STA5069Z Topic Ideas

Sian Wood - WDXSIA001

Table of contents

Redfin Data

This dataset originates from a study conducted along the Kars River by Keir Lynch in 2018. The aim of this study was to identify predictors of the presence and abundance of the redfin fish (Morch and Griffiths 2025).

The dataset contains 47 variables collected at 56 sites along the river. Two variables are binary, indicating whether any indigenous species were detected and indicating whether any non-indigenous species were detected at each site. The remaining variables are numeric. Six are count data for the number of fish of given species which were detected at each site. Variables relating to water quality, geographic features, substrates, debris, macrophyte presence, canopy and river depth and width are available. The variables relating to substrates, debris, macrophyte presence, canopy and river depth and width are values given as the proportion of the site with a given characteristic. The ranges of each of these variables can be seen in Table 1.

The research question I wish to answer using these data is "What are the predictors of the presence and abundance of six fish species at sites on the Kars River?". This would necessitate a multivariate analysis with at least six response variables and up to 41 explanatory variables. The answer to this question could help to identify aspects of the environment which may be advantageous or disadvantageous to the success of our indigenous fish species, as well as those which may be correlated with the over-abundance of non-indigenous species.

Table 1: The ranges of every variable available in the Redfin Dataset

	Minimum	Maximum
Sites	1.00	56.00
pH	4.20	10.25
EC	0.22	20.50
DO	43.87	126.40
Temp	11.70	27.87
Ammonia	0.00	0.94
Phosphorous	0.00	2.60
Nitrite	0.00	21.00
Nitrate	0.00	4.30
Total_Iron	0.00	3.16
Phosphonate	0.00	2.89
TDS	5.00	751.00
Inorg_Nitrogen	0.10	91.81
Elevation	58.00	306.00
Flow	0.00	0.90
Slope	0.03	36.86
Silt-Sand Substrate	0.00	1.00

Table 1: The ranges of every variable available in the Redfin Dataset

	Minimum	Maximum
Gravel Substrate	0.00	0.73
Cobble Substrate	0.00	0.03
Boulder Substrate	0.00	0.60
Bedrock Substrate	0.00	0.87
Woody Debris	0.00	0.70
No Woody Debris	0.30	1.00
Undercut Bank	0.00	0.60
No Undercut Bank	0.40	6.00
No Macrophytes	0.00	1.00
Scarce Macrophytes	0.00	1.00
Moderate Macrophytes	0.00	0.43
Abundant Macrophytes	0.00	0.47
Open Canopy	0.27	1.00
Partial Canopy	0.00	0.57
Closed Canopy	0.00	0.23
Shallow Water Depth (0 - 50)	0.00	1.00
Moderate Water Depth (51-100)	0.00	0.93
Deep Water Depth (100 - 180)	0.00	1.00
Very Deep Water Depth (<180)	0.00	0.90
Narrow River Width (0 - 3)	0.00	1.00
Moderate River Width (3 - 6)	0.00	1.00
Wide River Width (6 - 10)	0.00	1.00
Very Wide Width (<10)	0.00	1.00
Heuningnes Redfin	0.00	225.00
Cape Kurper	0.00	138.00
Cape Galaxias	0.00	46.00
Spotted Bass	0.00	5.00
Bluegill Sunfish	0.00	115.00
Common Carp	0.00	0.00
Native	0.00	1.00
Non-Native	0.00	1.00

Mtb Data

This dataset includes metadata regarding the participants of an Adolescent Cohort Study, as well as cellular responses which were measured after stimulating Peripheral blood mononuclear cells (PBMCs) of the study participants (Lloyd, Mpande, et al. 2021). PBMCs were collected

from the same participants at 6-monthly intervals, adding a longitudinal aspect to these data. This could be incorporated into the analysis, or the data could be filtered for a single time point.

A multidimensional analysis has already been conducted on this dataset by Lloyd, Steigler, et al. (2021).

The research question I wish to answer using these data is "Are cellular responses to the stimulation of PBMCs effective predictors of QFT value and status?". Other potential response variables could be BMI and type of stimulation used. QFT value provides an indication of whether an individual has experienced recent TB infection. The ability to identify individuals who have recently been infected with TB will allow healthcare providers to provide targeted treatment for the prevention of progression to TB disease.

