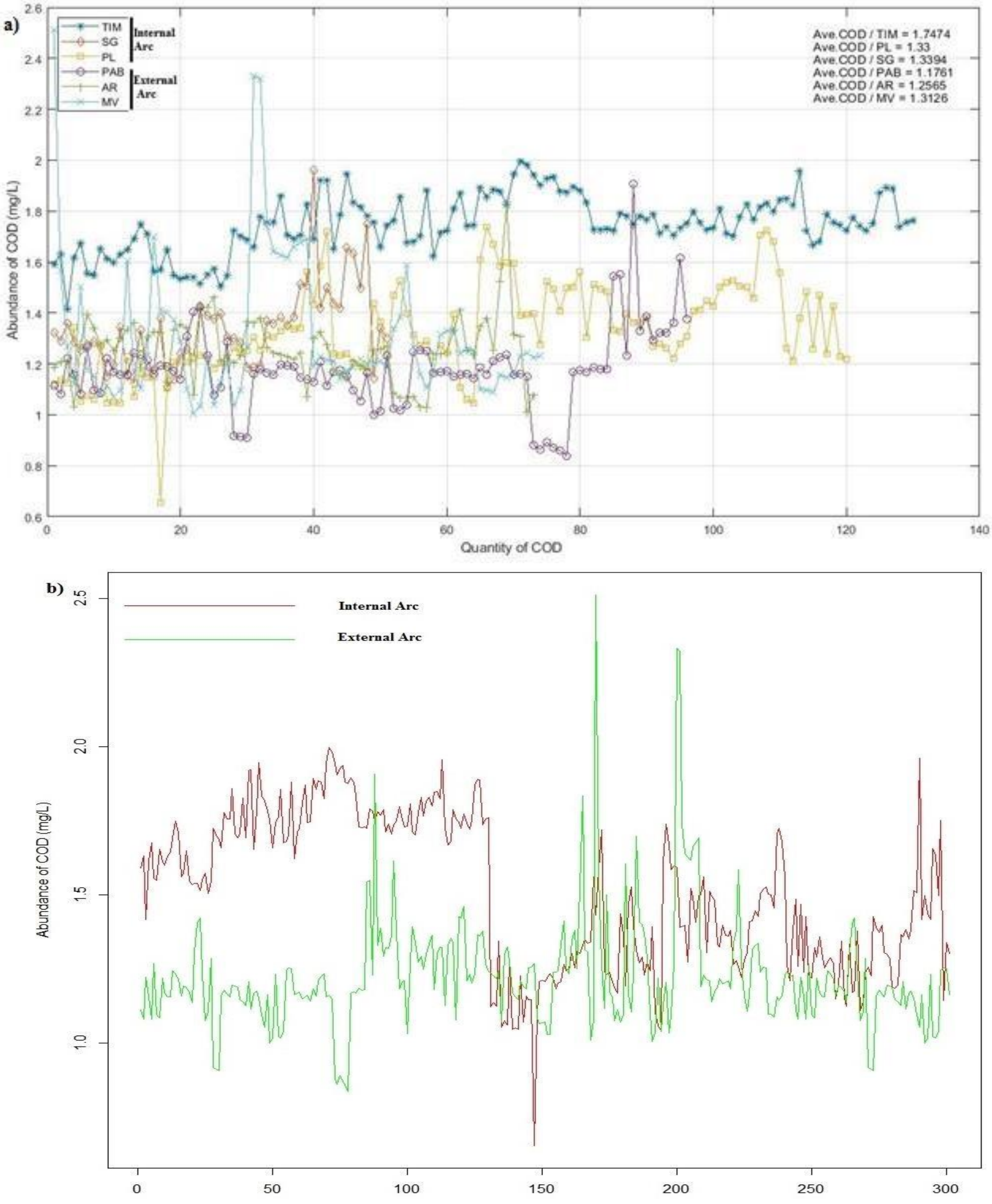


Fig. 69. DOC abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

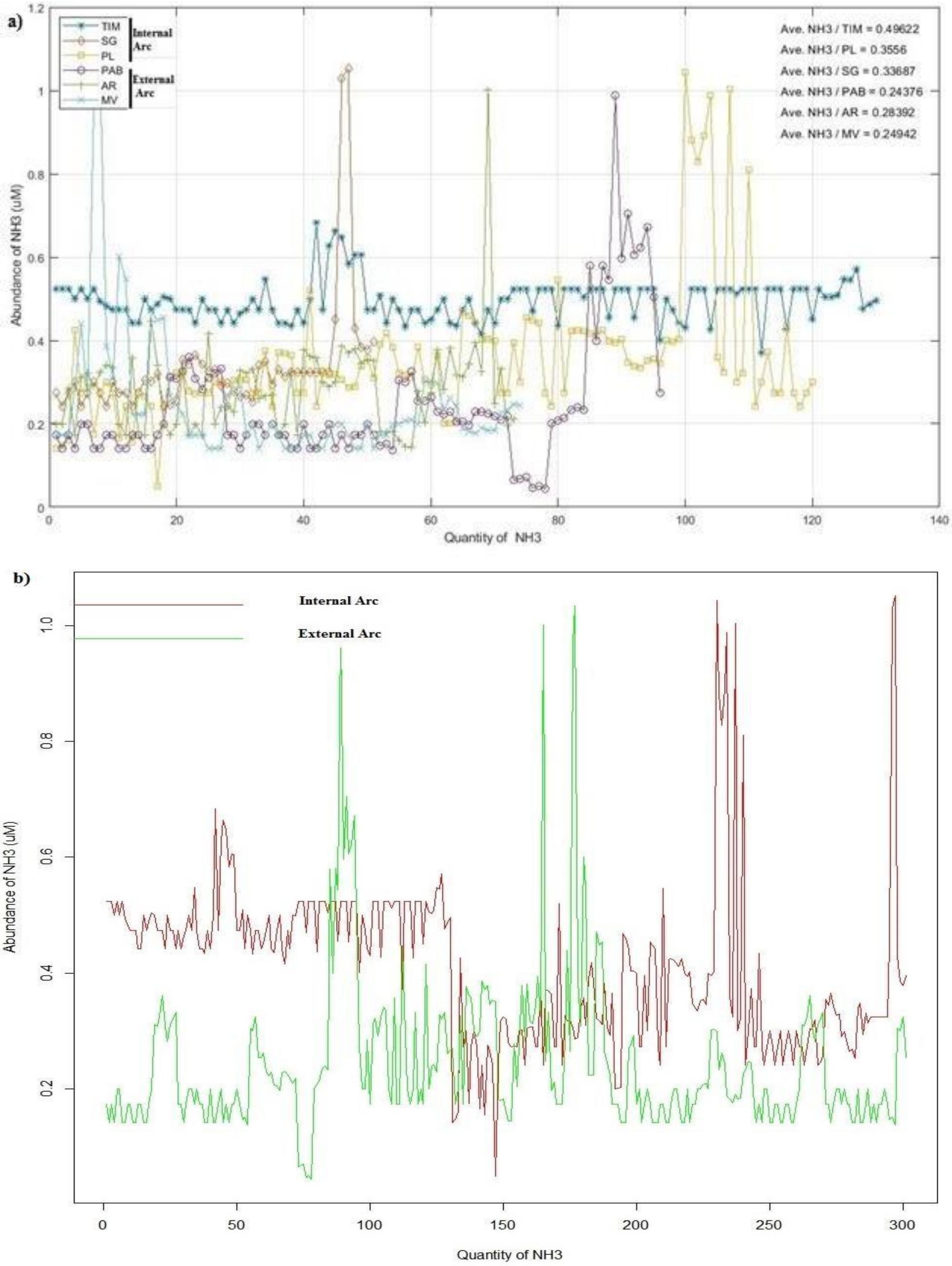


Fig. 70. NH3 abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

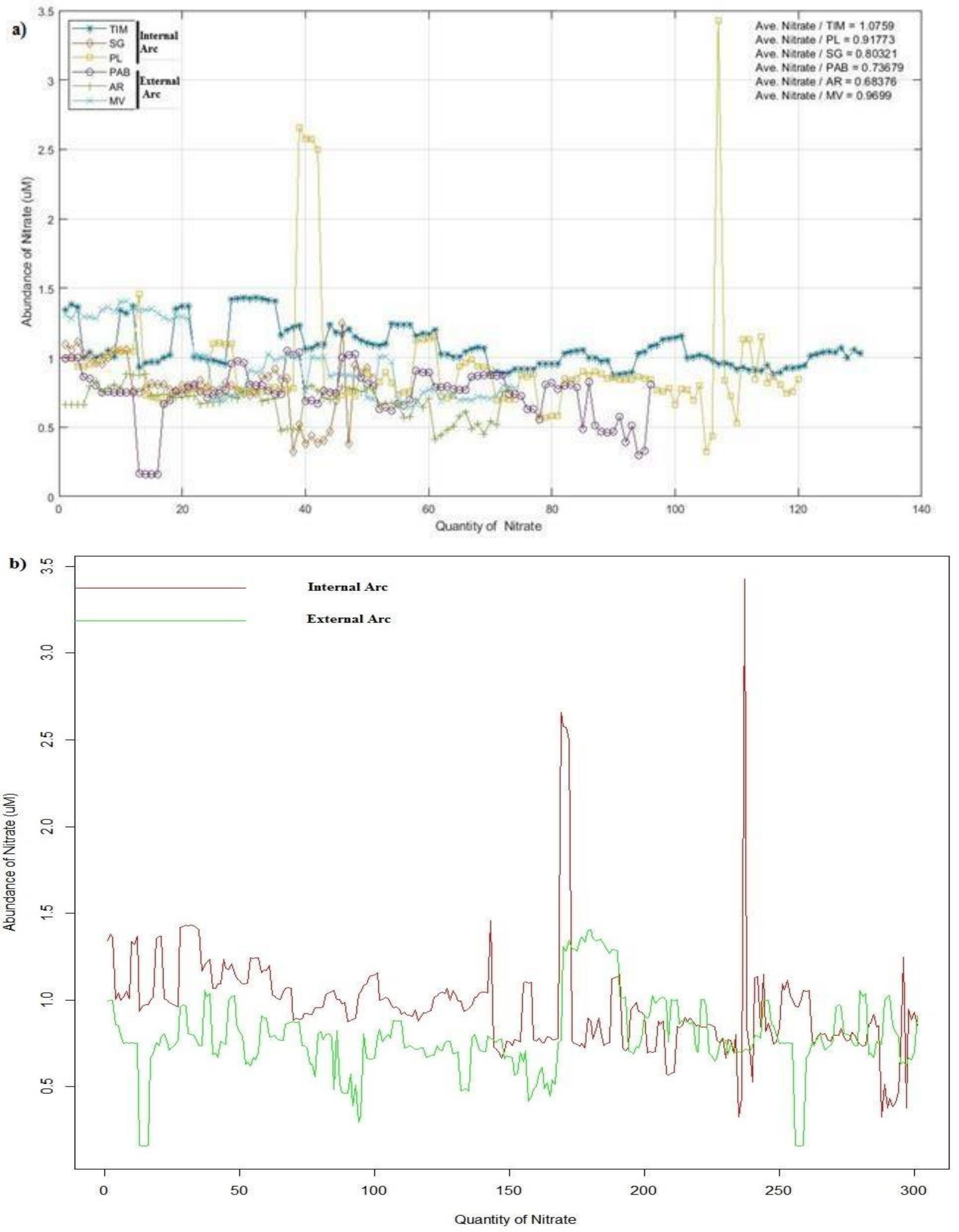


Fig. 71. Nitrate abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

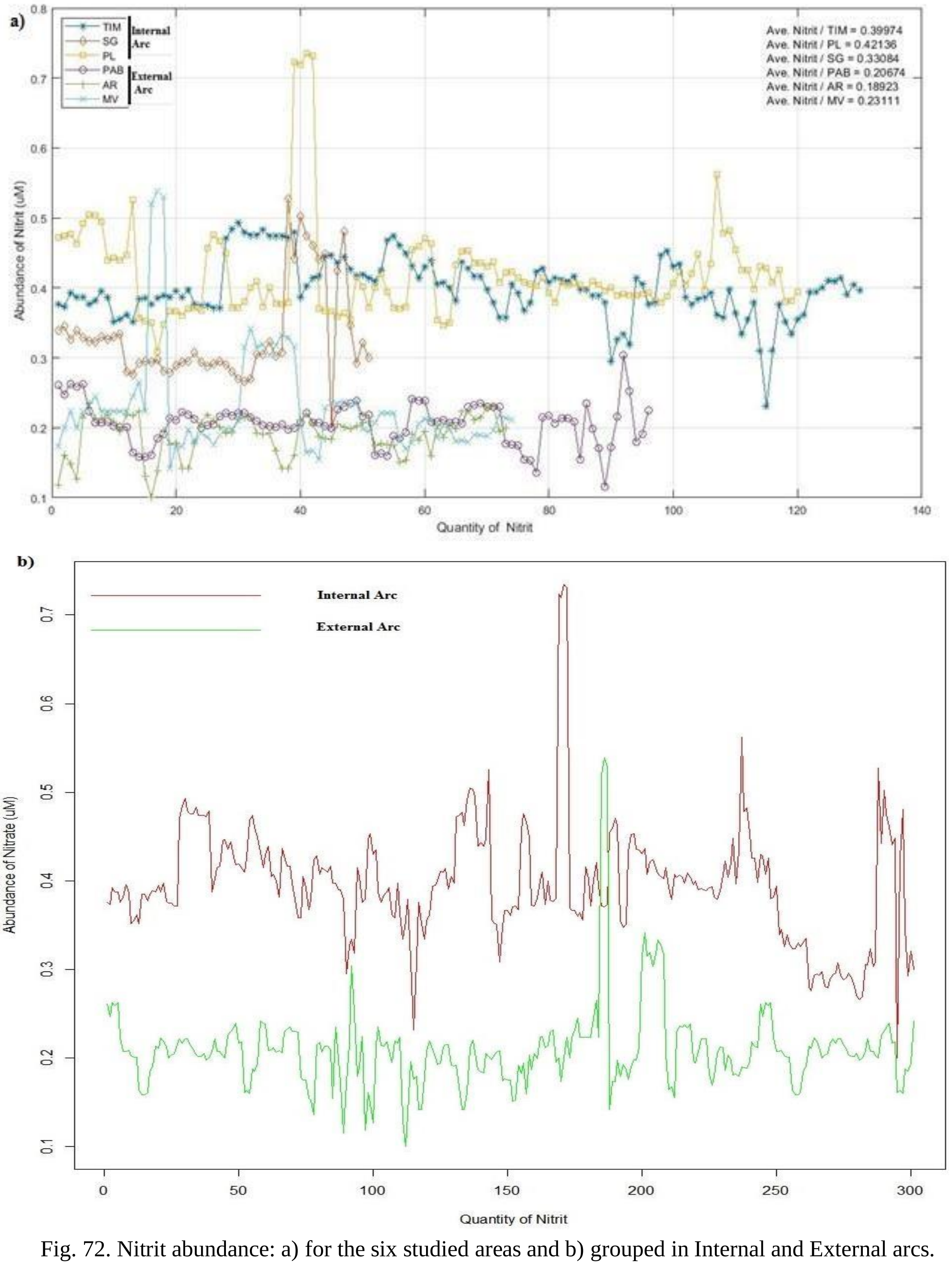


Fig. 72. Nitrit abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

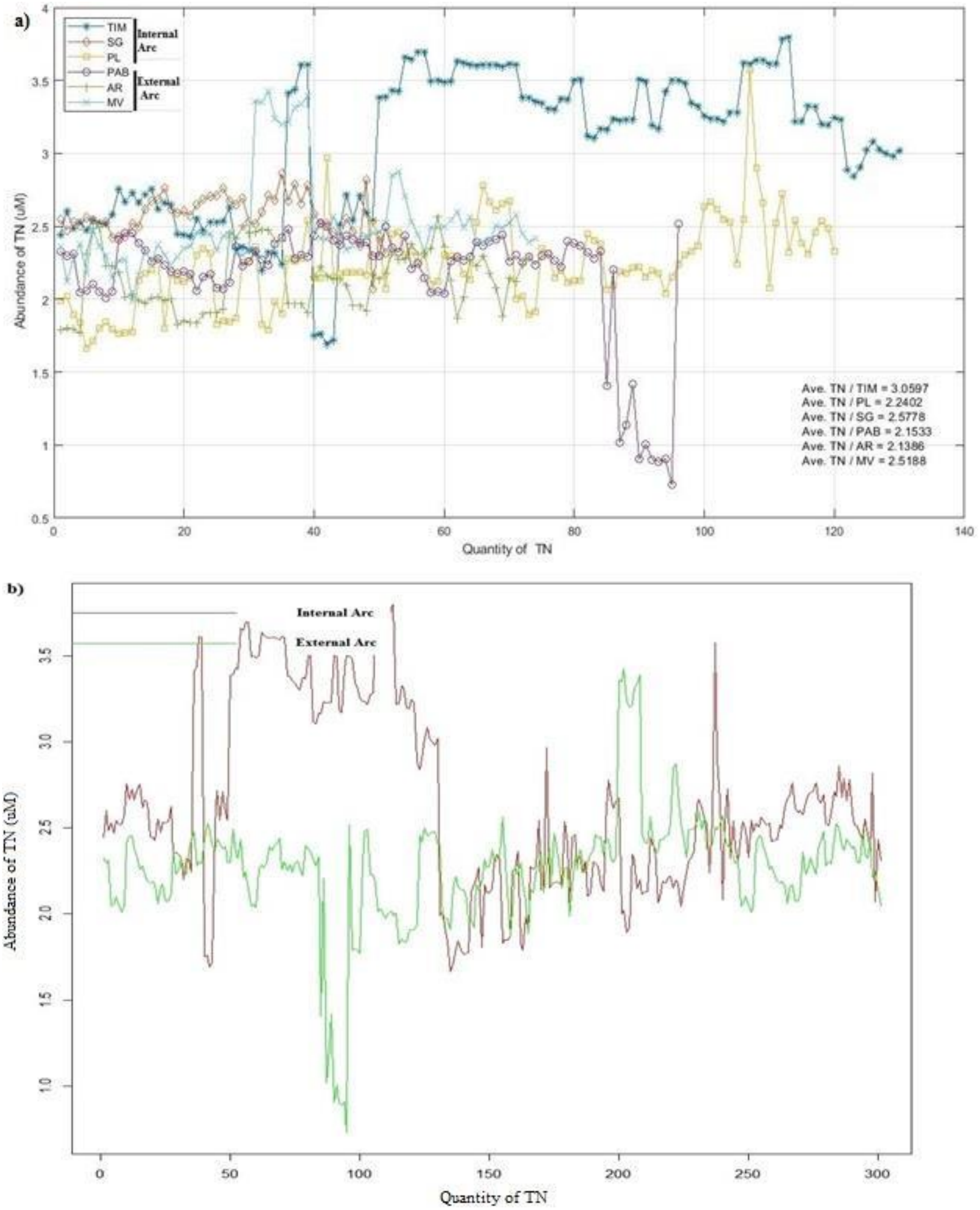


Fig. 73. TN abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

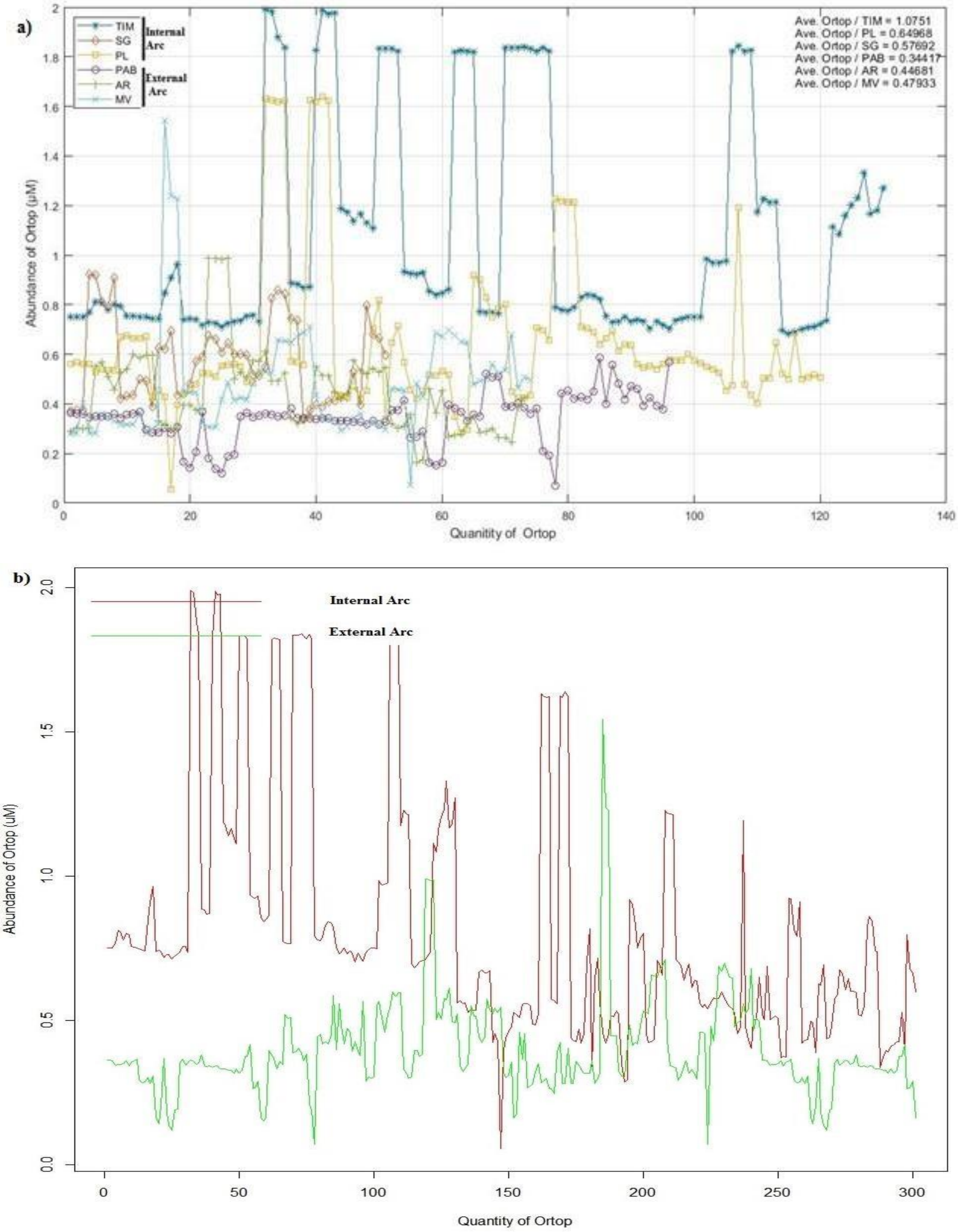


Fig. 74. Ortop abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

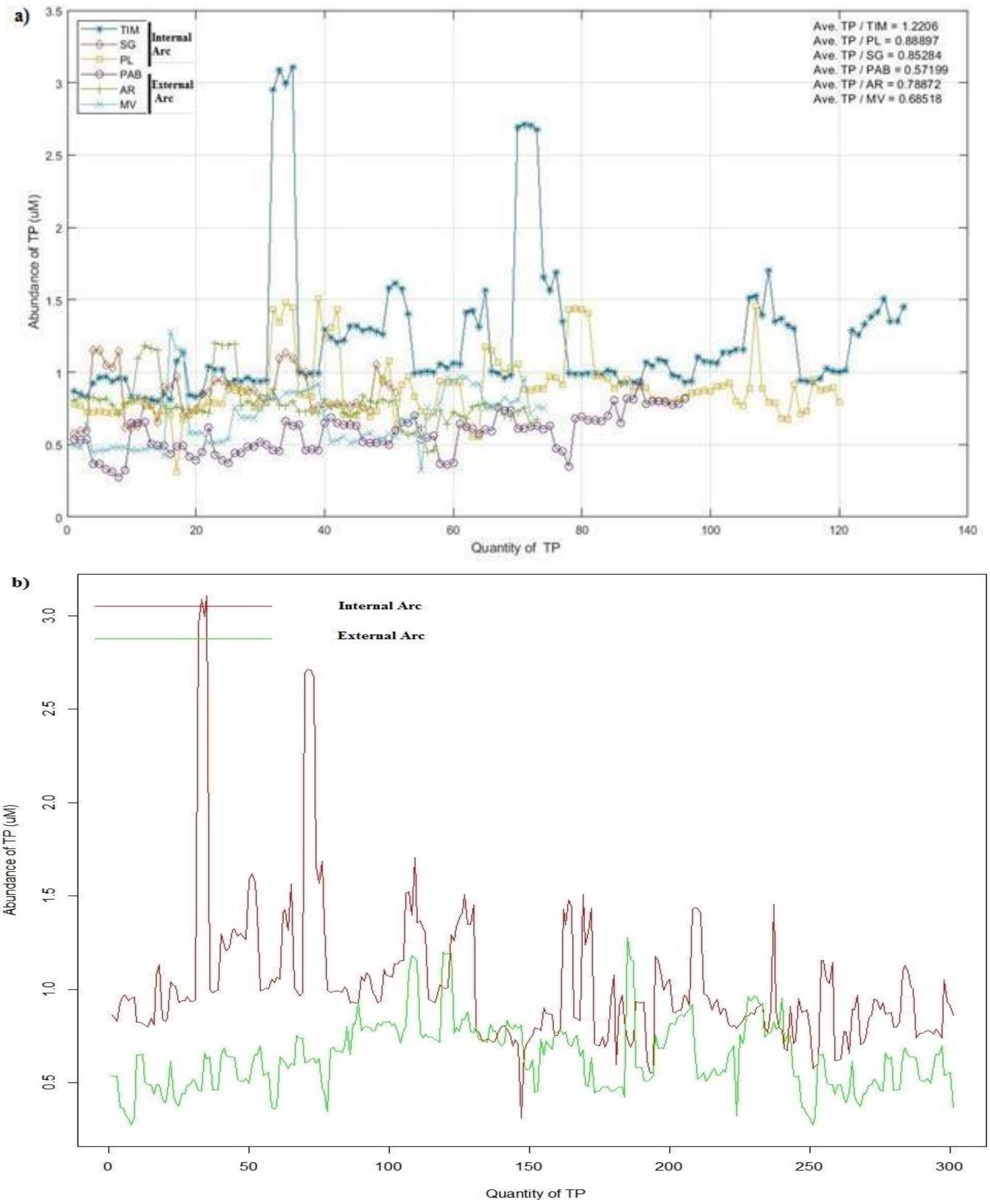


Fig. 75. TP abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

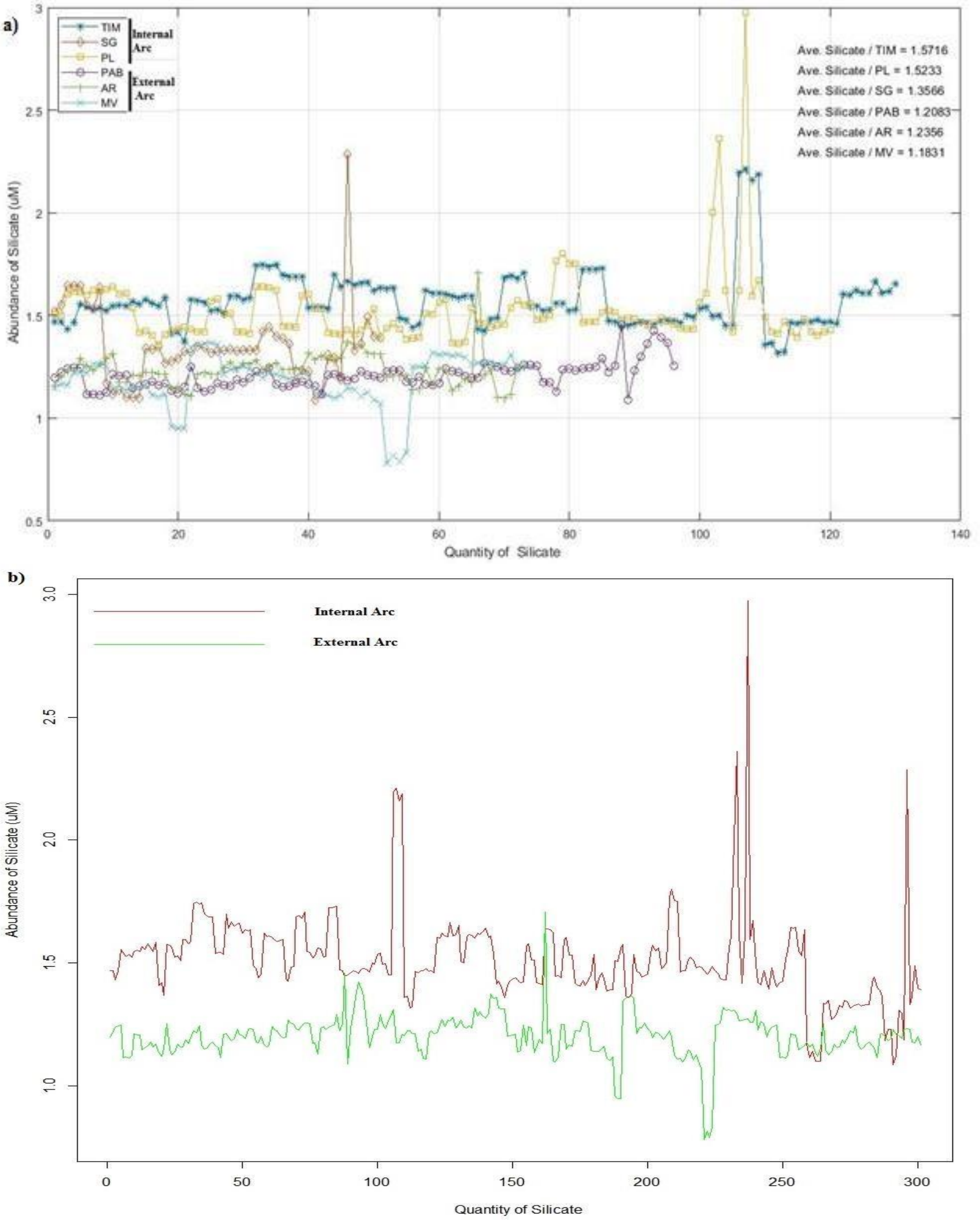


Fig. 76. Silicate abundance: a) for the six studied areas and b) grouped in Internal and External arcs.

3.4. Results' Discussion

The aim of this study was to model the influence of environmental parameters on the marine microbial abundance of the Abrolhos coral reefs using Random Forest and Boost Regression Tree.

The predictive performance of the BRT and RF models based on the 10- fold cross-validation, including the MAE, RMSE and R^2 values suggest that model RF was highly predictive based on the range of R^2 values (from 0.66 to 0.84) in compare to BRT with R^2

