# Capstone: Examining environmental elements relative to mortality sites in Arizona, USA

December 2024

## Introduction

Between 2011 and 2020, the US Border Patrol averaged 80,082 annual apprehensions of undocumented immigrants (UDIs) in the Tucson Sector of the USA–Mexico border (Martinez et al. 2021). This 262-mile border zone, stretching from Yuma County, AZ, to the New Mexico border, has since seen sharp increases in UDI encounters, with 173,400 reported in 2021 and 230,200 in 2022 (US Customs and Border Protection 2022a, b). This rise is concerning, as the Tucson Sector is recognized as an increasingly dangerous migration route. Despite fluctuations in border crossings, the ratio of deaths to crossings has grown (Boyce et al. 2019). Of the more than 3,300 UDI remains recovered in this sector between 1990 and 2020, 87% are linked to environmental exposure in the Sonora Desert, marked by extreme heat, rugged terrain, and scarce surface water (Martinez et al. 2021).

Since environmental exposure is the leading cause of UDI mortality in this sector (Figure 1; Martinez et al. 2021) and remains a central concern for Border Patrol, I examine how environmental features predict the location of UDI remains in a simplified and general manner. Using data from the Pima County Office of the Medical Examiner (PCOME) that maps UDI remains, I build on research emphasizing spatial data within the sector (Giordano and Spradley 2017).

Terrain significantly impacts surface and atmospheric heat and water distribution at local and regional levels, creating environmental stresses that challenge human physiology. Terrain indices—quantitative measures of Earth's surface derived from satellite-based digital elevation models (DEMs)—are widely applied in Earth sciences (Foster et al. 2019). Since terrain can predict elements like solar radiation and ambient air temperature, it provides a valuable lens to assess the link between the sector's extreme conditions and UDI mortality.

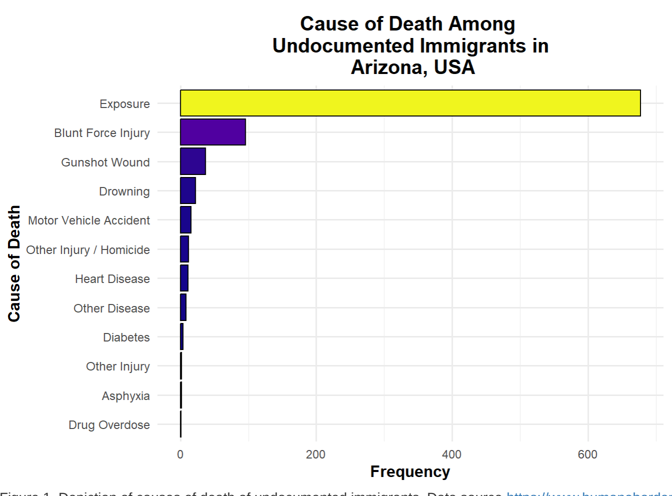
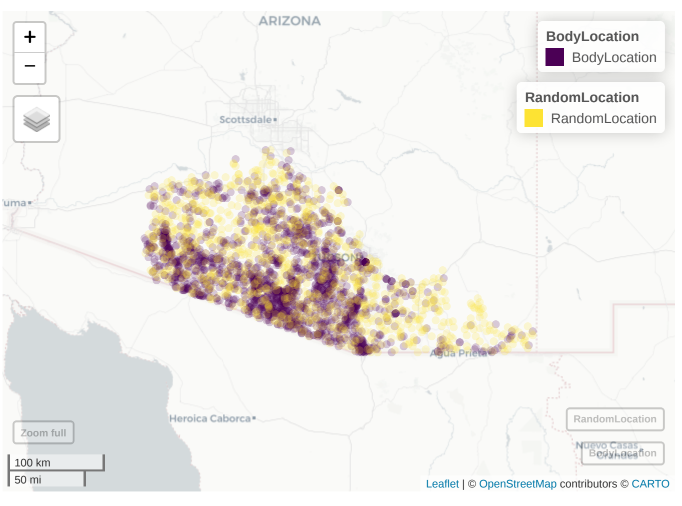



Figure 1. Depiction of causes of death of undocumented immigrants. Data source https://www.humaneborders.org/)

## Analysis

I defined the area using UDI mortality data from the Pima County Office of the Medical Examiner (n = 1,877; Humane Borders, Inc. 2022). This dataset includes mortality sites in Pima, Santa Cruz, Cochise, Graham, and La Paz Counties from 1981 to 2023. I also generated 998 random points for later prediction purposes.




