Appendix S

**Table S1.** The number of samples per country in the different soil datasets.

**ORIGINAL DATASET**

|  |  |
| --- | --- |
| **Country** | **N** |
| Argentina | 5 |
| Australia | 85 |
| Chile | 7 |
| Mexico | 8 |
| Morocco | 9 |
| Panama | 5 |
| China | 6 |
| Spain | 19 |
| Tunisia | 6 |
| United Kingdom | 11 |
| USA | 62 |

**INTERPOLATION**

|  |  |
| --- | --- |
| **Country** | **N** |
| Algeria | 2 |
| Argentina | 11 |
| Australia | 100 |
| Brazil | 1 |
| Chile | 11 |
| Libya | 1 |
| Mexico | 13 |
| Morocco | 12 |
| Panama | 5 |
| China | 10 |
| Russia | 2 |
| Spain | 23 |
| Tunisia | 10 |
| United Kingdom | 15 |
| USA | 122 |

**INTERPOLATION AND SMOTE**

|  |  |
| --- | --- |
| **Country** | **N** |
| Algeria | 2 |
| Argentina | 93 |
| Australia | 100 |
| Brazil | 1 |
| Chile | 66 |
| Libya | 1 |
| Mexico | 69 |
| Morocco | 56 |
| Panama | 42 |
| China | 70 |
| Russia | 2 |
| Spain | 51 |
| Tunisia | 70 |
| United Kingdom | 57 |
| USA | 122 |

**2X INTERPOLATION AND THEN SMOTE**

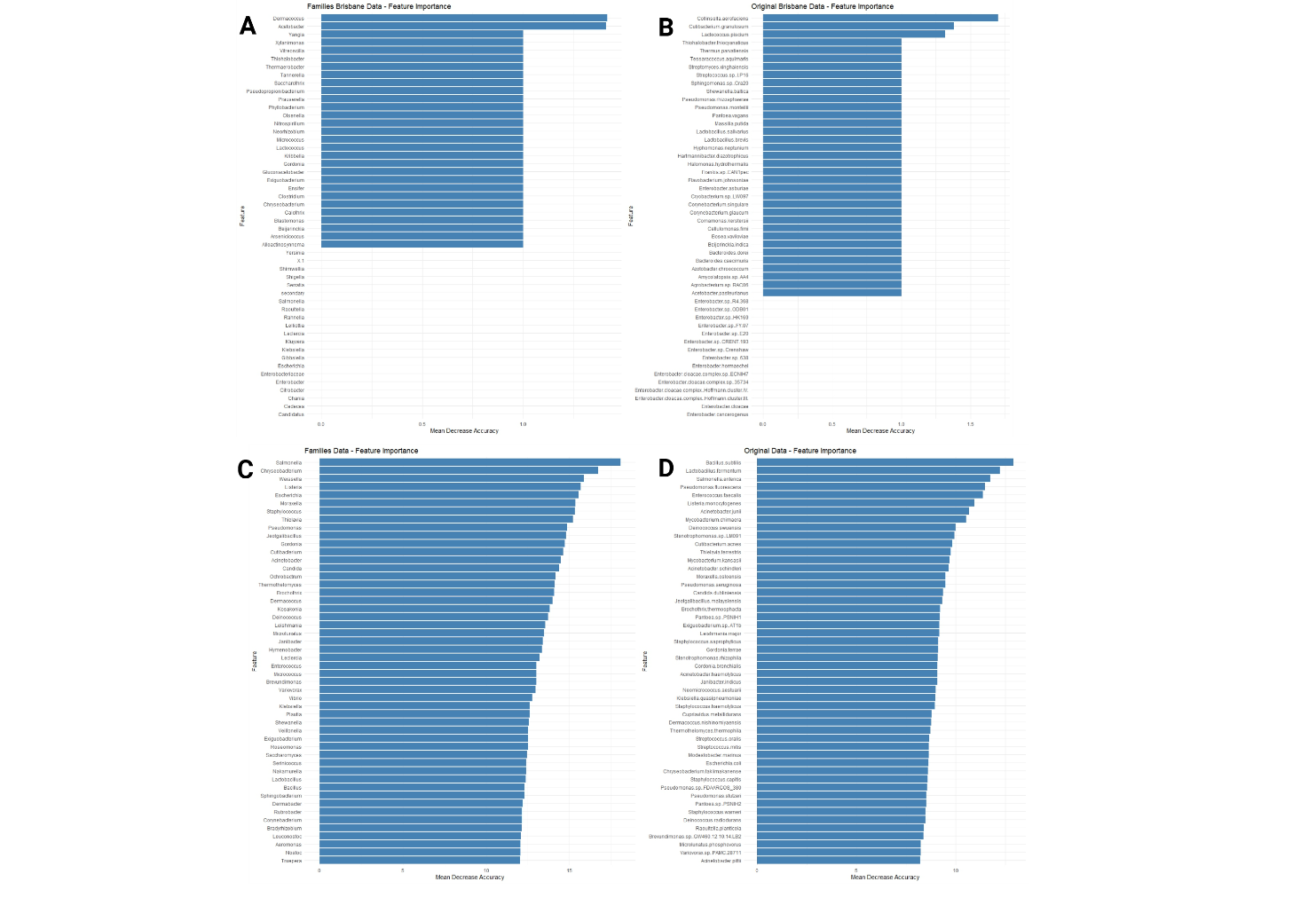
|  |  |
| --- | --- |
| **Country** | **N** |
| Algeria | 4 |
| Argentina | 144 |
| Australia | 115 |
| Brazil | 2 |
| Chile | 91 |
| Libya | 2 |
| Mexico | 95 |
| Morocco | 70 |
| Panama | 42 |
| China | 99 |
| Russia | 4 |
| Spain | 60 |
| Tunisia | 99 |
| United Kingdom | 73 |
| USA | 188 |

