Melanie Sattler, Khoi Tran, Haley Blair, and Bryce Runey

Modeling the impact of Weather on Distance Traveled by Lost Persons*

Abstract— Missing Persons cases are a race against time, where every minute is critical to save a life. The more information a Search and Rescue (SAR) team has to work with, the more likely the success of the search. dbS Productions created a Search and Rescue database with over 20,000 search and rescue cases across the world to assist rescuers in their SAR efforts. The database includes search-specific information such as location, eco-division, and limited weather information. It also includes personal data, including sex, age, clothing, and equipment, as well as various characterizations of the missing person, such as whether they are a hunter, a hiker, or have various medical conditions, such as dementia. All of these factors can be used to determine where a missing person may have headed while they were lost and try to locate them more efficiently.

The primary goal of this research is to create a predictive model by augmenting existing spatial models implemented by dbS Productions with additional weather features, determining how weather conditions impact the distance traveled by lost persons, thus improving the efficiency of search and rescue operations. This process was established through regression modeling and other machine learning methods. Several models included in order to determine the effect of weather on the distance traveled, including regression models, models using support vector machines (SVM), and the most successful model using XGBoost.

The results showed that there was a relationship between the distance traveled and the maximum temperature and the minimum temperature. Overall showing that the weather extremes have a significant impact on the distance traveled by lost persons.

I. Introduction

Search and Rescue missions are time-critical, and wisely spent resources can be the difference between life or death. Search and Rescue efforts are often completed by volunteer organizations, supported by local police and fire departments, but are still often limited in resources that can be expended per-mission.

For these teams, having specialized data or tools to direct their search could potentially minimize the frequency of terminated unsuccessful searches. Due to the limited resources and volunteer nature of these efforts, research in this field is limited, outdated, and based upon small case analyses.

Within the field of search and rescue, William Syrotuck was an early theorist, publishing multiple books on topics including "grid search techniques for locating lost individuals in wilderness areas" and "[lost] person behavior" [9]. In his first publication, Syrotuck established guidelines and suggestions for search directors on how to best establish a grid plot, allowing for quick and efficient location of lost persons by marking geographic locations through a grid pattern [9]. This research was promptly followed by the publication of a deeper analysis on human behavior, and its role in search and rescue operations [4]. This publication discusses strategy in reducing the field of search. Syrotuck's publications greatly contributed to the concept of a more efficient search, but with limited distribution and implementation of these resources, search and rescue groups are frequently conduct search and rescue efforts without established scientific methods and strategies.

Another expansive piece of historical research completed on lost persons is in an analysis completed by the United States Department of the Air Force in 1985 [11]. This research was ultimately compiled into something of a survival manual for Air Force personnel, discussing survival tactics and lost persons psychology, however, it does not consider the effect of weather upon the travel of the lost person. The manual does indicate that harsh or extreme weather conditions require additional conditioning for people to survive but does not elaborate extensively [11].

Other research in the field includes that of optimal route planning, specifically in small scale rescue operations such as buildings in hazardous conditions, where the optimal pathway to find a missing individual can be derived due to the localized scale of the search [12]. The localized scale of the search and rescue effort makes it practical to determine the distance traveled by the missing individual, but these search methods do not scale well to searches of increasing size and give few insights into scenarios with an unknown or unobstructed pathway. In addition to the sparse research on how far lost persons travel, even less research has gone into any potential interaction between weather and search and rescue activities, or weather and a lost person's distance traveled.

Perhaps the ultimate resource in the field of search and rescue is the International Search and Rescue Incident Database (ISRID) [2]. ISRID is a database compiling research conducted by the SARBayes project, a previous project analyzing the survivability of lost persons, based on mission data provided by the search and rescue teams [10]. ISRID became a place to collect and evaluate search and

* Research supported by dbS Productions.
Melanie Sattler MSDS candidate University of Virginia (email: ms9py@virginia.edu)
Khoi Tran MSDS candidate University of Virginia (email: kt2np@virginia.edu)
Haley Blair MSDS candidate University of Virginia (email: hab6qc@virginia.edu)
Bryce Runey MSDS candidate University of Virginia (email: bmr4ru@virginia.edu)

rescue data and has now grown to conglomerate data from 40 different sources, amounting to 165,000 search cases, providing data on many different aspects of a search and rescue case [2].