My environmental variables were generated via a terrain model (a 10mx10m-pixel digital elevation model) downloaded from the National Map Viewer (https://apps.nationalmap.gov/downloader/), which encompasses the sector's topography and relief, and generates subsequent derivatives, which include for each pixel: slope, elevation, aspect, ruggedness, proportion of the daily amount of sun received at 9:00AM, 2:00PM, and 6:00PM, the duration of direct solar received in hours, and the amount of diffuse radiation, direct radiation, and total solar radiation received during the longest day of the year in WH/m^2. Also included were an estimation of vegetation density using MSAVI2 (Modified Soil-Adjusted Vegetation Index II) and total aridity via potential evapotranspiration from https://www.climatologylab.org/gridmet.html)

Investigating patterns in the data via a basic correlation plot to inform upcoming generalized linear regression models (GLMs) and avoid multicollinearity.

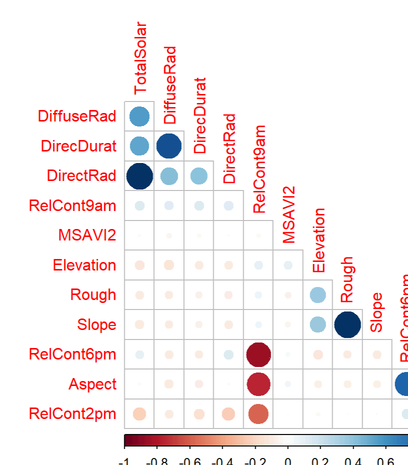



Figure 3. Correlation plot of continuous variables in the dataset

Next, I used AIC models to examine the relative quality of different generalized linear models (GLMs) by comparing their goodness-of-fit while penalizing for model complexity, thereby helping to identify the most parsimonious model that can be used to predict the difference between a mortality location and a random location.

As depicted in the AICc table below, the model (labelled AZ_Mod.9) constructed as: PointType ~ DirectSolarDuration + I(DirectSolarDuration^2) + TerrainRoughness yielded the best fit.

```
##
## Model selection based on AICc:
##
##            K    AICc Delta_AICc AICcWt Cum.Wt       LL
## AZ_Mod.9  4 3135.63       0.00   0.96   0.96 -1563.81
## AZ_Mod.6  4 3142.74       7.11   0.03   0.99 -1567.36
## AZ_Mod.5  4 3145.85      10.22   0.01   0.99 -1568.91
## AZ_Mod.1  3 3147.39      11.76   0.00   1.00 -1570.69
## AZ_Mod.12 3 3148.10      12.47   0.00   1.00 -1571.05
## AZ_Mod.3  5 3148.85      13.23   0.00   1.00 -1569.42
## AZ_Mod.8  3 3153.50      17.87   0.00   1.00 -1573.75
## AZ_Mod.4  3 3158.83      23.20   0.00   1.00 -1576.41
## AZ_Mod.11 2 3160.31      24.68   0.00   1.00 -1578.15
## AZ_Mod.13 2 3160.98      25.35   0.00   1.00 -1578.49
## AZ_Mod.10 3 3161.50      25.87   0.00   1.00 -1577.75
## AZ_Mod.7  3 3162.05      26.43   0.00   1.00 -1578.02
## AZ_Mod.2  4 3163.28      27.65   0.00   1.00 -1577.63
## AZ_Mod.14 0 3389.49     253.86   0.00   1.00 -1694.74
```

```
##       DirecDurat I(DirecDurat^2)            Rough
##         42.96150        43.10172          1.02149
```

```
##
## Call:
## glm(formula = PointType ~ DirecDurat + I(DirecDurat^2) + Rough,
##     family = binomial, data = GLM_training)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.640184   0.334487  -1.914  0.05563 .
## DirecDurat      -0.312154   0.122551  -2.547  0.01086 *
## I(DirecDurat^2)  0.021951   0.007578   2.897  0.00377 **
## Rough            0.054410   0.012459   4.367 1.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3157.6  on 2444  degrees of freedom
## Residual deviance: 3127.6  on 2441  degrees of freedom
## AIC: 3135.6
##
## Number of Fisher Scoring iterations: 4
```

I then tested the predictive accuracy of the optimal GLM. VIFs are low (except for quadratic version of variable).

```
## Confusion Matrix and Statistics
##
##                 Reference
## Prediction       BodyLocation RandomLocation
##   BodyLocation            242            138
##   RandomLocation           39             11
##
##                Accuracy : 0.5884
##                  95% CI : (0.5402, 0.6353)
##     No Information Rate : 0.6535
##     P-Value [Acc > NIR] : 0.9979
##
##                   Kappa : -0.077
##
##  Mcnemar's Test P-Value : 1.757e-13
##
##             Sensitivity : 0.86121
##             Specificity : 0.07383
##          Pos Pred Value : 0.63684
##          Neg Pred Value : 0.22000
##              Prevalence : 0.65349
##          Detection Rate : 0.56279
##    Detection Prevalence : 0.88372
##       Balanced Accuracy : 0.46752
##
##        'Positive' Class : BodyLocation
##
```

