RESEARCH/PD ANNUAL REPORT - PROGRESS REPORT

2015 annual report - progress
Julia Parrish
Scaling Up Cost-Efficient Community Engagement in Coastal Resource Management
R/OLWD-1
Submitted On: 05/03/2016 02:21:12 PM

METRICS & MEASURES

Metric/Measure	Value	Note
Acres of coastal habitat	0	
Fishermen and seafood industry personnel	0	
Communities - economic and environmental development	0	
Stakeholders - sustainable approaches	0	
Informal education programs	0	
Stakeholders who receive information	60	presentations
Volunteer hours	80	Time for COASST participants to travel to, and attend, one of 4 focus group meetings
P-12 students reached	0	
P-12 educators	0	

REQUESTED INFORMATION

Publications

No **Publications** information reported

Students Supported

Jennifer Lang (Continuing Student)
jwlang10@gmail.com
University of Washington, SAFS

Field of Study:
Advisor: Julia K. Parrish
Degree Type: MS
Degree Year: 2016

Student Project Title: A Multi-variate Analysis of COASST Data

Involvement With Sea Grant This Period (capstone, fellow, intern, etc.): grad student

Post-Graduation Plans (employer, grad school, etc.):

Was this thesis/dissertation supported by Sea Grant?: Yes

Thesis / Dissertation:

New or Continuing?: continuing

Degree awarded this reporting period?: No

Financially supported?: Yes

Narratives

Parrish 2015 Report Narrative
Uploaded File: WCSG2015_report_narrative.docx

Partners This Period

University of California, Davis (UCD)
Types: Academic Institution
Scale: REGIONAL
Notes:

Oregon State University (OSU)
Types: Academic Institution
Scale: REGIONAL
Notes:

STANDARD QUESTIONS

Impacts and Accomplishments

(1)

Type	accomplishment
Title	West Coast Sea Grant research explores why volunteers are motivated to join and remain active as citizen scientists
Relevance	Many research projects depend heavily upon the skills of volunteers to provide rigorous, cost-effective data collection for analysis, monitoring and management. But factors that influence recruitment and retention of well-trained volunteers can be critical for developing successful citizen science programs—and seldom are clearly identified and articulated.
Response	Using the Coastal Observation and Seabird Survey Team (COASST) as a case study, the West Coast Sea Grant research team employed quantitative and qualitative methods to explore factors associated with volunteer retention. The findings could help citizen science project planners tailor their programs to successfully retain volunteers. In addition to constructing a multivariate model based on volunteer information, their data collection sites and their communities, the team jointly conducted nine focus

	groups to identify emergent themes linked to joining and staying involved.
Results	Results from quantitative analysis of retention indicated that the number of people on a survey team, how far they travel, and their age were all influential factors for retention. Older individuals surveying in pairs that don't travel far to their study beach stayed on as volunteers longest. Weather appeared to have no effect. Focus group participants valued a well-organized program that collected meaningful data and reported regularly to volunteers on the larger data patterns and uses. The team also made progress on in-practice products that can be used to develop or expand rigorous coastal citizen science programs.
Recap	Washington Sea Grant researchers uncovered key factors influencing volunteer recruitment and retention in coastal citizen science programs.
Comments	
Primary Focus Area	Ocean Literacy and Workforce Development
Secondary Focus Areas	Healthy Coastal Ecosystems
Goals	The public is ocean literate.
Partners	Oregon State University (OSU) University of California, Davis (UCD)
PI Draft	* Type accomplishment * Title West Coast Sea Grant research identifies factors that influence community engagement and retention of volunteers in citizen science projects * Relevance Involvement in citizen science can deepen an individual's knowledge, awareness, and sense of place, as well as inspire behavioral change. Citizen science can also provide rigorous, cost-effective data collection for research, monitoring, and management needs. While recognition of its potential contribution to the scientific enterprise has garnered much attention for citizen science in recent years, factors that influence recruitment and retention of well-trained volunteers have not been clearly identified. * Response The West Coast Sea Grant programs jointly funded research to quantitatively and qualitatively examine factors affecting three measurable aspects of successful citizen science:engagement, retention, and accuracy. Using COASST (Coastal Observation and Seabird Survey Team) as a "model" coastal citizen science project, the research team is conducting a detailed analysis to abstract fundamental features that can be translated and used by other existing or emerging coastal citizen science projects. Both a qualitative component (focus groups) and a quantitative component (multivariate analysis of existing data) were planned

