Boston University

Predicting soil moisture and temperature using regression analysis & PCA on hyperspectral data

Vicky Karkera

CS555: Foundations Of Machine Learning

Dr Hong Pan

April 27 2023

Introduction

Research Scenario:

Remote sensing is the use of sensors installed on aircraft or satellites to obtain information about the Earth and its surroundings. These sensors acquire data without physically touching the object or area under study. Among many other things, remote sensing data can be used to make maps, monitor changes in land use and land cover, discover weather patterns, and research natural resources. Images, spectral measurements, and other sorts of measurements can be collected as data. Agriculture, forestry, urban planning, and natural resource management all benefit from remote sensing technology. Sensors in remote sensing are designed to catch radiation in certain spectral bands such as visible, near-infrared, and thermal infrared. Hyperspectral remote sensing is a sort of remote sensing technology that collects and analyzes data from a variety of spectral bands, often encompassing a wide range of wavelengths [2]. Each spectral band represents a particular wavelength range that can be used to collect various types of information about the Earth's surface and atmosphere. The visible spectral band, for example, catches radiation in the range of 0.4 to 0.7 micrometers and can be used to differentiate between different types of vegetation. The thermal infrared spectral band detects radiation with wavelengths ranging from 8 to 15 micrometers and can be used to assess temperature fluctuations on Earth's surface.Hyperspectral sensors may collect data in hundreds of tiny, contiguous spectral bands, enabling more detailed and precise analysis of the objects or locations under study.Agriculture, environmental monitoring, mineral exploitation, and military intelligence are just few of the industries where this technology might be used [1].

Research questions:

(1)Can hyperspectral data be used to identify areas of soil with particularly high or low moisture content, which could be targeted for irrigation or other interventions?

(2)Can hyperspectral features be used to identify areas where plants may be at risk due to high or low temperatures?

(3)Can principal component analysis improve upon traditional statistical models for predicting soil moisture and temperature using hyperspectral data?

Methodology

Background

1. *Data Preprocessing:*

1.1. Standardization or Scaling data :
Standardization is an effective method for making data more comparable and easier to interpret, especially when dealing with variables measured on different scales or in different units. Standardization improves machine learning model accuracy by lowering the influence of outliers and making the model less sensitive to feature scale [4].

2. *Statistical methods*

2.1. Principal Component Analysis:
Principal Component Analysis (PCA) is a statistical technique extensively used in machine learning for dimensionality reduction. By transforming the original set of variables into a new set of variables known as principal components, PCA assists in identifying the underlying structure in the data. The main components are ranked by the amount of variance they explain in the original data.PCA is useful in machine learning because it can aid in the reduction of the number of variables in a dataset. This can make working with data easier and increase the accuracy of machine learning models. By lowering the dimensionality of the data, PCA can also help to overcome the curse of dimensionality, which refers to the difficulties of analyzing large amounts of data [6].

2.2. Multiple Linear Regression
Multiple linear regression is a statistical approach for modeling the connection between one or more independent variables and a dependent variable. The purpose is to estimate the coefficients of the regression equation to discover the linear relationship that best reflects the data. The coefficients indicate the slope and intercept of the line, and they are used to forecast the value of the dependent variable based on the independent variables' values. Linear regression is a straightforward and extensively used statistical and machine learning method, notably for prediction and forecasting. It is frequently used to examine the relationship between variables in a data set and to forecast future observations [7].

3. *Evaluation metrics:*

3.1. R squared value:

The coefficient of determination, often known as R-squared, is a statistical metric that quantifies the proportion of the variance in the dependent variable that can be explained by the independent variables in a regression model. It reveals how well the model fits the data. R-squared values vary from 0 to 1, with a greater value near 1 indicating a better fit [8].

3.2. RMSE

RMSE is an abbreviation for Root Mean Squared Error, which is a statistic used to calculate the average difference between predicted and actual values in a regression problem. The lower the RMSE value, the better the model's performance.The RMSE value can be used to assess the accuracy of a regression model because it measures how well the model fits the data [9].

Dataset Description:

The soil moisture and temperature dataset we are using was measured in a five-day field campaign that took place in May 2017 in Karlsruhe, Germany and lasted for two weeks. The soil sample consists of bare soil without any vegetation. The dataset comprises 679 data records captured using a Cubert UHD 285 camera. Each record contains 125 hyperspectral bands that range from 450 nm to 950 nm. Additionally, each record has readings of soil moisture and temperature levels that allow for predictive analysis using regression based models [3].

Data column name	**Data type**	**Description**
soil_moisture	numeric	Response variable that quantifies moisture content in soil.
soil_temperature	numeric	Response variable that quantifies the temperature reading of the soil
454 - 634 (125 columns)	numeric	Numerous spectral bands with numeric values for radiation emitted within the

		range of the spectral band

Fig 1. Table of feature(s) description

Data cleaning and Preprocessing:

The dataset was thoroughly reviewed for inconsistencies such as null values, redundant columns, upon inspection no invalid values were found, although there were columns not essential to our analysis such as datetime and index that were removed from the dataset.

Results

Experimental Setup

In this study, the experimental setup consisted of creating two datasets from a single soil characteristics dataset. The first dataset had soil moisture as the response variable, while the second dataset had soil temperature as the response variable. The independent features for both datasets were the same, namely the hyperspectral measurement columns.To evaluate the performance of the models, both datasets were split into 80% training and 20% testing sets. Multiple linear regression models were then trained on each dataset separately, considering the 125 hyperspectral columns as independent variables.After obtaining the metrics, specifically the root mean square error (RMSE) and R-squared, we performed principal component analysis (PCA) on both datasets. PCA was applied to reduce the dimensionality of the hyperspectral measurement columns.To determine the optimal number of principal components, we conducted a loop that iterated from 1 to the length of the number of principal component columns generated by PCA. For each value of n components, we trained a linear regression model using the limited set of feature columns and computed the RMSE between the predictions and the test set.The RMSE values obtained for different numbers of principal components were then plotted to visualize the relationship between the number of components included and the resulting RMSE.This experimental setup allowed us to compare the performance of multiple linear regression models on the soil moisture and soil temperature datasets, assess the impact of dimensionality reduction using PCA, and identify the best RMSE achieved by varying the number of principal components used in the linear regression models.