values from 0.55 to 0.78. In general, the descriptive statistics indicate approximately similar levels of prediction accuracy of the BRT and RF models, based on the R^2 (0.65 vs. 0.75, respectively). The R^2 values suggest that both models can explain approximately 70% of the total microbial abundance variability and both approaches accurately predicted the microbial abundance, based on R^2 values of approximately 0.7.

The relative importance of each predictor was determined by performing 100 runs of the BRT and RF models. The relative importance of the predictors differed a little between the two models, but DOC (dissolved organic carbon) was the most influential factor in the BRT and RF models. In both, the three most important predictors, based on their mean values, were DOC, TN (total nitrogen) and Silicate.

The difference of biophysical and biochemical parameters and microbial abundance in Internal and External arc of Abrolhos has been also analysed. The water temperature showed an annual profile of variation. The difference of temperature values between Internal and External arcs and in years 2011 to 2016 showed that the external arc of Abrolhos (MV and PAB) presented the lowest temperature values along the years. On the other side, the difference of hydrodynamic velocity between Internal and External arcs in 2010 showed that the AR external arc presents the highest hydrodynamic velocity, and the internal arc PL presents the lowest hydrodynamic velocity. Furthermore, the microbial abundance and biochemical parameters showed that in the Internal arcs (TIM, PLand SG), the total amount of microbial abundance and biochemical variables, such as nutrients, are higher than in the External arcs (AR, PAB and MV).

The findings of this Section 3 corroborate our initial hypothesis that the environmental variables are the primary environmental indicators representing the spatial variability of microbial abundance in the internal and external arcs of Abrolhos, Brazil. The BRT and RF models showed that the variability in microbial abundance can be primarily explained by DOC, followed by TN (total nitrogen) and Silicate (Fig. 56). The absorption of organic matter by microbial is a major route of carbon flux, and its variability can change the overall patterns of carbon flow (Azam 1998).