The data contain information from 57 participants, most of whom had blood samples collected and tested at four time points. Each blood sample was stimulated using four different methods, recorded as 'Esp', 'Mtb-lysate', 'E6C10' and 'SEB'. A final part of each sample was left unstimulated.

Each observation in the TB Dataset is made up of the participant ID, their QFT status, the month post recruitment at which the sample was taken (0, 6, 12, 18), the form of stimulation, the cell population (CD4 or CD8) and the cellular responses. The ranges of each of these responses are detailed in Table 2. A separate dataset containing the participants' metadata (e.g. participant ID, age, BMI, exposure to household contacts etc.) is also available.

Table 2: The ranges of every variable available in the TB Dataset

	Minimum	Maximum
IL2+CD107+CD154+IFNg+TNF+	0	43
IL2+CD107+CD154+IFNg+TNF-	0	5
IL2+CD107+CD154+IFNg-TNF+	0	53
IL2+CD107+CD154+IFNg-TNF-	0	41
IL2+CD107+CD154-IFNg+TNF+	0	192
IL2+CD107+CD154-IFNg+TNF-	0	52
IL2+CD107+CD154-IFNg-TNF+	0	9
IL2+CD107+CD154-IFNg-TNF-	0	30
IL2+CD107-CD154+IFNg+TNF+	0	8319
IL2+CD107-CD154+IFNg+TNF-	0	1690
IL2+CD107-CD154+IFNg-TNF+	0	12981
IL2+CD107-CD154+IFNg-TNF-	0	10736
IL2+CD107-CD154-IFNg+TNF+	0	3370
IL2+CD107-CD154-IFNg+TNF-	0	1299
IL2+CD107-CD154-IFNg-TNF+	0	2789
IL2+CD107-CD154-IFNg-TNF-	0	10050

Table 2: The ranges of every variable available in the TB Dataset

	Minimum	Maximum
IL2-CD107+CD154+IFNg+TNF+	0	864
IL2-CD107+CD154+IFNg+TNF-	0	115
IL2-CD107+CD154+IFNg-TNF+	0	88
IL2-CD107+CD154+IFNg-TNF-	0	617
IL2-CD107+CD154-IFNg+TNF+	0	20929
IL2-CD107+CD154-IFNg+TNF-	0	8891
IL2-CD107+CD154-IFNg-TNF+	0	10327
IL2-CD107+CD154-IFNg-TNF-	0	21298
IL2-CD107-CD154+IFNg+TNF+	0	14534
IL2-CD107-CD154+IFNg+TNF-	0	7237
IL2-CD107-CD154+IFNg-TNF+	0	24650
IL2-CD107-CD154+IFNg-TNF-	0	82990
IL2-CD107-CD154-IFNg+TNF+	0	40146
IL2-CD107-CD154-IFNg+TNF-	0	20168
IL2-CD107-CD154-IFNg-TNF+	0	21591
IL2-CD107-CD154-IFNg-TNF-	0	915000

References

Lloyd, Tessa, Cheleka Mpande, Elisa Nemes, Thomas Scriba, Virginie Rozot, Timothy Reid, Mark Hatherill, et al. 2021. "Innate and adaptive immune cell responses to recent M.tb infection," February. https://doi.org/10.25375/uct.13693699.v3.

Lloyd, Tessa, Pia Steigler, Cheleka A. M. Mpande, Virginie Rozot, Boitumelo Mosito, Constance Shreuder, Timothy D. Reid, et al. 2021. "Multidimensional Analysis of Immune Response Identified Biomarkers of Recent Mycobacterium Tuberculosis Infection." *medRxiv*. https://doi.org/10.1101/2021.01.27.21250605.

Morch, Casper, and Charles Griffiths. 2025. "Between a Bass and a Hard Place: The Fragmented Distribution of an Endangered Redfin in the Heuningnes River System of the Cape Fold Ecoregion," January. https://doi.org/10.25375/uct.25195838.v1.