Robert Koester expanded on many of these ideas in his research, in his book titled "Lost Person Behavior" [5]. Here, many factors are agglomerated – including distance – as variables within a search and rescue attempt. In analyzing psychological factors of human behavior, notably in lost persons' navigation, he created a plan and algorithm for a more navigable search and rescue planning method [5]. Based on research in the book, Hunters tend to travel further out from the Initial Planning Point in comparison to other lost persons [5]. This could be due to multiple additional human factors, such as age and health, that could also play a role in how far a lost person travels [5]. This research did evaluate distance, but it did not have weather data available at the time of publication, so the impact of weather was not considered in this research.

The primary goal of this research is to determine how weather conditions impact the distance traveled by lost persons to increase the efficiency of search and rescue operations. This was established through regression modeling and other machine learning methods. The modeling in this paper used the data available in ISRID to evaluate the problem as well as augmented data from resources such as the National Oceanic and Atmospheric Association (NOAA) [7]. Search and rescue research is limited, so this paper attempts to expand on the research available on search and rescue to continue to improve rescue missions.

II. Methods

A. Data Sourcing

A major issue the team faced was the vast amount of missing data. Provided with nearly 23,000 entries of lost persons, we were unable to use the majority of these due to their incompleteness. As a team, we concluded that any observation missing values in more than fifty percent of the columns should be discarded. We also implemented criteria where some variables were more important than others, so if those were missing values, we would discard those observations as well. These columns and categories included weather information, such as temperature, a binary variable indicating rain or snow, etc. Our response variable of distance traveled from the original location of the missing person was also vital, therefore observations missing this value were mostly discarded. However, we did have access to some general location data based on zip code, and sometimes the specific latitude and longitude. Also, we were able to recover some of the observations missing official distance through calculations. Following the completion of data removal, we relied upon imputation methods to augment the dataset.

ISRID provided the majority of data from 16,683 searches. In addition to this, the data for this project also includes more recent search data from a variety of sources, including state-level searches from Nevada and Oregon, along with data from the Great Smoky Mountains. Additional data was augmented to try to supplement the existing weather data in ISRID by obtaining additional data from the NOAA API [7].

B. Data Processing

As the data came from a wide variety of sources, and search and rescue groups do not share any standards for data collection, the data was often very fragmented and incomplete, even within a dataset from a singular source. The data required extensive cleaning, ranging from correcting individual string values of typo-induced mistakes, which would lead to a location being a character off, or a coordinate being a digit off, to forcing all location data to the same standard of units. While much of the location data were in latitude and longitude coordinate format most are familiar with, much of it also resided in variations of the UTM coordinate format, which required extensive cleaning, especially with frequent typographical errors. Additionally, data points attempting to classify lost persons, known as "Subject Categories" were simplified, eliminating inconsistencies and typographical errors, and transformed into dummy categorical variables, to account for the fact that many subjects fell within more than one category. For example, a subject category originally labeled "child swimming" would be reclassified to the child category and the water category.

TABLE I. Representation of Missingness in Data

Variable	*Missing Values*	*% of Total Values*
cat_night	2401	28.56%
dist_mi	5302	63.07%
eco_div	391	4.65%
lat	2401	28.56%
lon	2401	28.56%
month	2401	28.56%
outcome	56	0.67%
rain	2401	28.56%
season	2401	28.56%
snow_depth	2401	28.56%
temp_avg	4201	49.97%
temp_max	4201	49.97%
temp_min	4201	49.97%
terrain	992	11.80%
year	2401	28.56%

Much of the work in data processing was to solve or alleviate missingness in data. Most variables in their original datasets were upward of fifty or even ninety percent incomplete, leaving little complete data to work with – and most notably, records of weather conditions, which are the primary focus of this project, were largely nonexistent. Through the use of NOAA and Meteostat's APIs, weather data for most search attempts were found, given the location, date, and time of the incident. However, most location and datetime data suffered from a high degree of missingness, or was fragmented and not usable for the purpose of extracting weather data. In order to alleviate this, approximated coordinates, derived from a zip code (when applicable) were used alongside heavily cleaned datetimes in accessing NOAA and Meteostat's APIs. In lieu of adequate location data or available zip codes, zip codes could often be derived using a provided town or county name alongside the geolocator functions within the GeoPy package.