	for the study. * Results Focus groups were conducted in Bellingham (WA), Port Angeles (WA), Eureka (CA), and Brookings (OR) and all focus groups were coded. In addition, data fields, coding categories, and missing data were compiled for the multivariate analysis, and a quantitative analysis was completed. * Recap West Coast regional research is identifying factors contributing to recruitment, retention, and scientific success of volunteers engaged in citizen science. Comments Primary Focus Area Ocean Literacy and Workforce Development Secondary Focus Areas Healthy Coastal Ecosystems Goals The public is ocean literate. Partners Oregon State University (OSU) University of California, Davis (UCD)

Tools, Technologies, Information Services / Sea Grant Products

No **Tools, Technologies, Information Services / Sea Grant Products** information reported

Economic Impacts

No **Economic Impacts** information reported

Community Hazard Resilience

No **Community Hazard Resilience** information reported

Meetings, Workshops, Presentations

(1)

Type of Event	Public or professional presentation
Description	UC Davis School of Education Seminar - Mastery, Mental Models Appropriation and Identity: The COASST Program
Event Date	11-30-2015
Number of Attendees	25

(2)

Type of Event	Public or professional presentation
Description	NSF Seminar - Mastery, Mental Models, Appropriation and Identity: The COASST Program
Event Date	01-11-2016
Number of Attendees	35

Leveraged Funds

No **Leveraged Funds** information reported

Parrish 2015-16 Narrative
Scaling Up Cost-Efficient Community Engagement in Coastal Resource Management

This project is a cooperative effort including three PIs:
Parrish – UW
Rowe – OSU
Ballard – UC Davis

The objectives are to:

1. Determine how a successful citizen science program, specifically measured through engagement, retention, and accuracy, is influenced by four categories of variables present in any/all citizen science endeavors: individual-level participant demographics, organization-level factors, community-level contextual factors, and regional-level environmental factors.
2. Produce professional and in-practice products that can be used to develop or expand rigorous coastal citizen science programs, ultimately creating a network of rigorous citizen science programs linking coastal communities directly to institutions of higher education, research, and monitoring, and collecting vital information of the physical, biological, and human dimensions of the California Current Large Marine Ecosystem (CCLME).

During the reporting period:

1. The COASST office (Erika Frost, Volunteer Coordinator) scheduled all Focus Groups (Table 1), and followed through with COASST participants to make sure that there was adequate attendance.

Table 1. List of Focus Groups and Linked Trainings

Date	Location	# of COASSTers	Facilitator	COASST Staff
2/21/15	Brookings, CA	7	Heidi	Erika
2/22/15	Eureka, CA	10	Heidi	Erika
3/7/15	Port Angeles, WA	11	Heidi	Jane
3/14/15	Bellingham, WA	7	Heidi	Jane

2. The COASST office (Erika Frost, Volunteer Coordinator) scheduled all linked COASST training sessions, and followed through with COASST participants to make sure that there was adequate attendance.
3. COASST personnel (Frost, Dolliver) conducted all linked training sessions and attended Focus Groups, opening and closing the event, and otherwise assisting the facilitator (Ballard) as needed, and according to the Focus Group protocol.
4. COASST interns, supervised by COASST personnel, transcribed all Focus Group materials into excel spreadsheets for review and coding by the PIs.
5. The PIs, Jennifer Metes (UC Davis graduate student), and COASST staff met to validate the Focus Group transcriptions, and initiate the creation of the coding system. The UC Davis contingent (Ballard, Metes) finalized the codebook and performed the initial coding.