Bacteria are key players in organic matter recycling in the aquatic ecosystems, mediating the flux of nutrients and energy to higher trophic levels (Azam, 1988), and the results of the present study are in agree with Smith *et al.* (2006), which verified that the DOC released by algae may promote the microbial growth that leads to coral death.

Based on the BRT and RF results, NH3 (ammonia), Orthophosphate and TP (total phosphate) ranked among the least important predictors for all response variables (Fig.

56). Nevertheless, the importance values obtained for these predictors suggested that each of the response variables depends differently on the concentrations of these nutrients to r grow. Based on these observations, we postulate that microbial abundance has been much more dependent on DOC, TN and silicate concentrations than the other independent variables.

Numerous studies show that hydrodynamics is an important factor acting on ecosystems and affecting microbial distribution (Frontalini *et al.,* 2009; Martins *et al.,* 2013; Damasio *et al.,* 2020). Our findings revealed that the internal arcs of Abrolhos, specially Pedra de Leste, have low hydrodynamics, longer residence time of the water, higher concentration of nutrients, greater proliferation of microbes and possibly more coral disease; while the external arcs of Abrolhos, specially Arquipélago, have high hydrodynamics - favoring the "washing" of the reefs, lower temperature, lower concentration of nutrients, and minor proliferation of microbes. The result of the present study is in agree with Santos *et al.* (2011), which verified that the strong currents with low temperature in the marine zone of the estuary of Ria de Aveiro (Portugal) promote vertical mixing, inhibiting the establishment of bacterial community (Santos *et al.*, 2011).

Furthermore, the abiotic variables are faced by the Abrolhos photosynthetic microbiome. Importantly, other forms of chlorophyll besides the one that was measured in this study (chlorophyll a) could have had associations with predictor variables that were different from those presented here. TP, Orthophosphate and NH3 ranked among the least important predictors for all response variables (Fig. 56). Nevertheless, the importance values obtained for these predictors suggested that each of the three response variables has depended differently on phosphorus concentrations for growth. NH3 had almost no importance to microbial abundance, while the opposite pattern was observed for DOC, which was more important to microbial abundance than it. Based on these observations, we postulated that microbial abundance has been much more dependent on DOC concentrations than the other response variables.