Even with extensive cleaning, and the use of heavy data augmentation from various packages, the final set of variables for modeling still suffered from a high degree of missingness. In order to alleviate this, the scikit-learn package was utilized, allowing for the use of various imputation methods, which were tested alongside the testing of various modeling techniques. Ultimately, KNN imputation with a k-neighbors value of 1, which was deemed to best align with the distribution of the original data, was settled upon.

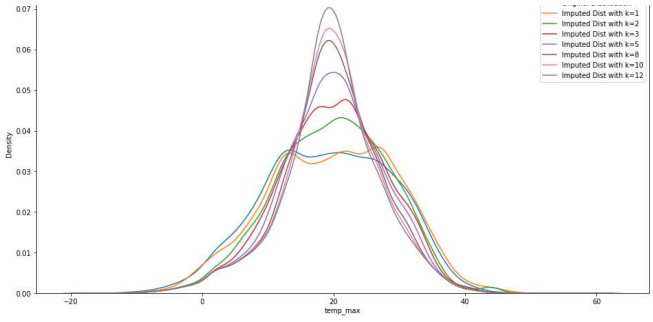

Figure 1. A graphical representation of a variable with scikit-learn's KNN imputation, conducted with varying k-neighbors values.

C. Feature Engineering

As part of the weather data augmentation, a generalized set of longitudinal and latitude variables were created, in order to accommodate for observations without provided location data. Alongside this, a precise nighttime binary categorical variable was also created, using the `astral` package to pinpoint sunrise and sunset times for a given location.

III. Modeling

A. Models

Several models were created to evaluate the impact of weather and other variables on the distance traveled by lost persons. Multiple linear regression showed that terrain-flat, total hours, and eco-division had a significant impact on distance traveled, while the weather variables were not significant, indicating low impact on distance traveled. An additional linear regression model was created using only weather components. The only weather variable with a significant impact was the minimum temperature. Forward and backward stepwise regression were also performed with the entire dataset and with the subset containing only weather variables to determine any indications of a significant relationship between weather and the distance traveled.

In order to use classification methods, the distance traveled outcome variable was binned into a categorical variable with five labels. As search and rescue procedures tend not to be very precise, estimating and predicting an approximate distance – binned as one, two, five, ten, or greater than ten-mile distances – were adequate in conforming to the goals of providing a guideline in how far to search for missing persons. SVM classification, boosted logistic regression, and random forest classifications were then performed on the data.

Various eXtreme Gradient Boost (XGBoost) models were also utilized, for the predictive power and flexibility of gradient boosting. Using the tune library from the ray package alongside XGBoost, an accelerated hyperparameter search was performed, testing randomized learning rates, max depths, estimator counts, and alpha values for gradient-boosted random forest regression, random forest classification, linear regression, and logistic regression, with the classification modeling techniques performed upon our binned distance variable, transformed into a categorical variable.

The hyperparameter search and XGBoost modeling techniques were performed on both the standard data – complete with missing values – and an imputed dataset, using KNN imputation.

B. Validation

Exploratory data analysis was initially conducted in order to determine potential shortfalls in the data; initial validation included verifying that all assumptions were met, in order to meet the statistical requirement of a linear regression model. A correlation matrix, using the seaborn python package, was then used to visualize any possibility of multiple collinearity. Finally, the modeling was conducted with multiple fold cross-validation to ensure robust and high performing models, using cross-validation packages in R for the linear regression models and scikit-learn in python.