Continuing analyses, I also utilized the Extreme Gradient Boosting Machine Learning Application, XGBOOST, which is useful as it efficiently supports regularization which helps avoid overfitting, to attempt to predict mortality sites vs random locations using environmental metrics.

```
## ##### xgb.Booster
## raw: 4.4 Mb
## call:
##   xgb.train(params = xgb_params, data = Boost_train_matrix, nrounds = 500,
##     watchlist = watchlist, verbose = 0, eta = 0.005, max.depth = 7,
##     subsample = 0.85)
## params (as set within xgb.train):
##   objective = "multi:softprob", eval_metric = "mlogloss", num_class = "2", eta = "0.005", max_dep
th = "7", subsample = "0.85", validate_parameters = "TRUE"
## xgb.attributes:
##   niter
## callbacks:
##   cb.evaluation.log()
## # of features: 13
## niter: 500
## nfeatures : 13
## evaluation_log:
##     iter Boost_train_mlogloss Boost_test_mlogloss
##        1            0.6915780           0.6921951
##        2            0.6900384           0.6912977
## ---
##      499            0.4131173           0.6004236
##      500            0.4128766           0.6004295
```

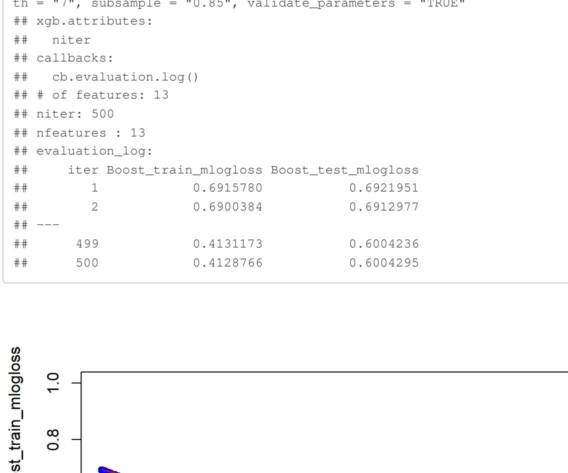



Figure 4. Training log-loss metric indicating training (blue) & testing (red) accuracy

```
## [1] "Minimum log-loss/training accuracy: "
```

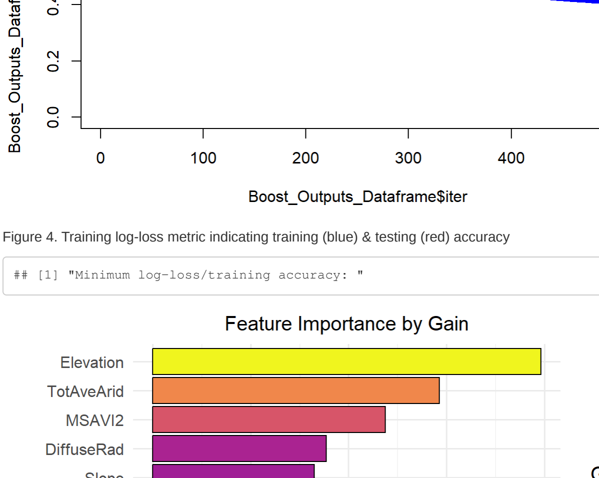



Figure 4. Feature importance as ranked by the gain metric.

```
## [1] "Confusion Matrix and Statistics for the XGBoost predictive approach"
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 186 101
##          1 189  98
##
##                Accuracy : 0.4948
##                  95% CI : (0.4531, 0.5365)
##     No Information Rate : 0.6533
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : -0.0105
##
##  Mcnemar's Test P-Value : 3.242e-07
##
##             Sensitivity : 0.4960
##             Specificity : 0.4925
##          Pos Pred Value : 0.6481
##          Neg Pred Value : 0.3415
##              Prevalence : 0.6533
##          Detection Rate : 0.3240
##    Detection Prevalence : 0.5000
##       Balanced Accuracy : 0.4942
##
##        'Positive' Class : 0
##
```