Hydrodynamic velocities and temperature influence the ecological processes in the Abrolhos environment. In the low temperature External arcs (AR, PAB and MV), intense water currents (specifically in AR) promote a strong mixing of water, circumventing the establishment of microbial communities. In contrast, in the high temperature of internal arcs, local hydrodynamic characteristics provide the necessary conditions for the proliferation of biochemical parameters such as nutrients and active microbial abundance. Temperature was ranked among the most important factors regulating the levels of the response variables, and strong positive associations between temperature and response variables were observed (Coutinho *et al.,* 2019). Analyses of microbial communities spanning multiple ecosystems have suggested that temperature is major factors shaping microbial community composition across aquatic habitats (Lozupone and Knight, 2007; Sunagawa *et al.*, 2015; Thompson *et al.*, 2017). The positive associations with temperature were likely a reflection of the increase in microbial metabolism brought by higher temperatures. Several physical and biological processes, including simple diffusion, turbulent mixing, in situ primary production, convection and upwelling of underlying waters (UW) (Liss & Duce, 1997), contribute to the enrichment of organic and inorganic nutrients as well as microorganisms at the water column.

Coutinho *et al.,* (2019) showed that physical parameters (i.e., temperature, salinity and transparency) were more relevant for determining the abundance of bacteria, chlorophyll and Vibrio than nutrients (i.e., TP and TN). Maybe temperature and hydrodynamic have acted together in determining the abundance of microbial and photosynthesizes through three major mechanisms: altering the taxonomic composition of the community, affecting their growth rates, and regulating the rates of photosynthesis and consequently primary productivity as reported in Coutinho *et al.* (2019).

Based on this evidence, we postulated that, due to the intense eutrophication at the internal arcs of Abrolhos, the microbial community had reached its maximum capacity for taking up and utilizing nutrients. Thus, the Abrolhos microbiome growth might have no longer been limited by nutrient availability. Instead, temperature, and hydrodinamic effect have acted together in determining the abundance of microbiomes and photosynthesizers through three major mechanisms: altering the taxonomic composition of the community, affecting their growth rates, and regulating the rates of photosynthesis and consequently primary productivity. In the original pristine conditions, the growth of the microbial community was likely limited by the availability of TP and NH3.

Our findings demonstrate that hydrodynamic and temperature can possibly regulate microbial abundance, providing important insights for a better environmental management of the Abrolhos Bank. These results are in accordance with previous findings that explored the associations between microbial abundance and these parameters (Constantin de Magny *et al.*, 2008; Haley *et al.*, 2014; Höfle *et al.*, 2015; Vezzulli *et al.*, 2016). Our results corroborated these findings while also elucidating the associations between microbial abundance with DOC, NT, and silicate. Therefore, the results suggest that microbe growth has been fed by DOC-rich sewage dumped into the Abrolhos reefs.

Understanding the associations between microbes and environmental conditions has been a fundamental step in predicting, preventing and mitigating the impacts of disease outbreaks associated with aquatic habitats (Lobitz *et al.,* 2000; Russek-cohen *et al.,* 2003). Together, the warmer and lower hydrodynamics waters of the innermost sections of Abrolhos have posed a higher risk for the local population than the colder and higher hydrodynamics waters present in the regions that receive higher inputs of oceanic waters. Thus, the innermost regions of the Abrolhos likely have been the most threatening because the high sewage input and low influence of oceanic waters in these regions has created an ideal environment for the proliferation of microbes and other potential pathogens. Therefore, developing strategies for minimizing pollution into these regions of Abrolhos that represent the biggest threat to public health should be a priority. Our findings provide insights for developing strategies for reversing the impacts to inner arc of Abrolhos. This could be achieved by a combination of proper sewage treatment, reduction of nutrient loads, minimizing deforestation and recovery of the surrounding and aquatic vegetation. Additionally, bioremediation strategies capable of reducing nutrient availability could be applied as well (Boesch *et al.*, 2001; Greening and Janicki, 2006; Little *et al.*, 2000; McGann *et al.*, 2003; Paerl, 2009; Walker *et al.*, 2013; Coutinho *et al.*, 2019). Furthermore, our results demonstrate that BRT and RF models could serve as tools to assess the threat level to public health posed by the aquatic ecosystem throughout changing environmental conditions, which could be used to predict and mitigate coral disease outbreaks (Yang *et al.,* 2016).

Our results are in agree with Coutinho *et al.* (2019), which concluded that microbiomes levels were primarily regulated by temperature. Also, these results were in accordance with previous findings that explored the associations between microbes and abiotic parameters (Paerl, 2009; Walker *et al.*, 2013; Coutinho *et al.*, 2019). Our results corroborated these findings while also elucidating the associations between microbial abundance, DOC, TN and silicate. Therefore, the results suggest that the growth of microbes and of other potentially pathogenic bacteria has been fuelled by nutrient concenterations that releases carbone, Nitrogen and silicate sources into the Abrolhos bank. This explanation concurs with previous analyses that have indicated that Vibrio and other copiotrophic and potentially pathogenic bacteria depend on DOC (Coutinho *et al.*, 2015).

Together, these findings indicated that warmer and less hydrodinamic velocity waters of the innermost sections of Abrolhos have posed a higher risk for the local population than the colder and more hydrodinamic velocity waters typical of the regions that receive higher inputs of oceanic waters. Thus, the innermost regions of the Abrolhos likely have been the most threatening because the high nutrients and low influence of oceanic waters in these regions has created an ideal environment for the proliferation of microbes and other potential pathogens. Therefore, developing strategies for minimizing pollution into these regions of Abrolhos that represent the biggest threat to public health should be a priority.

Despite decades of investments, the water pollution at Abrolhos is increasing. The continuous of temperature of this area lead to increased nutrient concentrations. Furthermore, climate change is expected to increase water temperatures in Abrolhos. Assuming that the observed associations between predictors and response variables remain stable through time, our results suggest that the aforementioned changes expected to affect this ecosystem in the future would lead to higher densities of microbes, potential pathogens and coral disease.

The models were used to verify the associations between biological variables and abiotic parameters regardless of how these associations change through time. More complex models that incorporate the aforementioned variables would require a much larger number of samples but could provide a more comprehensive understanding of the dynamics taking place within the microbial community that resides in Abrolhos. Likewise, the advancement of models approaches and perhaps the use of algorithms designed specifically to incorporate temporal trends (e.g., recurrent neural networks) could improve the precision of these models. Nevertheless, our work provides a stepping stone for future studies that aim to understand the dynamics of the Abrolhos microbiome through ecological modelling approaches.

4. Conclusions and Suggestions for Future Research

We concluded that hydrodynamic velocities and temperature influences the ecological processes in the Abrolhos environment. In the low temperature External arcs (AR, PAB and MV), intense water currents (specifically in AR) promote a strong mixing of water, circumventing the establishment of microbial communities. In contrast, in the high temperature of internal arcs, local hydrodynamic characteristics provide the necessary conditions for the proliferation of biochemical parameters such as nutrients and active microbial abundance.