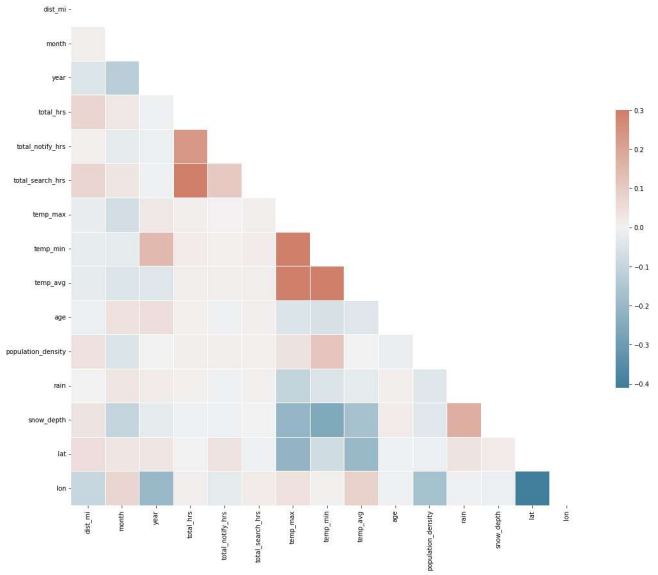

Figure 2. Correlation Matrix validation for multi-collinearity

C. Equations

Most of the models built during this research indicated that there was little to no connection between weather and the distance traveled by lost persons. The initial linear regression showed that weather was not a significant factor when looking at the complete dataset. Upon further examination of subsets of the data, some showed an improvement in the significance of weather variables. Similarly, one of the subsets of category types within the dataset did result in a significant model.

The most successful multiple linear regression model contained the maximum and minimum temperature, binary variables for mountainous, hilly, and water terrain, and a binary variable for missing runaways. The R^2 value of this model was 0.0215, while the F-score was 7.34 and a p-value of 8.835×10^{-8}. While the R^2 value appears small, with a large sample size the F-value is significant which supports that these factors influence the distance traveled by lost persons. The results establish that there could be a relationship that can be investigated with further research.

Additional models were constructed, using support vector machines, in order to attempt different approaches to modeling the data. The SVM models were unsuccessful, failing to establish a reasonable accuracy, likely due to the high degree of missingness and dimensionality of categorical data.

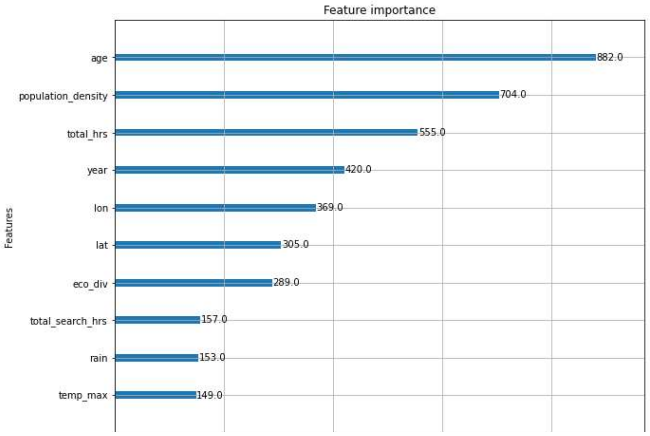

Figure 3. Random Forest Regression on KNN imputed data, showing age, time missing (total hrs) and aspects of location (population density, latitude/, longitude, snow, temp min/max/avg, eco -division as important features

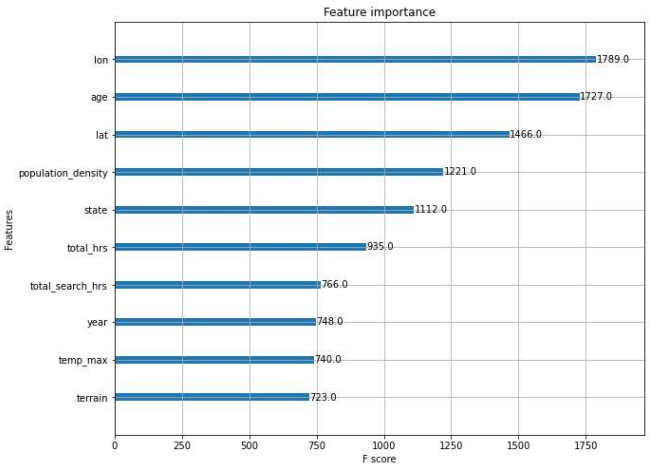

Figure 4. Logistic regression model output on non-imputed (standard) data, showing latitude, longitude, age, population density, state, and time missing (total hours) as important features

The XGBoost models were the models with the strongest predictive power. With the standard continuous outcome variable, a boosted random forest regression managed a 0.40616 R^2 value, far more successful than standard multiple linear regression. Notably, when handling the continuous outcome variable, the boosted method performed poorly with non-imputed data given the generally better handling of missing data by XGBoost. However, the regression model performed well on imputed data.