Results from base model Multiple linear regression

Model	Response variable			
	Soil moisture		Soil temperature	
	RMSE	R squared	RMSE	R squared
Linear Regression	0.386	0.855	0.473	0.790
Linear Regression + PCA	0.334 (40)	0.888 (40)	0.383 (38)	0.852 (38)

Fig 2. Table of multiple linear regression results

Results of combining Principal Component Analysis with Multiple linear regression

Soil Moisture

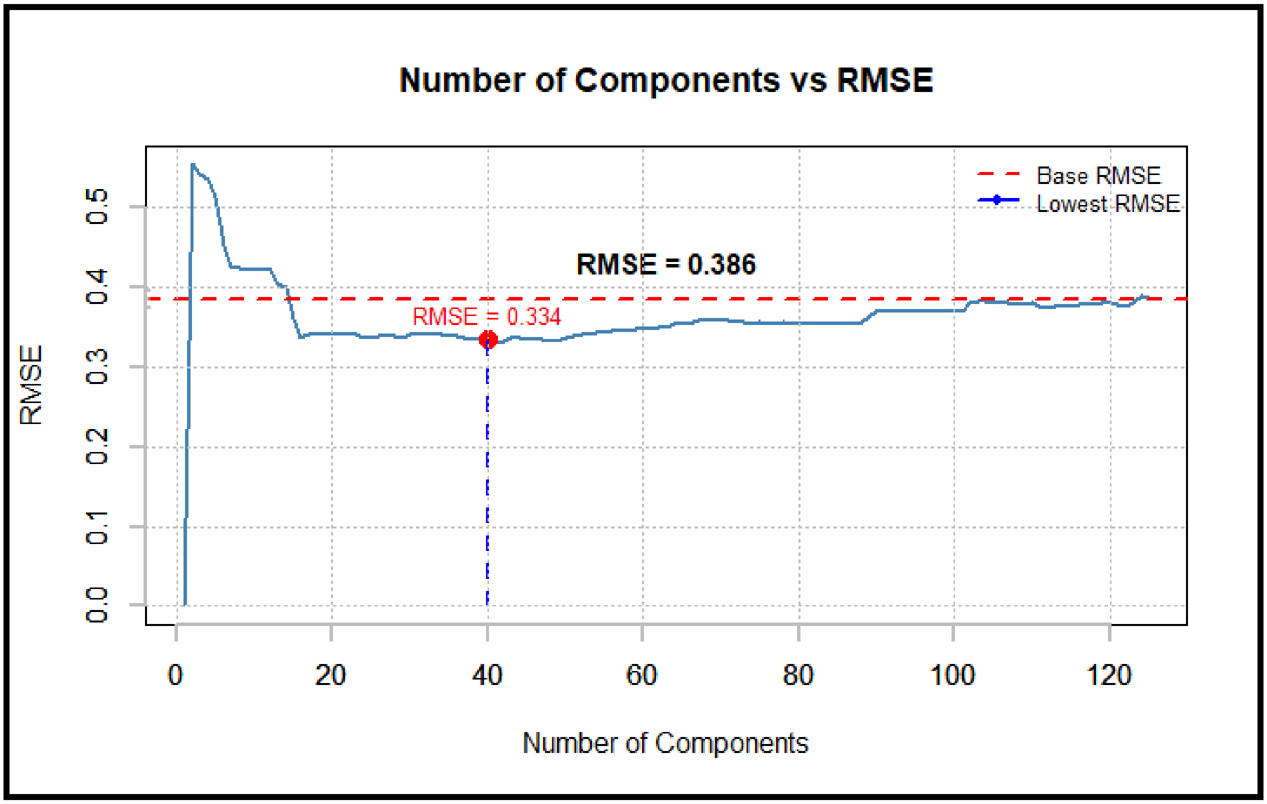


Fig 3. RMSE Vs Number of PCA components plot (soil moisture data)

Soil temperature

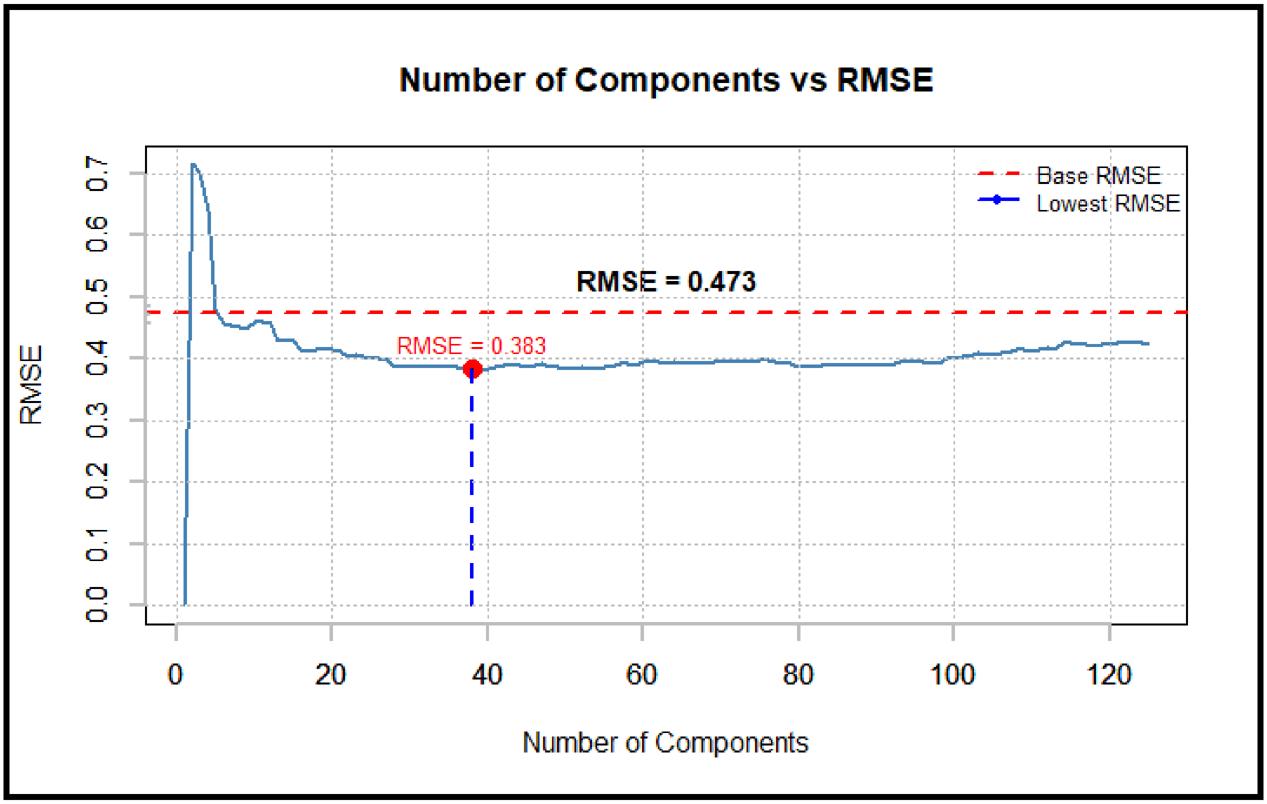


Fig 4. RMSE Vs Number of PCA components plot (soil temperature data)