6. Tim Jones, UW postdoctoral associate, performed the multivariate analyses on the quantitative dataset amalgamated by Jennifer Lang (née Ma) in the previous year (see below).

Tri-State Sea Grant Recruitment/Retention Multi-variate Analysis
Part 1. Results from the Exploration of "Why Stay?"
Tim Jones, Postdoc, COASST

The following text gives an overview of the analyses conducted to identify which environmental, beach/survey, socio-economic and personal factors influence the retention of COASST volunteers.

We examined two volunteer-specific measures of volunteer retention. The first response variable used as a measure of retention, ***duration***, was calculated for each volunteer as the number of days between the first survey completed and the last survey completed. A second binary response variable, ***1-year***, was calculated for each volunteer and took a value of one if that volunteer had completed surveys for a year or more, and zero if they withdrew less than a year after their first survey.

We wished to examine volunteer retention as a function of a suite of predictors from each of the environmental, beach, socio-economic and volunteer specific datasets. Socio-economic information, obtained from census data, was assigned to each individual based on the volunteer's home zip-code, and therefore represents the community in which they reside. Beach information was assigned to each volunteer based on the beach that they primarily survey. Weather information was assigned to each volunteer by identifying the weather station that was closest to the volunteer's primary beach, and therefore represents the average climate and weather conditions at the primary data collection site. The response variables ***duration*** and ***1-year*** were combined with the corresponding volunteer-specific environmental, beach, socio-economic and personal information. The set of predictors described in **Table E1** were included as potential candidates for influencing volunteer retention. We used machine learning methods to identify which factors are the best predictors of retention. We used the boosted regression tree (BRT) analysis method, which is based on the construction of decision trees to map a response onto a set of predictors.

Table E1. Names, description and types of the predictors used to model COASST volunteer retention.

Predictor	Description	Type
Personal		
Gender	Volunteer gender	factor
Trained.Age	Volunteer's age when trained	numeric
Involvement.1	Motivation for volunteer involvement	factor
Avg.Travel	Average travel time to get to and from survey location	numeric
Ave.Groupsize	Average size of data collection group	numeric
Ave.SurveyTime	Average survey duration	numeric
Primary Beach predictors		

Substrate.beach	Beach substrate type	factor
Length.beach	Beach length	numeric
Access.beach	Beach access type	factor
Width.beach	Beach width	factor
mean.primary.ER	Average encounter rate (birds / km)	numeric
mean.primary.PREV	Average prevalence (proportion of surveys with any birds)	numeric
primary.MAX	Maximum encounter rate (birds/km)	numeric
Weather		
MNTM.avg	Month-averaged mean temperature	numeric
DT90.sum	Number of days where maximum temperature > 90F	numeric
DT32.sum	Number of days where minimum temperature < 32F	numeric
DP01.sum	Number of days where rainfall > 0.1 inch	numeric
DP10.sum	Number of days where rainfall > 1 inch	numeric
TPCP.avg	Average total precipitation	numeric
Socio-economic predictors in home zip code		
edu_LessThanHighSchool	Proportion of population educated to less than High School standard	numeric
edu.University	Proportion of population educated to undergraduate level or above	numeric
laborforce_Unemployed	Proportion of population who are unemployed	numeric
Notlaborforce	Proportion of population not in labor force (retired, carers)	numeric
income_lt20G	Proportion of population that have an annual income < $20,000	numeric
income_gt100G	Proportion of population that have an annual income > $100,000	numeric
incomebelowpoverty_prop	Proportion of population whose income is below the poverty line	numeric
age_prop60plus	Proportion of population that are 60 or older	numeric