This work contributes toward a better understanding of the ecology of microbial communities at Abrolhos ecosystems. Boosted Regression Trees (BRT) and Random Forest (RF) models were very important to attest causal inference, providing supporting evidence of how the microbial abundance has been regulated by the environmental parameters. We have ranked the relative importance of these parameters over the response variables, and characterized the synergistic effects between variables. Furthermore, BRT and RF allowed us to infer the response of the microbial communities to changes in water quality conditions. Our findings provide insightful information on the dynamics of microbial communities for tropical ecosystems, for which little information is currently available. This approach could be easily applied to other similar datasets from other ecosystems and has served as a proof-of-principle of the usefulness of these models in the field of microbial ecology. This outcome is especially relevant considering the current scenarios of global climate changes and increasing environmental impacts.

In fact, understanding the associations between microbes and environmental conditions is a fundamental step in predicting, preventing and mitigating the impacts of disease outbreaks associated with aquatic habitats. We showed that the warmer and lower hydrodynamics waters of the innermost sections of Abrolhos have posed a higher risk for the local population than the colder and higher hydrodynamics waters present in the regions that receive higher inputs of oceanic waters: the innermost regions of Abrolhos present an ideal environment for the proliferation of microbes and other potential pathogens.

Therefore, developing strategies for minimizing pollution into these regionsthat represent the biggest threat to public health should be a priority, and this could be achieved by a combination of proper sewage treatment, reduction of nutrient loads, deforestation minimization, and recovery of the surrounding and aquatic vegetation.

Furthermore, our results demonstrate that BRT and RF models could serve as good tools to assess the threat level to public health posed by the aquatic ecosystem throughout changing environmental conditions, which could be used to predict and mitigate coral disease outbreaks.

Our findings can be very useful for the development of strategies to reducethe burden of waterborne diseases and for the remediation of Abrolhos and other aquaticecosystems, aiming to preserve their biodiversity as well as their economic, historical andaesthetic values.

As future research, we can suggest the use of a larger sample set, having more predictor variables, and incorporating associations between biotic variables to provide an even better description of the ecological associations taking place within the Abrolhos microbiome.

References

Bell, J. J. 2008. The functional roles of marine sponges. Estuarine, coastal and shelf science, 79(3), 341-353. https://doi.org/10.1016/j.ecss.2008.05.002.

Bishop, C., 1995. Neural Networks for Pattern Recognition. Oxford University Press, New York. ISBN:978-0-19-853864-6.

Boesch, D.F., Brinsfield, R.B., Magnien, R.E., 2001. Chesapeake Bay eutrophication: Scientific Understanding, Ecosystem Restoration, and Challenges for Agriculture. J. Environ. Qual., 30(2), 303–320. https://doi.org/10.2134/jeq2001.302303x.

Breiman, L., 2001. Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Bruce, T., Meirelles, P. M., Garcia, G., Paranhos, R., Rezende, C. E., de Moura, R. L., Francini-Filho, R, Coni, E. O. C., Vasconcelos, A. T., Amado Filho, G., Hatay, M., Schmieder, R., Edwards, R., Dinsdale, E., Thompson, F. L. 2012. Abrolhos bank reef health evaluated by means of water quality, microbial diversity, benthic cover, and fish biomass data. PloS One, 7(6), e36687. https://doi.org/10.1371/journal.pone.0036687.

Caiqing, ZH., Ruonan, Qi., Zhiwen, Qi., 2008. Comparing BP and RBF Neural Network for Forecasting the Resident Consumer Level by MATLAB. International Conference on Computer and Electrical Engineering, 169-172, https://doi.org/10.1109/ICCEE.2008.35.

Constantin de Magny, G., Murtugudde, R., Sapiano, M.R.P., Nizam, A., Brown, C.W., Busalacchi, A.J., Yunus, M., Nair, G.B., Gil, A.I., Lanata, C.F., Calkins, J., Manna, B., Rajendran, K., Bhattacharya, M.K., Huq, A., Sack, R.B., Colwell, R.R., 2008. Environmental signatures associated with cholera epidemics. Proc. Natl. Acad. Sci. USA, 105, 17676–17681. https://doi.org/10.1073/pnas.0809654105.

Coutinho, F. H., Thompson, C. C., Cabral, A. S., Paranhos, R., Dutilh, B. E., Thompson, F. L. (2019). Modelling the influence of environmental parameters over marine planktonic microbial communities using artificial neural networks. Science of the Total Environment, 677, 205-214. https://doi.org/10.1016/j.scitotenv.2019.04.009.

Damasio, B. V., Timoszczuk, C. T., Kim, B. S. M., e Sousa, S. H. D. M., Bícego, M. C., Siegle, E., Figueira, R. C. L. (2020). Impacts of hydrodynamics and pollutants on foraminiferal fauna distribution in the Santos Estuary (SE Brazil). Journal of Sedimentary Environments, 5(1), 61-86. https://doi.org/10.1007/s43217-020-00003-w.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification. Wiley-Interscience, New York. ISBN: 978-0-471-05669-0.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Francini-Filho, R. B., Coni, E. O., Meirelles, P. M., Amado-Filho, G. M., Thompson, F. L., Pereira-Filho, G. H., Bastos, A. C., Abrantes, D. P., Ferreira, C. M., Gibran, F. Z., Güth, A. Z., Sumida, P. Y. G., Oliveira, N. L., Kaufman, L., Minte-Vera, C. V., Moura, R. L. 2013. Dynamics of coral reef benthic assemblages of the Abrolhos Bank, eastern Brazil: inferences on natural and anthropogenic drivers. PloS One, 8(1), e54260. https://doi.org/10.1371/journal.pone.0054260.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat., 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal., 38(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2.

Friedman, J.H., Meulman, J.J., 2003. Multiple additive regression trees with application in epidemiology. Stat. Med. 22(9), 1365–1381. https://doi.org/10.1002/sim.1501.

Froeschke, J.T., Froeschke, B.F., 2011. Spatio-temporal predictive model based on environmental factors for juvenile spotted sea trout in Texas estuaries using boosted regression trees. Fish. Res., 111(3), 131–138. https://doi.org/10.1016/j.fishres.2011.07.008.

Goulder, R. 1980. Seasonal variation in heterotrophic activity and population density of planktonic bacteria in a clean river. The Journal of Ecology, 68(2), 349-363. https://doi.org/10.2307/2259410.

Gower, J. C. 2014. Principal coordinates analysis. Wiley StatsRef: Statistics Reference Online. https://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat05670.

Greening, H., Janicki, A., 2006. Toward reversal of eutrophic conditions in a subtropical estuary: water quality and seagrass response to nitrogen loading reductions in Tampa Bay, Florida, USA. Environ. Manag., 38, 163–178. https://doi.org/10.1007/s00267- 005-0079-4.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island-digital soil mapping using Random Forests analysis. Geoderma 146, 102–113. https://doi.org/10.1016/j.geoderma.2008.05.008.

Haley, B.J., Kokashvili, T., Tskshvediani, A., Janelidze, N., Mitaishvili, N., Grim, C.J., de Magny, G.C., Chen, A.J., Taviani, E., Eliashvili, T., Tediashvili, M., Whitehouse, C.A., Colwell, R.R., Huq, A., 2014. Molecular diversity and predictability of Vibrio parahaemolyticus along the Georgian coastal zone of the Black Sea. Front. Microbiol., 5, 1–9. https://doi.org/10.3389/fmicb.2014.00045.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer- Verlag, New York. https://web.stanford.edu/~hastie/Papers/ESLII.pdf.