**Table S2.** Subway cities.

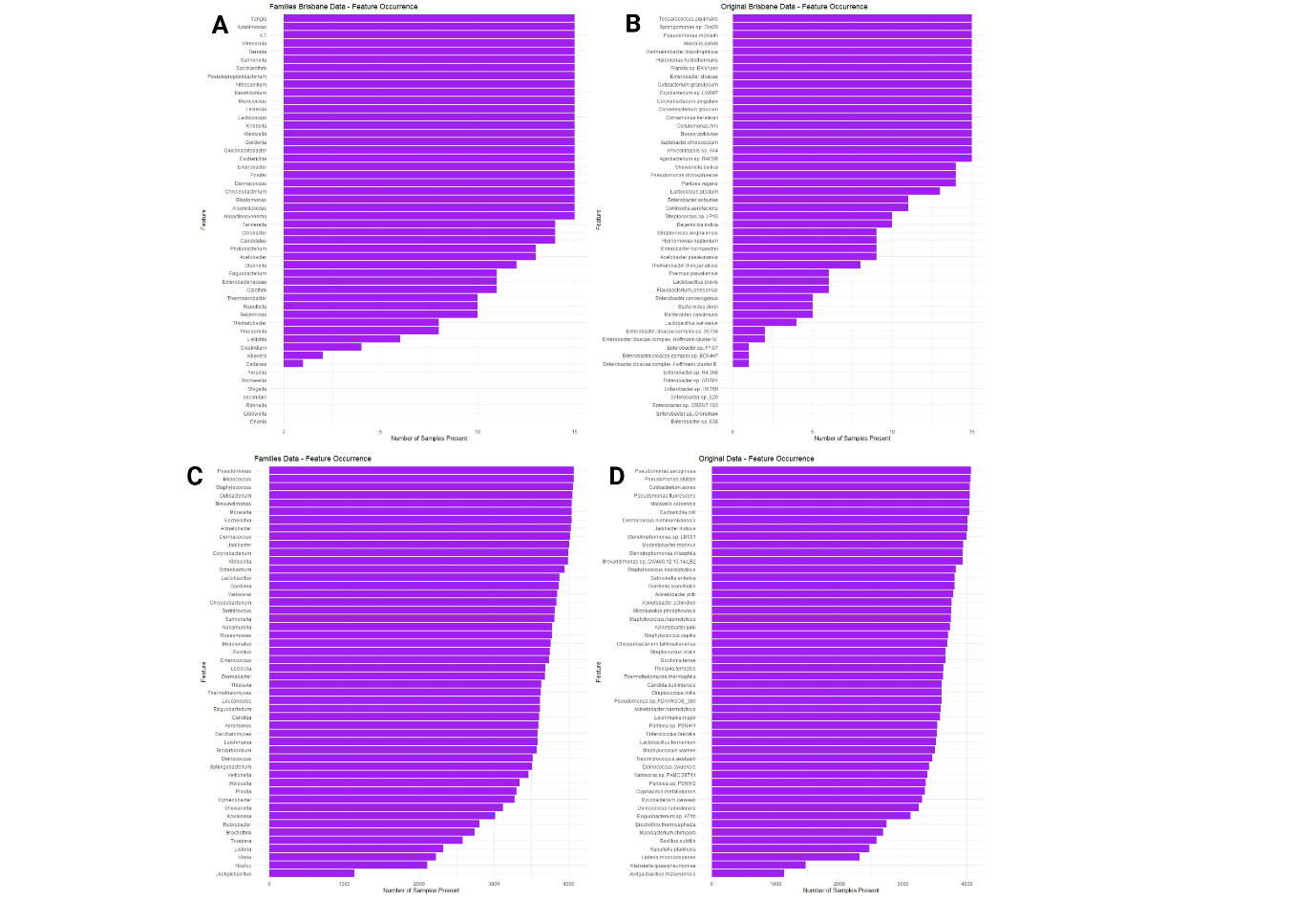
|  |  |
| --- | --- |
| **City** | **N** |
| Auckland | 15 |
| Baltimore | 13 |
| Barcelona | 125 |
| Berlin | 52 |
| Bogota | 15 |
| Brisbane | 15 |
| Denver | 72 |
| Doha | 73 |
| Fairbanks | 97 |
| Hamilton | 16 |
| Hanoi | 16 |
| Hong Kong | 793 |
| Ilorin | 271 |
| Kuala Lumur | 30 |
| Kyiv | 97 |
| Lisbon | 51 |
| London | 618 |
| Marseille | 16 |
| Minneapolis | 14 |
| Naples | 16 |
| New York | 507 |
| Offa | 25 |
| Oslo | 87 |
| Paris | 16 |
| Porto | 112 |
| Rio De Janeiro | 34 |
| Sacramento | 16 |
| San Fransisco | 28 |
| Santiago | 26 |
| Sao Paolo | 29 |
| Sendai | 29 |
| Seoul | 79 |
| Singapore | 186 |
| Sofia | 16 |
| Stockholm | 115 |
| Taipei | 94 |
| Tokyo | 152 |
| Vienna | 16 |
| Yamaguchi | 9 |
| Zurich | 79 |

**Table S3.** The number of features interpolated by each interpolation method.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Interpolation Method** | **N** | **Mean R-squared** | **Median R-squared** | **Min R-squared** | **Max R-squared** |
| GAM | 212 | 0.371 | 0.357 | 0.00266 | 0.789 |
| IDW | 213 | 0.487 | 0.494 | 0.0185 | 0.847 |
| Nearest Neighbour | 23 | 0.413 | 0.417 | 0.124 | 0.723 |
| TPS | 63 | 0.501 | 0.511 | 0.0376 | 0.829 |



**Figure S1.** Top 50 most important features for A) Brisbane in the species dataset. B) Brisbane in the original (phyla) dataset. C) In the complete species dataset. D) In the complete original dataset. The feature importance was quantified using a RFC. Each feature’s values are scrambled and the loss of classification accuracy is evaluated for the scrambling of each feature. The more accuracy that is lost, the more important the feature is.



**Figure S2.** Feature occurrence, sorted by importance for A) Brisbane in the species dataset. B) Brisbane in the original dataset. C) The complete species dataset. D) The complete original dataset.