Discussion

Interpretation of results

The results obtained after applying PCA with 40 components for predicting soil moisture and 38 components for predicting soil temperature individually, show an improvement in both the RMSE and R-squared values compared to their respective multiple linear regressions without dimensionality reduction. The RMSE value decreased from 0.386 to 0.334 for the case of soil moisture and decreased from 0.473 to 0.383 for the case of soil temperature, which indicates a reduction in the average prediction error of both the models. A lower RMSE suggests that the predictions are closer to the actual values, indicating improved accuracy.Additionally, the R-squared value increased from 0.855 to 0.888 for the case of soil moisture , and 0.790 to 0.852 for the case of soil temperature . R-squared represents the proportion of the variance in the dependent variable that is explained by the independent variables. An R-squared value closer to 1 indicates a better fit of the model to the data. Therefore, the increase in the R-squared value suggests that the models with PCA are able to explain more of the variance in the dependent variable, indicating a better overall fit.Overall, the results indicate that applying PCA for dimensionality reduction has led to improved performance in the regression models. The reduced number of components captured by PCA seems to be capturing more meaningful information and improving the model's ability to make accurate predictions. Additionally, the ANOVA test for both use cases (soil moisture and soil temperature), yielded the following results: a) testing models with and without pca for soil moisture, a p-value of 0.003 < 0.05 was obtained; b) testing models with and without pca, for soil temperature a p-value of 0.0002 < 0.05 was obtained, in both cases the p values indicate significant difference of model performance (See Appendix B) coupled with previous results this indicates we were able to obtain better models than baseline (without PCA) , that had significant difference in performance for the better. There were other notable graphs and plots generated for the preliminary analysis including scree plot, biplot (for PCA) (see Appendix A), summary reports for regression models (See Appendix C) ,and finally, residual plots for the regression models (See Appendix D).

Limitations and Assumptions

Linear regression assumes that the independent variables and the dependent variable have a linear relationship. It is critical to determine whether this assumption holds true for the given data, as non-linear correlations may be overlooked by the model.Linear regression is based on the assumption that the observations are independent of one another. If there are dependencies or correlations among the data points, this assumption may be violated, resulting in skewed or incorrect results [10]. While using PCA for dimensionality reduction might help to simplify the model and improve interpretability, it can also cause information loss. The fewer components may not capture all of the significant information contained in the original variables, thus resulting in a loss of predictive value [11]. The measurements of the experiment are unique to the dataset and variables used. It is critical to assess the predictive model to different datasets or real-world circumstances. When applied to different datasets with varying properties, the model's performance may vary. The sample size may have an impact on the significance of the results. A bigger sample size is more likely to produce strong and representative results. It is critical to determine whether the sample size used in the experiment is sufficient for making relevant findings. Even though dimensionality reduction by PCA can help decrease multicollinearity, it is still necessary to determine whether there are any remaining high correlations among the independent variables. Multicollinearity can have an effect on the interpretation of the coefficients as well as the model's stability [12].

Future works

Given the multiple linear regression assumption of linearity, researching non-linear regression models, such as polynomial regression or nonparametric regression techniques, can be a fruitful direction. These models are capable of capturing more complicated interactions between independent and dependent variables. Apply rigorous model selection approaches, such as cross-validation, to determine the model's generalizability and the ideal amount of components or variables to include in the regression model. This helps to avoid overfitting and delivers more trustworthy performance predictions.Validate the model with external datasets or gather new data to evaluate its performance in various circumstances. This aids in determining the model's generalizability and usefulness outside the unique dataset utilized in the experiment. Other dimensionality reduction techniques, such as independent component analysis (ICA) or factor analysis, should be investigated and compared in terms of capturing useful information and enhancing the regression model.Investigate more advanced machine learning methods, such as decision trees, random forests, and support vector regression, which can handle non-linear correlations, interactions, and complicated data patterns. When compared to ordinary linear regression, these models may yield superior predicted accuracy.By addressing these future studies, researchers can improve the regression model's knowledge and applicability, increase its performance, and overcome the limits and assumptions indicated in the current experiment.

Conclusion

In conclusion, our experiment aimed to investigate the relationship between soil moisture and soil temperature with regards to various hyperspectral features. We performed multiple linear regression analysis and subsequently applied principal component analysis (PCA) for dimensionality reduction to improve the predictive performance of the regression models.
The initial multiple linear regression model achieved an RMSE of 0.386 and an R-squared value of 0.855 for soil moisture prediction, and an RMSE of 0.473 and an R-squared value of 0.790 for soil temperature prediction , indicating a reasonably good fit to the data. However, to address the high dimensionality and potential multicollinearity among the predictors, we employed PCA to reduce the number of features to 40 components for the soil moisture case and 38 for the case of soil temperature.After applying PCA, we observed a notable improvement in the regression model's performance. The reduced models achieved an RMSE of 0.334 and an R-squared value of 0.888 for soil moisture and an RMSE of 0.383 and an R-squared value of 0.852. This reduction in RMSE and increase in R-squared indicate that the PCA-based

dimensionality reduction effectively captured the essential information in the hyperspectral features, leading to a more accurate and predictive regression models.These findings highlight the usefulness of PCA in reducing dimensionality and selecting informative components for regression analysis. By retaining the most significant components, we were able to capture the essential variability in the data and improve the model's predictive ability.It is important to note that the experiment had some limitations. The assumption of linearity in the multiple linear regression model may not hold in all cases, and there could be other factors influencing soil moisture and soil temperature that were not included in our analysis. Additionally, the experiment relied on a specific dataset, and the findings may not generalize to other soil types or geographical locations.Further research should explore non-linear regression models, incorporate additional variables, and validate the model on external datasets. Additionally, investigating alternative dimensionality reduction techniques and advanced machine learning algorithms can contribute to more robust and accurate models.Overall, our experiment demonstrates the effectiveness of PCA in improving the predictive performance of a multiple linear regression model for soil moisture and soil temperature prediction using hyperspectral features. These findings have implications for soil science, agriculture, and environmental monitoring, providing insights into the relationship between soil moisture and spectral characteristics for improved understanding and management of soil systems.

References

[1] S. L. Ustin et al., "Hyperspectral remote sensing for ecosystem science and natural resource management," in Ecosystems, vol. 12, no. 3, pp. 125-136, May 2009.

[2] T. M. Lillesand, R. W. Kiefer, and J. W. Chipman, "Remote Sensing and Image Interpretation," Wiley, 2014.

[3] Riese, F.M.; Keller, S. Hyperspectral benchmark dataset on soil moisture. 2018. Available online: https://zenodo.org/record/1227837#.XfnK2vx5vIV (accessed on 7 December 2019).

[4] Contributors to Wikimedia projects, “Standard score,” *Wikipedia*, Mar. 29, 2023. https://en.wikipedia.org/wiki/Standard_score#Standardizing_in_mathematical_statistics (accessed Apr. 27, 2023).

[5] Contributors to Wikimedia projects, “Pearson correlation coefficient,” Wikipedia, Mar. 22, 2023. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient (accessed Apr. 27, 2023).

[6] Contributors to Wikimedia projects, “Principal component analysis,” *Wikipedia*, Apr. 18, 2023. https://en.wikipedia.org/wiki/Principal_component_analysis (accessed Apr. 27, 2023).