I then implemented a slightly more complicated technique - a neural network - from package 'nnet'.

```
## # weights:  151
## initial  value 763.895241
## iter  10 value 519.661462
## iter  20 value 511.022660
## iter  30 value 491.357866
## iter  40 value 485.041019
## iter  50 value 480.332839
## iter  60 value 476.581810
## iter  70 value 473.404369
## iter  80 value 469.616012
## iter  90 value 466.737888
## iter 100 value 464.567146
## iter 110 value 462.983377
## iter 120 value 461.643434
## iter 130 value 460.493253
## iter 140 value 459.568997
## iter 150 value 458.488288
## iter 160 value 457.312711
## iter 170 value 456.507555
## iter 180 value 455.905934
## iter 190 value 455.434966
## iter 200 value 455.112506
## final  value 455.112506
## stopped after 200 iterations
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  21  15
##          1 178 360
##
##                Accuracy : 0.6638
##                  95% CI : (0.6235, 0.7024)
##     No Information Rate : 0.6533
##     P-Value [Acc > NIR] : 0.316
##
##                   Kappa : 0.0811
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.10553
##             Specificity : 0.96000
##          Pos Pred Value : 0.58333
##          Neg Pred Value : 0.66914
##              Prevalence : 0.34669
##          Detection Rate : 0.03659
##    Detection Prevalence : 0.06272
##       Balanced Accuracy : 0.53276
##
##        'Positive' Class : 0
##
```

## Results

The relationships between the environmental variables as discerned by the correlation plot were as follows: all forms of radiation were positively correlated, slope and terrain roughness were positively correlated, aspect was negatively correlated with the relative contribution of incoming solar radiation at 9:00AM and positively correlated with the relative contribution of incoming solar radiation at 6:00pm. The relative contribution of incoming solar radiation received at 2:00PM and 6:00PM negatively correlates to that at 9:00AM. The best-fit GLM identified the relationships between the variables as follows. The results show that duration of direct solar radiation exposure has a significant negative linear effect ($p<0.01$) and a significant positive quadratic effect ($p<0.002$), indicating a non-linear relationship with the point type. Roughness is also a significant positive predictor ($p<0.001$), suggesting that rougher terrain is associated with higher likelihoods of the point being a mortality site. Overall, the model identifies the duration of direct solar radiation exposure and terrain roughness as impactful environmental factors influencing the response variable, with a notable quadratic pattern for duration of solar radiation received by that location.The predictive accuracy of the GLM was 59% overall. The training and testing XGBoost logloss values (<0.61) were better than random (which would be closer to 0.70). Elevation was the most important environmental feature as measured by gain, followed by the average aridity and MSAVI2 vegetation index. The predictive accuracy of the XGBOOST model was ~50%. The neural network ceased after 200 iterations and resulted in a 66% accuracy, the highest of the attempted methods.

## Conclusion

Here, I explored various methods used to explore how environmental variables predict UDI mortality locations, with the ultimate goal of identifying key predictors to improve understanding of risk areas. By integrating terrain generated from a high-resolution digital elevation model and subsequent derivatives — including solar radiation, vegetation density, and aridity - I evaluated patterns in the data and tested their predictive power using GLMs, AIC scores, XGBoost, and neural nets. Results showed that elevation, aridity, and vegetation were among the most influential factors, with predictive accuracies ranging from 50% to 66% depending on the method used. Ultimately, the neural net was the most accurate, although: with additional variable examinations, data preprocessing, and tuning, it is possible that these accuracy values could increase. These findings highlight the value of incorporating detailed environmental metrics into predictive models to better understand the spatial risks associated with UDI mortality, offering potential applications for more targeted mitigation efforts in the Tucson Sector.

## References


Martinez, Daniel, Robin Reineke, Bruce Anderson, Gregory Hess, and Bruce Parks. "Migrant Deaths in Southern Arizona: Recovered Undocumented Border Crosser Remains Investigated by the Pima County Office of the Medical Examiner, 1990-2020." J Hum Secur 1(2021): 257–286. US Customs and Border Protection. 2022a. "Nationwide Enforcement Encounters: Title 8 Enforcement Actions and Title 42 Expulsions." https://www.cbp.gov/newsroom/stats/cbp-enforcement-statistics/title-8-and-title-42-statistics-fy2020. US Customs and Border Protection. 2022b. "Southwest Land Border Encounters (By Component). Fiscal Year 2022." https://www.cbp.gov/newsroom/stats/southwest-land-border-encounters-by-component. Boyce, G. A., S. N. Chambers, and S. Launius. 2019. "Bodily Inertia and the Weaponization of the Sonoran Desert in US Boundary Enforcement: A GIS Modeling of Migration Routes through Arizona's Altar Valley." Journal on Migration and Human Security 7(1): 23–35.

Giordano, Alberto, and Katherine Spradley. 2017. "Migrant Deaths at the Arizona–Mexico Border: Spatial Trends of a Mass Disaster." Forensic Science International 280(1):200–212.

Foster, Kevin M., Brendon A. Bradley, Christopher R. McGann, and Liam M. Wotherspoon. 2019. "A VS30 Map for New Zealand Based on Geologic and Terrain Proxy Variables and Field Measurements." Earthquake Spectra 35(4):1865–1897.