For each of the response variables, ***duration*** (4^{th}-root transformed to approximate a Gaussian error distribution) and ***1-year*** (binomial response), the dataset was partitioned into training (80% of data-points) and test (20% of data-points) datasets with data-points selected at random. A BRT model with 10,000 trees was constructed based on the training dataset, and the models predictive capacity was optimised by obtaining predictions from the BRT model for the test dataset using only the first X_t trees added in the construction of the model. This allowed us to identify the number of trees, *nt*, at which the BRT model begins overfitting relative to the test dataset, and identifies the point at which the model has the greatest predictive capacity. This procedure was repeated on 30 different partitions of training and test datasets and results were averaged across the resultant 30 models. The BRT models fitted to the ***duration*** response variable achieved a maximum deviance explained, D^2, of 0.203, whereas the BRT models fitted to the ***1-year*** response variable achieved a maximum deviance explained of 0.139. Personal and survey-related predictors including average group size, average travel time, and the age at which the individual was trained came out as the most influential based on relative influence scores (a measure of how often a variable is chosen to be split) (**Table E2**). Beach related predictors including beach length, access type, mean encounter rate and prevalence were also relatively influential, but less so than survey related factors (**Table E2**). Predictors related to average climate were relatively less influential, with average temperature (MNTM.avg) and the number of days receiving a minimum of 0.1 inches of rain (DP01.sum) being the most influential of the six weather predictors trialled (**Table E2**). Socio-economic factors were also relatively uninfluential, with the exception of the proportion of individuals aged 60 or more in the home zip code (**Table E2**).

Partial dependency plots for the eight most important factors (determined by relative importance) were plotted illustrating the response of retention across that predictors range, integrating across the response of all other predictors. The response profiles produced by the BRT models for each predictor were similar between the response variables ***duration*** and ***1-year***. The response profile for average group-size shows that individuals that always perform surveys alone (group-size = 1) have a lower retention than individuals who have at any point performed surveys with others (group-size > 1) (**Figure E1**). However, individuals performing surveys with group-sizes ~ 2 have on average the highest retention, with a sharp decrease in retention for individuals who participate in surveys in group-sizes > 2 (**Figure E1**). This may indicate that individuals in larger groups feel less engaged in the process of data collection, and are less likely to stay involved with COASST, as in most surveys a maximum of two individuals (one data recorder and one measurer) can be active at any one time. The response for average travel time is complex, but shows that individuals who travel for less than 30 minutes have a higher retention than individuals who travel for 30-60 minutes (**Figure E1**). Beyond a travel time of 60 minutes there is an increase in retention up to a peak at ~ 160 minutes, and then decreases from 160-230 minutes (**Figure E1**). This second peak at ~160 minutes is perhaps suggestive of a group of volunteers that are very dedicated and are likely to travel great distances, and are likely to remain as volunteers for longer (**Figure E1**). The response of age when trained shows that younger individuals (age < 30) have a lower average retention than individuals aged 30 +, with a general increase in retention as age increases (**Figure E1**). This may reflect the fact that older individuals are much more likely to remain in one area, and will likely have a more stable job and/or home-life than younger individuals. Similarly, the proportion of the population aged 60+ has a positive effect on retention, perhaps due to a larger community of active retired volunteers leading to a greater feeling of engagement within the community leading to greater retention (**Figure E1**).

Table E2. Summary of the relative influence of predictors for BRT models fitted to the response variable ***duration***. Relative influence scores are presented as the mean and the range (minimum and maximum) across the 30 data partitions the models were evaluated on.