[7] Contributors to Wikimedia projects, “Linear regression,” *Wikipedia*, Mar. 25, 2023. https://en.wikipedia.org/wiki/Linear_regression (accessed Apr. 27, 2023).

[8] Contributors to Wikimedia projects, “Coefficient of determination,” *Wikipedia*, Apr. 04, 2023. https://en.wikipedia.org/wiki/Coefficient_of_determination (accessed Apr. 27, 2023).

[9] Contributors to Wikimedia projects, “Root-mean-square deviation,” *Wikipedia*, Mar. 20, 2023. https://en.wikipedia.org/wiki/Root-mean-square_deviation (accessed Apr. 27, 2023).

[10] “Simple Linear Regression.” https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html (accessed Apr. 27, 2023).

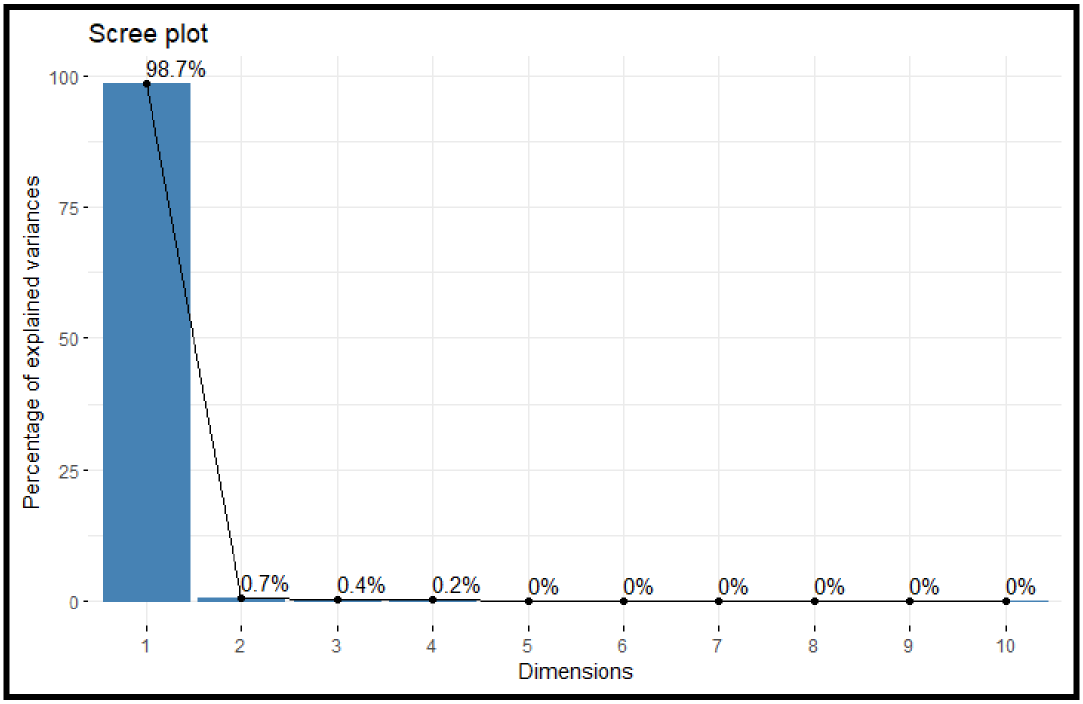
[11] B. C. Geiger and G. Kubin, “Relative information loss in the PCA,” in 2012 IEEE Information Theory Workshop, Sep. 2012. Accessed: Apr. 27, 2023. [Online]. Available: http://dx.doi.org/10.1109/itw.2012.6404738

[12] S.Q. Lafi, J.B. Kaneene,An explanation of the use of principal-components analysis to detect and correct for multicollinearity,Preventive Veterinary Medicine,Volume 13, Issue 4, 1992, Pages 261-275, ISSN 0167-5877, https://doi.org/10.1016/0167-5877(92)90041-D.

Appendices

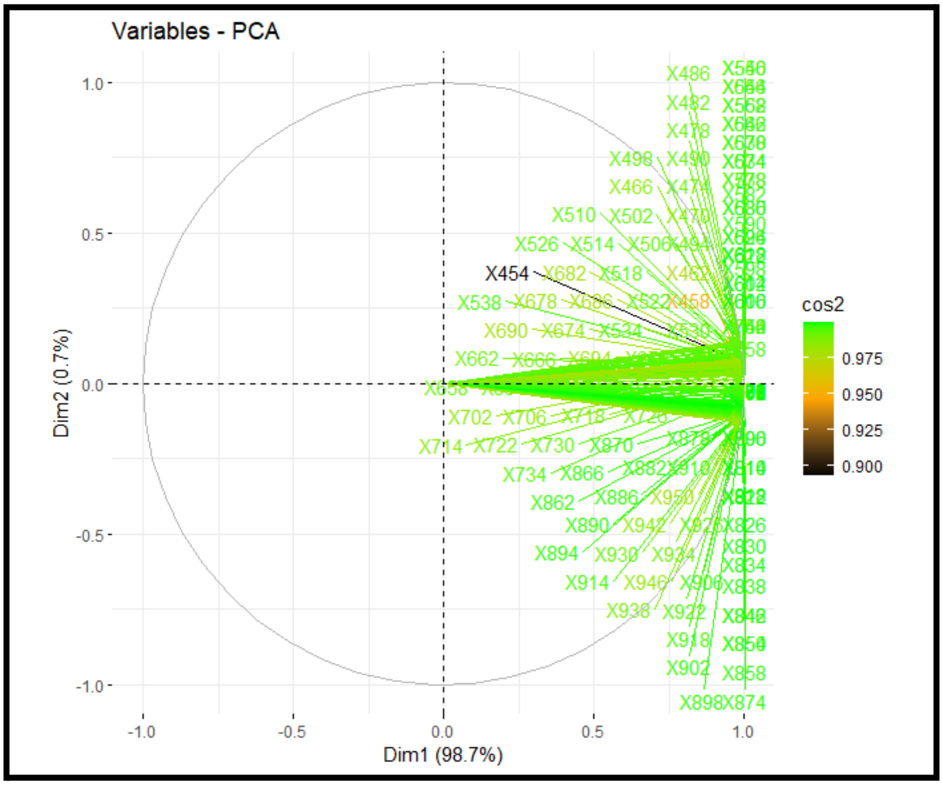
Appendix A. Plots for Principal Component Analysis (PCA)

Scree plot for PCA



Note: First component is able to explain 98.7% of variance in data.