When performed on the data with a binned categorical outcome variable, XGBoost models showed inverse results compared to that with the continuous outcome variable, showing its strong handling of missing data, but also showing relatively poor results with imputed data. Gradient boosted classification displayed upwards of 68% accuracy.

		Evaluation		Hyperparameters			
Data	**Model**	***Score***	***Scoring Metric***	***Learning Rate***	***Max Depth***	***Estimators***	***Alpha***
Standard	Random Forest Regression	-0.02629	R^2	0.73928	1	132	0.10562
KNN (*k=1*) Imputed	Random Forest Regression	0.40616	R^2	0.81856	5	179	0.07798
Standard	Boosted Regression	-0.05166	R^2	0.06196	3	148	0.00217
KNN (*k=1*) Imputed	Boosted Regression	0.39714	R^2	0.09067	3	130	0.00105
Standard	Random Forest Classification	0.68371	Accuracy	0.73829	6	169	0.73964
KNN (*k=1*) Imputed	Random Forest Classification	0.39596	Accuracy	0.21703	9	102	0.68936
Standard	Boosted Classification	0.68922	Accuracy	0.02604	4	196	0.93840
KNN (*k=1*) Imputed	Boosted Classification	0.43703	Accuracy	0.58287	8	310	0.58286

Figure 5. Displaying the results of gradient-boosted modeling via XGBoost

IV. Discussion

A. Discussion of Results

The initial assumption was that weather would have a significant impact on the distance traveled, the initial model's results indicate that it is not the case. This was based on logical reasoning based on human behavior in outdoor activities in the initial literature review. Looking at the full dataset with all the variables did not produce significant results. The factor that were repeatedly was significant in the initial builds was time as well as the eco-division that lost persons was in.

While looking for other connections within the dataset, subset datasets, based upon specific subject categories were created in order to determine weather conditions' potential localized impact on specific groups. The results indicated a relationship between minimum and maximum temperatures on the distance traveled by lost persons. Overall low overall model accuracy contrasting with significant values for weather variables in multiple linear regression, makes the impact of weather conditions upon distance traveled tenuous.

However, weather conditions may have predictive value in context of various classifications of individuals, or subject categories. Here, further research would aid in expanding upon these initial results in determining what types of people are impacted more by extreme weather conditions and that could impact their distance traveled.

With the degree of missingness in the data, it was difficult to evaluate the validity of the models we were able to create. A significant amount of lost person cases had to be dropped from the data in order to create a significant model, while some variables were kept, even with high degrees of missingness, due to how crucial they were in modeling different aspects of location, time, weather, and distance, and were thus imputed.

Overall, the results indicate a possibility of weather conditions' impact on the distance traveled, conditional upon the type or "subject category" of missing person. Weather also seems to impact lost persons differently based on the eco-division they are in, and what terrain they are traversing, but varying weather conditions do not seem to have a universal impact. Further research would be needed to determine the true impact.

B. Use Cases

Currently, search and rescue is still largely completed by people in the field, making intuitive assumptions and predictions, based upon unstructured data and knowledge of the situation and the missing person at hand. Search and rescue groups are often undermanned or short staffed, often making search and rescue efforts something of an optimization problem, where every resource is applied with a great amount of efficiency.

The applicability of this model is clear as a tool to aid search and rescue crews. Having the input of who the person is, and what subject category they fall into as well as other factors that define a search and rescue case can help narrow the field of search by a significant margin [8]. This helps increase survivability of a case as time is a great factor in survivability of a lost person which was shown in research by Robert Koester as well as the previous capstone's research project [5][6]. Being able to conquer these factors would also allow for wise spending of the resources during a search and rescue case and allow limited resources to go further in each case.