Variable		Relative Influence	
Name	Type	Mean	Range
Ave.Groupsize	Personal/Survey	16.9	13.2 - 22
Avg.Travel	Personal/Survey	9.9	8 - 11.7
Trained.Age	Personal	8.9	6.2 - 12.6

age_prop60plus	Socio-economic	6.8	4.1 - 10.7
Length.beach	Beach	6.2	3.3 - 8.2
Access.beach	Beach	4.8	2.8 - 6.4
Ave.SurveyTime	Personal/Survey	4.5	3.1 - 6.5
mean.primary.ER	Beach	4.0	2.4 - 5.5
Involvement.1	Personal	3.1	1.5 - 4.4
mean.primary.PREV	Beach	3.1	1.4 - 4.3
MNTM.avg	Weather	3.1	1.8 - 4.9
income_lt20G	Socio-economic	2.8	1.7 - 5
primary.MAX	Beach	2.6	1.5 - 3.9
DP01.sum	Weather	2.4	1.4 - 4
income_gt100G	Socio-economic	2.3	1.2 - 3.3
edu_LessThanHighSchool	Socio-economic	2.2	0.9 - 3.1
TPCP.avg	Weather	2.0	1 - 3.5
laborforce_Unemployed	Socio-economic	2.0	0.9 - 3
DP10.sum	Weather	1.9	0.8 - 2.7
Notlaborforce	Socio-economic	1.8	0.8 - 2.9
edu.University	Socio-economic	1.7	0.9 - 2.4
DT32.sum	Weather	1.7	0.9 - 2.8
DT90.sum	Weather	1.6	0.8 - 2.6
Width.beach	Beach	1.3	0.4 - 2.2
Gender	Personal	1.1	0.3 - 4.1
incomebelowpoverty_prop	Socio-economic	0.9	0.2 - 1.6
Substrate.beach	Beach	0.6	0.2 - 1.6

The response of beach length with respect to retention show that individuals surveying shorter beaches have lower retention than those surveying longer beaches (**Figure E1**). The effect of beach access illustrates that individuals that have to drive (responses of drive and walk/drive), to get to their beach have a lower retention on average than individuals using any other mode of transport (**Figure E1**). The response for average survey time shows that individuals carrying out shorter surveys (< 60 minutes) have a lower retention than individuals carrying out surveys averaging 60 – 110 minutes, with a slight reduction in retention for individuals carrying out surveys averaging > 110 minutes (**Figure E1**). Finally, individuals surveying beaches with lower bird encounter rates (0 – 2 birds/km) have higher retention than individuals surveying beaches with encounter rates > 2 birds/km (**Figure E1**).