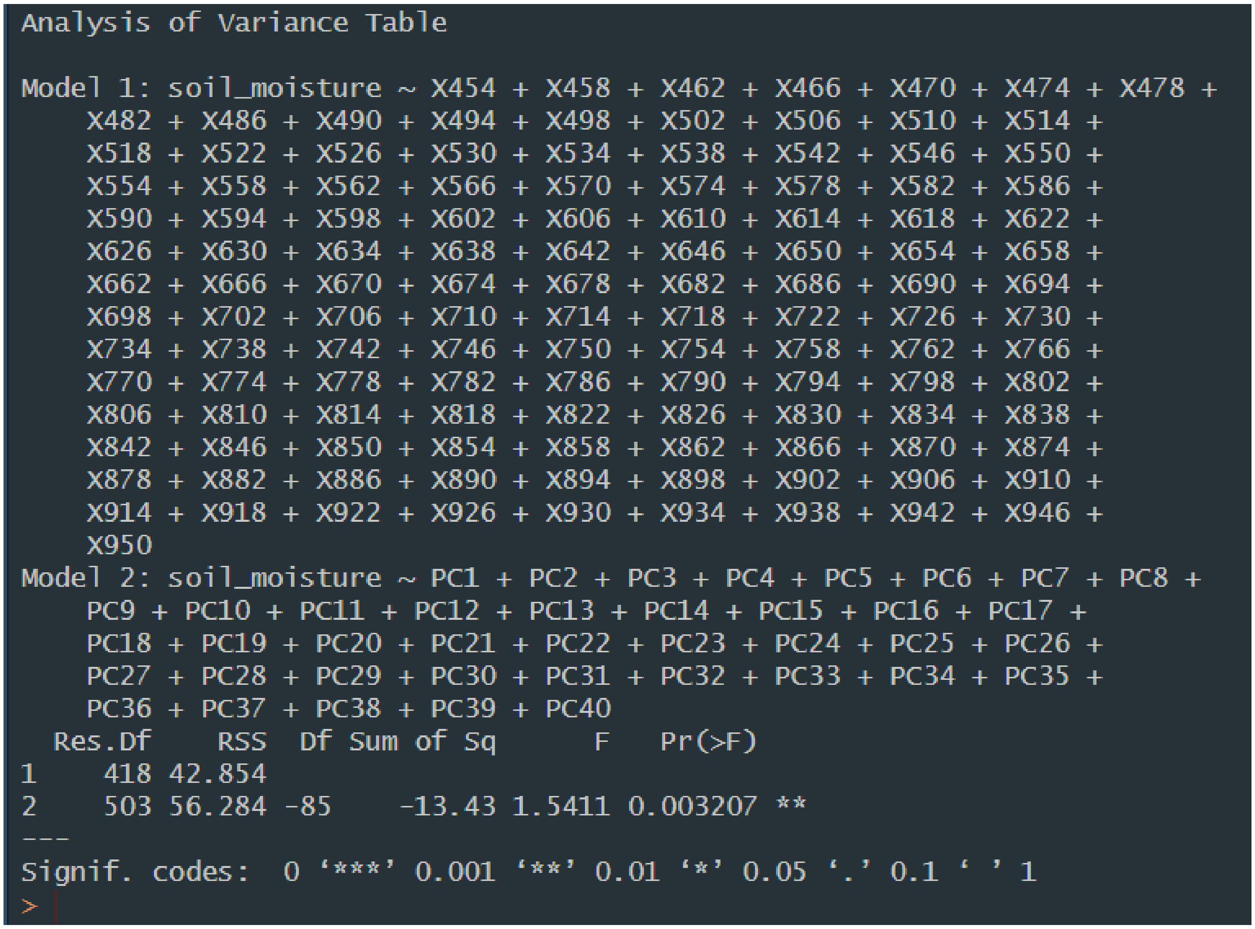
Bi Plot for PCA



Appendix B. Results of ANOVA test on models

<u>Case 1: Soil moisture prediction</u>

Model 1: Multiple linear regression model without PCA
Model 2 : Multiple linear regression model with PCA (40 components)

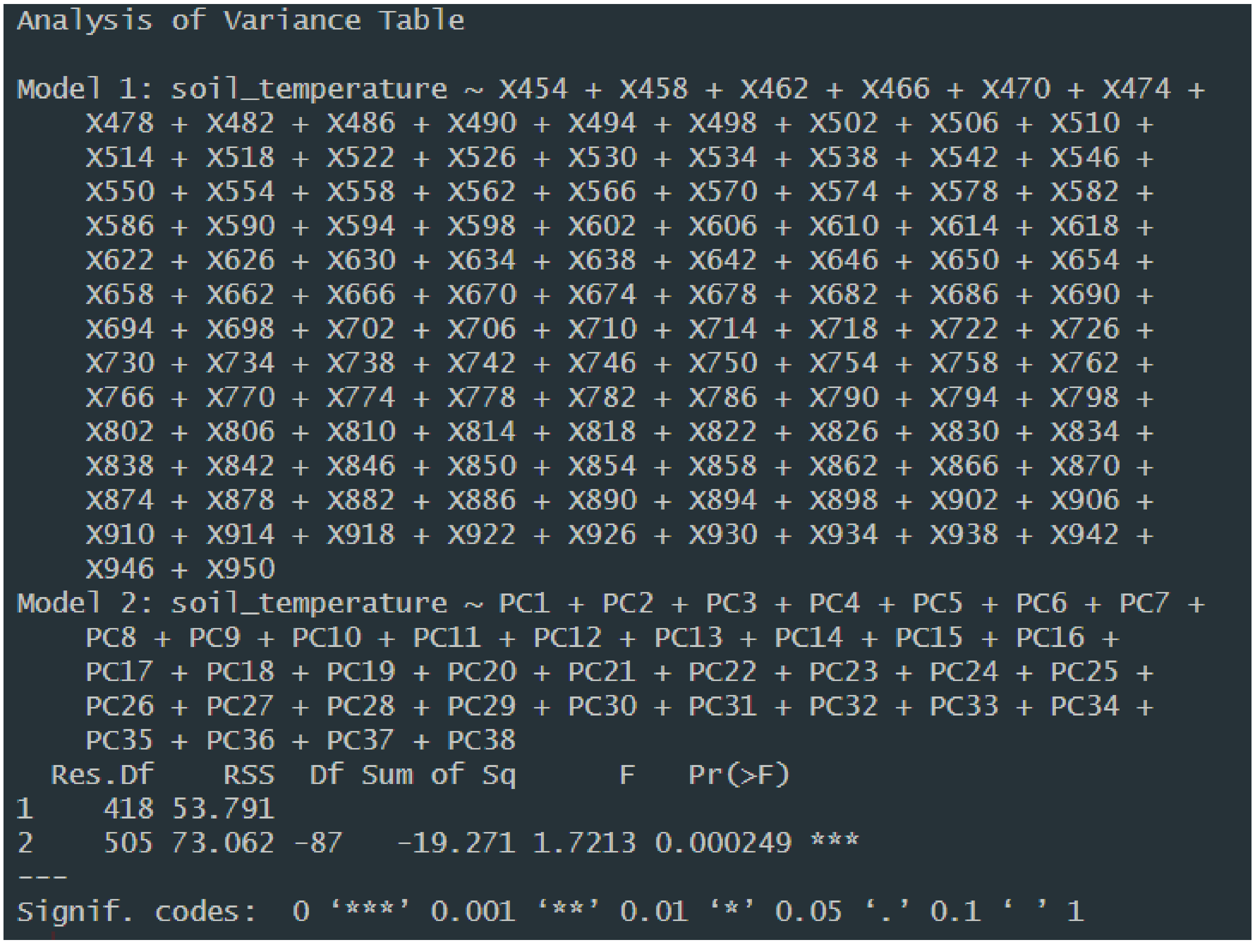


```
Analysis of Variance Table

Model 1: soil_moisture ~ X454 + X458 + X462 + X466 + X470 + X474 + X478 +
    X482 + X486 + X490 + X494 + X498 + X502 + X506 + X510 + X514 +
    X518 + X522 + X526 + X530 + X534 + X538 + X542 + X546 + X550 +
    X554 + X558 + X562 + X566 + X570 + X574 + X578 + X582 + X586 +
    X590 + X594 + X598 + X602 + X606 + X610 + X614 + X618 + X622 +
    X626 + X630 + X634 + X638 + X642 + X646 + X650 + X654 + X658 +
    X662 + X666 + X670 + X674 + X678 + X682 + X686 + X690 + X694 +
    X698 + X702 + X706 + X710 + X714 + X718 + X722 + X726 + X730 +
    X734 + X738 + X742 + X746 + X750 + X754 + X758 + X762 + X766 +
    X770 + X774 + X778 + X782 + X786 + X790 + X794 + X798 + X802 +
    X806 + X810 + X814 + X818 + X822 + X826 + X830 + X834 + X838 +
    X842 + X846 + X850 + X854 + X858 + X862 + X866 + X870 + X874 +
    X878 + X882 + X886 + X890 + X894 + X898 + X902 + X906 + X910 +
    X914 + X918 + X922 + X926 + X930 + X934 + X938 + X942 + X946 +
    X950
Model 2: soil_moisture ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 +
    PC9 + PC10 + PC11 + PC12 + PC13 + PC14 + PC15 + PC16 + PC17 +
    PC18 + PC19 + PC20 + PC21 + PC22 + PC23 + PC24 + PC25 + PC26 +
    PC27 + PC28 + PC29 + PC30 + PC31 + PC32 + PC33 + PC34 + PC35 +
    PC36 + PC37 + PC38 + PC39 + PC40
  Res.Df    RSS  Df Sum of Sq      F   Pr(>F)
1    418 42.854
2    503 56.284 -85    -13.43 1.5411 0.003207 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

<u>Case 2: Soil temperature prediction</u>

Model 1 : Multiple linear regression model without PCA
Model 2 : Multiple linear regression with PCA (38 components)



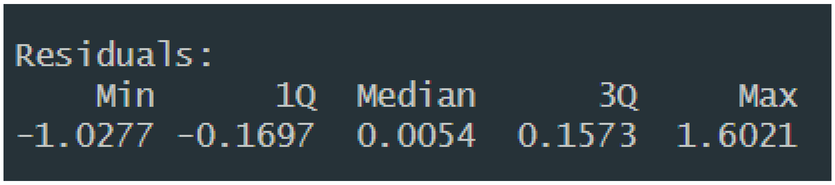
```
Analysis of Variance Table