Another future application of this research is as search and rescue groups start to transition into a modern era of search and rescue, unmanned aerial vehicles (UAV) are becoming more accessible and usable in search and rescue cases [3]. As they get used more often, they will need help guiding and fine tuning these UAVs to search in the correct area, and the correct radius around the search initiation [5][3]. They can do this by having access to these spatial assessments made through this research.

Ultimately, while this is only a small start, this research has applicability in the search and rescue field. As ISRID continues to grow and collect more information, it can hopefully establish more complete data standards, and further expand research and new applications to be used in the field.

V. Conclusion and Further Research

Based upon the results of this research, it is clear that extreme temperatures impact the distance traveled by lost persons, with minimum and maximum temperature values showing significance in all models. In order to improve the predictive value of the impact of weather, more complete data needs to be collected during search and rescue cases. There is difficulty collecting additional information throughout searches, as volunteer rescuers tend to not document or record their searches thoroughly. Documentation is frequently ceased upon the completion of a search and rescue mission, and crucial variables as simple as final times and locations may not be recorded, along with

more detailed information, such as what the lost person was wearing, or carrying with them. Additionally, standardization in data formats and practices used by search and rescue organizations will vastly improve uniform access to data.

In the future, there are many opportunities for further research in this area. More precise location-based data in cases of lost persons would greatly assist in determining the impact of weather on distance, as weather in itself is a localized event. Investigating further upon different subject categories, such as lost persons with varying health conditions, may provide further insight to the impact of weather on varying types of individuals. Acknowledgment

Robert J. Koester, Ph.D., of dbS Productions funded the research project and contributed to the creation of the International Search and Rescue Incident Database (ISRID).

Gerard P. Learmonth Sr., Ph.D., Professor in the School of Data Science at the University of Virginia, mentored and guided us throughout the project.

REFERENCES

[1] Choutri, K., Mohand, L., & Dala, L. (2020). Design of search and rescue system using autonomous Multi-UAVs. Intelligent Decision Technologies, 14(4), 553–564. https://doi.org/10.3233/IDT-190138W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] dbS Productions LLC. Incident Search & Rescue Incident Database. https://www.dbs-sar.com/SAR Research/ISRID.htm, 2011.

[3] Goodrich, M. A., Morse, B. S., Gerhardt, D., Cooper, J. L., Quigley, M., Adams, J. A., & Humphrey, C. (2008). Supporting wilderness search and rescue using a camera-equipped mini UAV. Journal of Field Robotics, 25(1/2), 89–110. https://doi.org/10.1002/rob.20226

[4] Hill, K. Anthony., Syrotuck, William., & National Search and Rescue Program (Canada). National Search and Rescue Secretariat. (1999). Lost person behaviour.

[5] Koester, R. J. (2014). Lost Person Behavior. Charlottesville, VA: dbS Productions.

[6] M. Pajewski, C. Kulkarni, N. Daga, and R. Rojhwani, "Predicting Survivability in Lost Person Cases," Proceedings of the 2021 Systems and Information Engineering Design Symposium, Charlottesville, Virginia, April 2021.

[7] NOAA Online Weather Data (nowdata): Interactive Data Query System : Public Fact Sheet. Washington, D.C.: National Oceanic and Atmospheric Administration, 2006. Internet resource

[8] Shabani, A., Asgarian, B., Gharebaghi, S. A., Salido, M. A., & Giret, A. (2019). A New Optimization Algorithm Based on Search and Rescue Operations. Mathematical Problems in Engineering, 1–23. https://doi.org/10.1155/2019/2482543

[9] Syrotuck, W. G. (1974). Some grid search techniques for locating lost individuals in wilderness areas.

[10] Twardy, C. R., Koester, R., & Gatt, S. (2006). Missing Person Behaviour An Australian Study.

[11] United States Department of the Air Force (1985). Search and Rescue: Survival Training. Washington, DC: Dept. of the Air Force.

[12] Zverovich, V., Mahdjoubi, L., Boguslawski, P., & Fadli, F. (2017). Analytic Prioritization of Indoor Routes for Search and Rescue Operations in Hazardous Environments. Computer-Aided Civil & Infrastructure Engineering, 32(9), 727–747. https://doi.org/10.1111/mice.12260H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.