Höfle, M., Pezzati, E., Vezzulli, L., Brettar, I., Pruzzo, C., 2015. Effects of Global Warming on Vibrio Ecology. Microbiol. Spectr., 3(3). https://doi.org/10.1128/microbiolspec.ve-0004- 2014.

Little, J.L., Hall, R.I., Quinlan, R., Smol, J.P., 2000. Past trophic status and hypolimnetic anoxia during eutrophicaton and remediation of Gravenhurst Bay, Ontario: comparison of diatoms, chironomids, and historical records. Can. J. Fish. Aquat. Sci. 57, 333–341. https://doi.org/10.1139/f99-235.

Lobitz, B., Beck, L., Huq, A.,Wood, B., Fuchs, G., Faruque, A.S., Colwell, R., 2000. Climate and infectious disease: use of remote sensing for detection of Vibrio cholerae by indirect measurement. Proc. Natl. Acad. Sci. USA, 97, 1438–1443. https://doi.org10.1073/pnas.97.4.1438.

Lozupone, C.A., Knight, R., 2007. Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. USA, 104, 11436–11440. https://doi.org/10.1073/pnas.0611525104.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News2, 18–22. [Online]. Available on: https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf.

Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. Geoderma 170,70–79. https://doi.org/10.1016/j.geoderma.2011.10.010.

Lin, L., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45, 255–268. https://doi.org/10.2307/2532051.

Liss, P. S.; Duce, R. A., 1997. The Sea Surface and Global Change. Cambridge University Press, Cambridge. ISBN 9780511525025. https://doi.org/10.1017/CBO9780511525025.

McGann, M., Alexander, C.R., Bay, S.M., 2003. Response of benthic foraminifers to sewage discharge and remediation in Santa Monica Bay, California. Mar. Environ. Res. 56, 299–342. https://doi.org/10.1016/S0141-1136(02)00336-7.

Milner, C. R., Goulder, R. 1986. The abundance, heterotrophic activity and taxonomy of bacteria in a stream subject to pollution by chlorophenols, nitrophenols and phenoxyalkanoic acids. Water Research, 20(1), 85-90.

Morikawa, K. 1984. Seasonal fluctuation in the number of aerobic heterotrophic bacteria and its relation to environmental factors in the upstream area of the Tamagawa River. Jap. J. Limnol., 45, 69-78. https://doi.org/10.3739/rikusui.45.69.

Paerl, H.W., 2009. Controlling eutrophication along the freshwater-marine continuum: dual nutrient (N and P) reductions are essential. Estuar. Coasts 32, 593–601. https://doi.org/10.1007/s12237-009-9158-8.

Peters, J., Verhoest, N., Samson, R., Boeckx, P., De Baets, B., 2008. Wetland vegeta-tion distribution modelling for the identification of constraining environmental variables. Landscape Ecol., 23, 1049–1065. https://doi.org/10.1007/s10980-008-9261-4.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression treetechniques: bagging and random forests for ecological prediction. Ecosystems, 9, 181–199. https://doi.org/10.1007/s10021-005-0054-1.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. [Online]. Available on: http://lib.stat.cmu.edu/R/CRAN/doc/manuals/rdevel/fullrefman.pdf

Russek-cohen, E., Choopun, N., Rivera, I.N.G., Gangle, B., Jiang, S.C., Rubin, A., Patz, J. a, Huq, A., Colwell, R.R., 2003. Predictability of Vibrio cholerae in Chesapeake Bay Vale. Appl. Environ. Microbiol., 69, 2773–2785. https://doi.org/10.1128/AEM.69.5.2773.

Santos L, Santos AL, Coelho FJ, Gomes NC, Dias JM, Cunha Â, Almeida A. Relation between bacterial activity in the surface microlayer and estuarine hydrodynamics. 2011. FEMS Microbiology Ecol., 77(3):636-646. https://doi.org/10.1111/j.1574-6941.2011.01147.x.

Schlink, U., Dorling, S., Pelikan, E., Nunnari, G., Cawley, G., Junninen, H., Vondracek, J., Greig, A., Foxall, R., Eben, K., Chatterton, T., Vondracek, J., Richter, M., Dostal, M., Bertucco, L., Kolehmainen, M., Doyle, M. 2003. A rigorous inter-comparison of ground-level ozone predictions. Atmospheric Environment, 37(23), 3237-3253. https://doi.org/10.1016/S1352-2310(03)00330-3

Shamseldin, A. Y., O'connor, K. M. 2001. A non-linear neural network technique for updating of river flow forecasts. Hydrology and Earth System Sciences, 5(4), 577-598. https://doi.org/10.5194/hess-5-577-2001.

Skurichina, M., Duin, R.P., 2002. Bagging, boosting and the random subspace methodfor linear classifiers. Pattern Anal. Appl., 5, 121–135. https://doi.org/10.1007/s100440200011.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., Cornejo-castillo, F.M., Costea, P.I., Cruaud, C., Ovidio, F., Engelen, S., Ferrera, I., Gasol, J.M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., 2015. Structure and function of the global ocean microbiome. Science, 348, 1–10. https://doi.org/10.1126/science.1261359 80.

Ryu, C. 2019. Package dlookr: Tools for Data Diagnosis, Exploration, Transformation. [Online]. Available on: https://cran.r-project.org/web/packages/dlookr/dlookr.pdf.

Thompson, B. 2005. Canonical correlation analysis, in Encyclopedia of Statistics in Behavioral Science. Hoboken, NJ, USA: Wiley. [Online]. Available on: https://onlinelibrary.wiley.com.

Thompson, L., Sanders, J., McDonald, D. et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature, 551, 457–463 (2017). https://doi.org/10.1038/nature24621.

Tranmer, M., Elliot, M. 2008. Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR), Working Papers. [Online]. Available on: http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2008/2008-19-multiple-linear-regression.pdf.

Vezzulli, L., Grande, C., Reid, P.C., Hélaouët, P., Edwards, M., Höfle, M.G., Brettar, I., Colwell, R.R., Pruzzo, C., 2016. Climate influence on Vibrio and associated human diseases during the past half-century in the coastal North Atlantic. Proc. Natl. Acad. Sci., 113, E5062–E5071. https://doi.org/10.1073/pnas.1609157113.

Walker, T.R., MacAskill, D., Rushton, T., Thalheimer, A., Weaver, P., 2013. Monitoring effects of remediation on natural sediment recovery in Sydney Harbour, Nova Scotia. Environ. Monit. Assess., 185, 8089–8107. https://doi.org/10.1007/s10661-013-3157-8.

Wold, S., Esbensen, K., Geladi, P. 1987. Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), 37-52. http://dx.doi.org/10.1016/0169-7439(87)80084-9.

Yang, R. M., Zhang, G. L., Liu, F., Lu, Y. Y., Yang, F., Yang, F., Yang, M., Zhao, Y. G., Li, D. C. (2016). Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecological Indicators, 60, 870-878. https://doi.org/10.1016/j.ecolind.2015.0.