Model 1: soil_temperature ~ X454 + X458 + X462 + X466 + X470 + X474 +
    X478 + X482 + X486 + X490 + X494 + X498 + X502 + X506 + X510 +
    X514 + X518 + X522 + X526 + X530 + X534 + X538 + X542 + X546 +
    X550 + X554 + X558 + X562 + X566 + X570 + X574 + X578 + X582 +
    X586 + X590 + X594 + X598 + X602 + X606 + X610 + X614 + X618 +
    X622 + X626 + X630 + X634 + X638 + X642 + X646 + X650 + X654 +
    X658 + X662 + X666 + X670 + X674 + X678 + X682 + X686 + X690 +
    X694 + X698 + X702 + X706 + X710 + X714 + X718 + X722 + X726 +
    X730 + X734 + X738 + X742 + X746 + X750 + X754 + X758 + X762 +
    X766 + X770 + X774 + X778 + X782 + X786 + X790 + X794 + X798 +
    X802 + X806 + X810 + X814 + X818 + X822 + X826 + X830 + X834 +
    X838 + X842 + X846 + X850 + X854 + X858 + X862 + X866 + X870 +
    X874 + X878 + X882 + X886 + X890 + X894 + X898 + X902 + X906 +
    X910 + X914 + X918 + X922 + X926 + X930 + X934 + X938 + X942 +
    X946 + X950
Model 2: soil_temperature ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 +
    PC8 + PC9 + PC10 + PC11 + PC12 + PC13 + PC14 + PC15 + PC16 +
    PC17 + PC18 + PC19 + PC20 + PC21 + PC22 + PC23 + PC24 + PC25 +
    PC26 + PC27 + PC28 + PC29 + PC30 + PC31 + PC32 + PC33 + PC34 +
    PC35 + PC36 + PC37 + PC38
  Res.Df    RSS  Df Sum of Sq      F   Pr(>F)
1    418 53.791
2    505 73.062 -87   -19.271 1.7213 0.000249 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

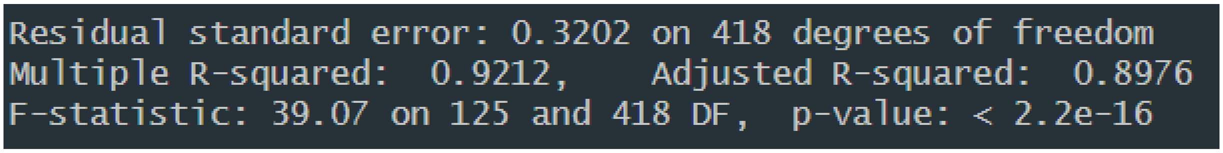
Appendix C. Summary Reports for linear regression models

Case 1: Soil moisture prediction

Multiple linear regression model without PCA

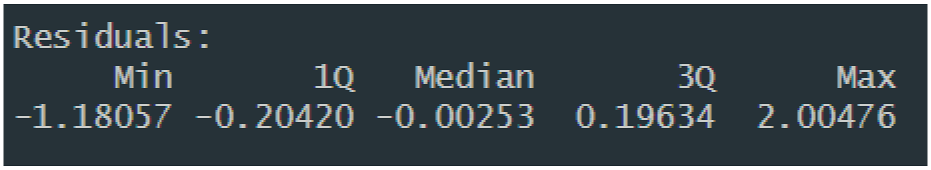


```
Residuals:
    Min      1Q  Median      3Q     Max
-1.0277 -0.1697  0.0054  0.1573  1.6021
```

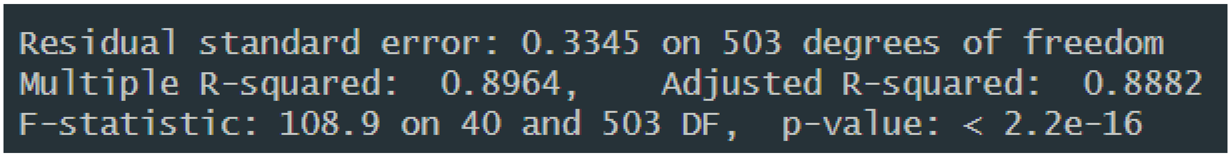


```
Residual standard error: 0.3202 on 418 degrees of freedom
Multiple R-squared:  0.9212,    Adjusted R-squared:  0.8976
F-statistic: 39.07 on 125 and 418 DF,  p-value: < 2.2e-16
```

Multiple linear regression model with PCA (40 components)



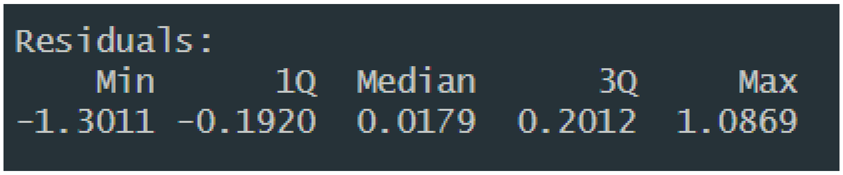
```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.18057 -0.20420 -0.00253  0.19634  2.00476 
```



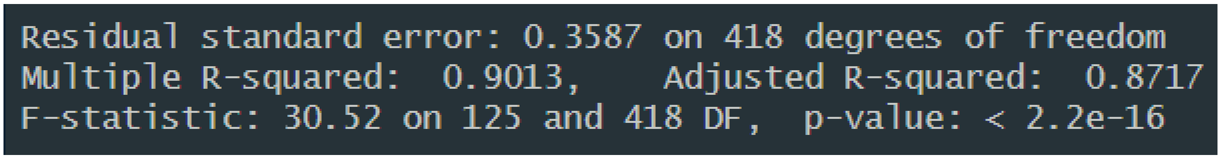
```
Residual standard error: 0.3345 on 503 degrees of freedom
Multiple R-squared:  0.8964,	Adjusted R-squared:  0.8882 
F-statistic: 108.9 on 40 and 503 DF,  p-value: < 2.2e-16
```

Case 2: Soil temperature prediction

Multiple linear regression without PCA

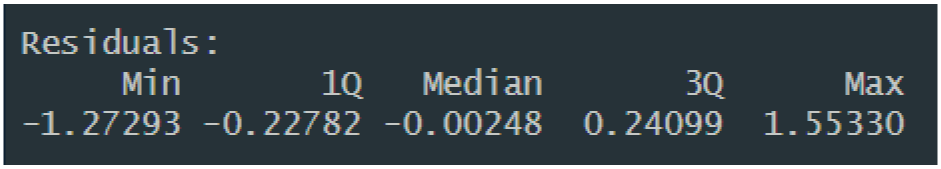


```
Residuals:
    Min      1Q  Median      3Q     Max 
-1.3011 -0.1920  0.0179  0.2012  1.0869 
```

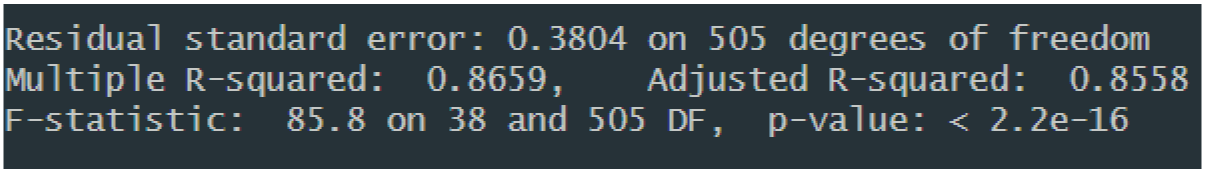


```
Residual standard error: 0.3587 on 418 degrees of freedom
Multiple R-squared:  0.9013,	Adjusted R-squared:  0.8717 
F-statistic: 30.52 on 125 and 418 DF,  p-value: < 2.2e-16
```

Multiple linear regression with PCA (38 components)



```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.27293 -0.22782 -0.00248  0.24099  1.55330 
```

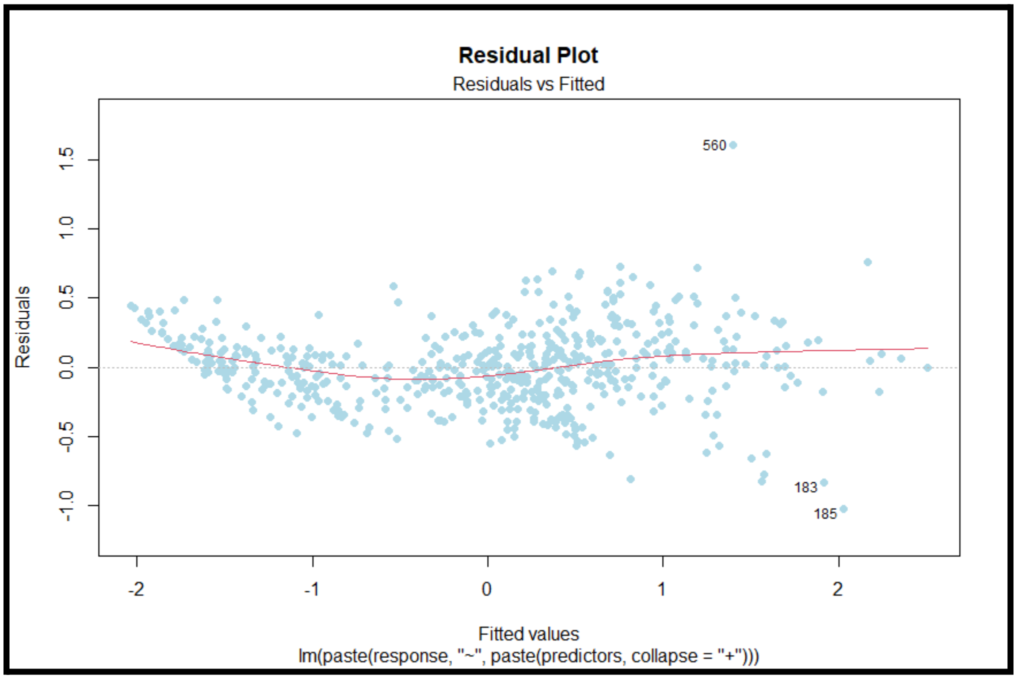


```
Residual standard error: 0.3804 on 505 degrees of freedom
Multiple R-squared:  0.8659,	Adjusted R-squared:  0.8558 
F-statistic:  85.8 on 38 and 505 DF,  p-value: < 2.2e-16
```

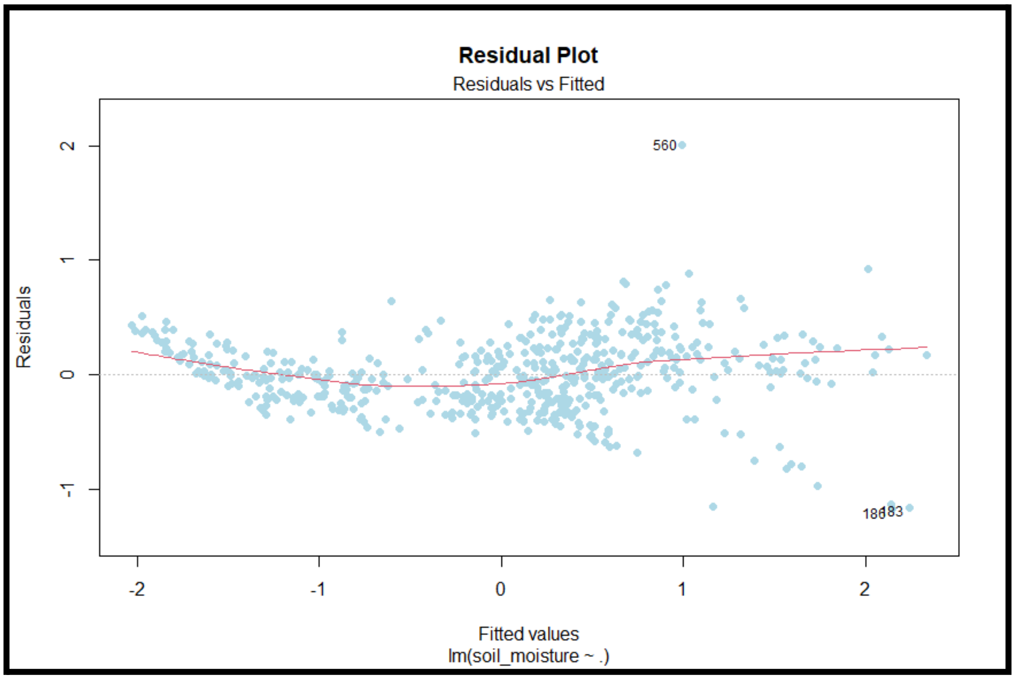
Appendix D. Residual Plots for Regression

Case 1: Soil moisture prediction

Multiple linear regression model without PCA

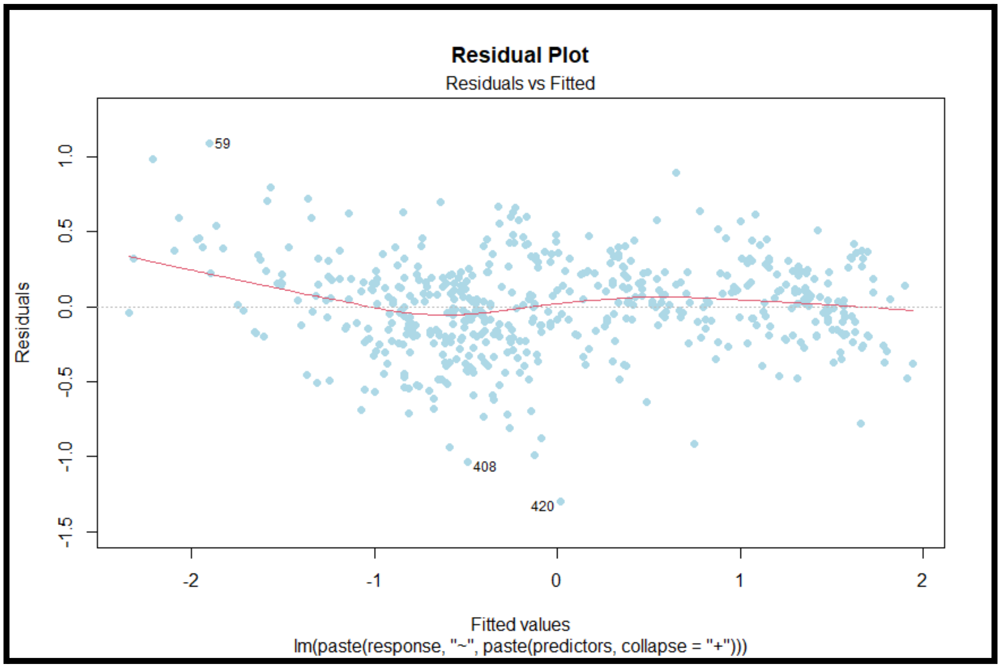


Multiple linear regression model with PCA (40 components)

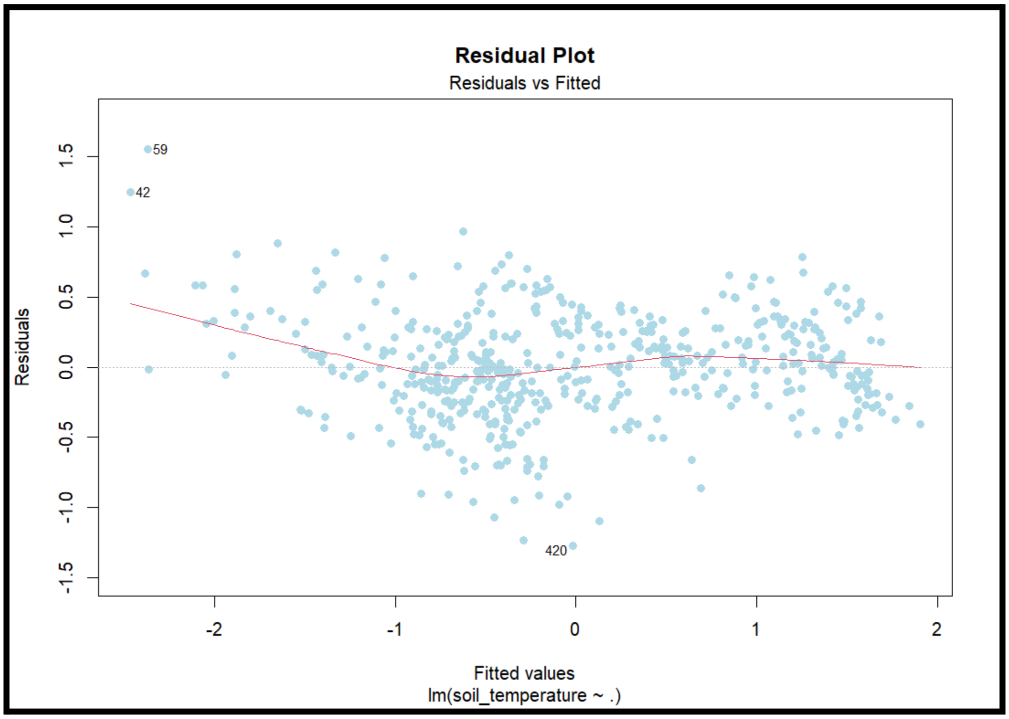


Case 2: Soil temperature prediction

Multiple linear regression without PCA



Multiple linear regression with PCA (38 components)


Residual Plot
Residuals vs Fitted
Residuals
Fitted values
lm(soil_temperature ~ .)