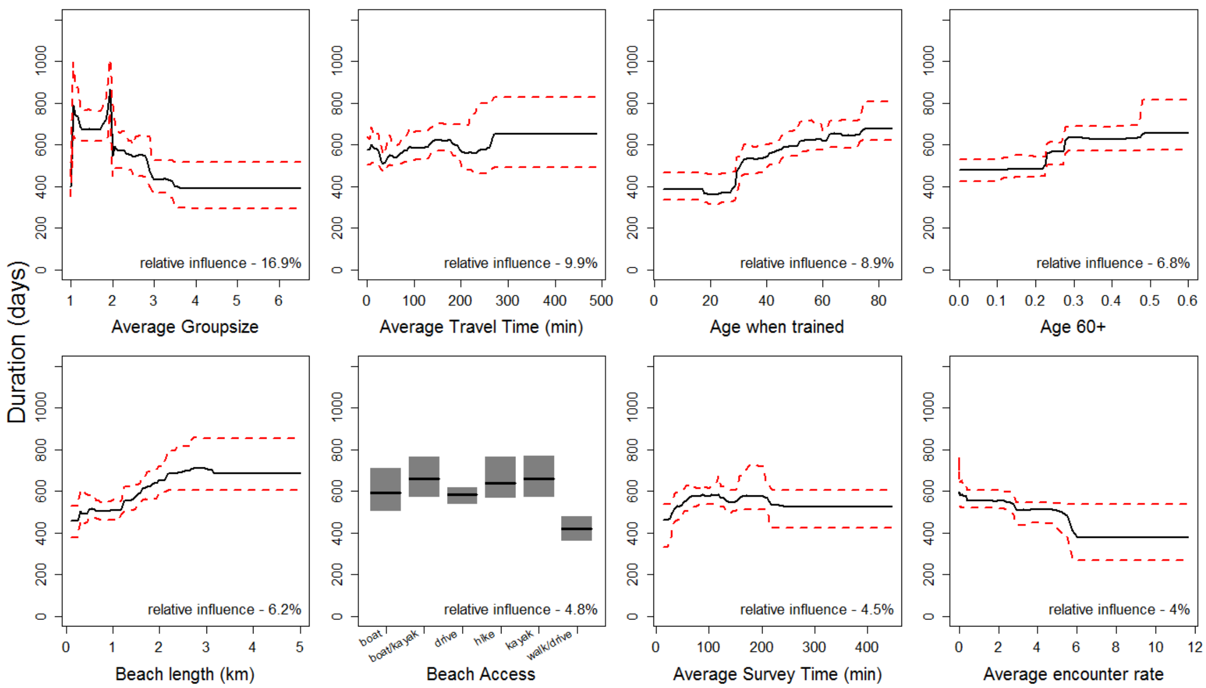

Figure E1. Response profiles for the eight most influential predictors for the BRT model fitted to the response variable, ***duration***. Response profiles represent the average response for each variable integrated across the response of all other predictors. The black line illustrates the average response profile across the 30 BRT models fitted to each of the 30 data partitions, and red dotted lines, and grey boxes for factorial variables, show the maximum and minimum predicted responses.

Description of data and data processing

Several sources of information were compiled to represent the environmental, socio-economic and personal factors that may influence the duration of time that COASST volunteers remain active. These were:

- Weather station data: provides information on long-term average climate, mostly with respect to temperature and precipitation (data on average wind strength was not available). Data were available for 49 weather stations from California, Oregon and Washington State.
- Census data: provides information on income, housing, age-demographics, ethnicity, education and employment for individuals residing in California, Oregon and Washington State. Data were broken down by zip code, of which we were able to obtain information for 285 unique zip codes.
- Beach data: provides information on length, width, substrate, latitude, longitude, access type (e.g. drive, hike etc), long-term average bird encounter rate (average number of birds encountered per km for that beach, averaged across all surveys carried out from 1999-2015), average prevalence (proportion of all surveys where any birds were encountered for that beach) and maximum encounter rate for each beach surveyed by COASST. Information was available for 545 beaches from California, Oregon and Washington State and was representative of all surveys completed between 1999 and 2014 inclusive.
- Volunteer data: provides information on home city, state, zip-code, age, trained date, first/last survey dates, occupation, birding experience, number of surveys performed, number of birds found, number of unique species/families found, primary beach surveyed, average travel time, average survey time and average group size for each volunteer. Data were available for 2411 individuals from California, Oregon and Washington State and the information (e.g. number of surveys, birds found etc) was representative of all surveys completed between 1999 and 2014 inclusive.

We wished to examine volunteer retention as a function of a suite of predictors from each of these four datasets. The response variable considered was specific to each individual volunteer within the volunteer dataset. Predictors from the weather, beach and census datasets were assigned to each individual in the following way. Census information was assigned to each individual volunteer based on the volunteer's home zip-code, and therefore represents the community in which they reside. Beach information was assigned to each volunteer based on the beach that they primarily survey. Although many volunteers survey, or have surveyed, multiple beaches (42% of individuals who performed surveys had performed surveys on more than one beach), the majority have a primary beach that they survey (91% of individuals who performed surveys had performed 50% or more of their surveys on a single, or primary, beach), and so the primary beach factors represent the factors that each volunteer encounters when carrying out surveys. Weather information was assigned to each volunteer by identifying the weather station that was closest to the volunteer's primary beach, and therefore represents the average climate and weather conditions at the primary data collection site. This resulted in a single dataset with each datapoint specific to each volunteer recorded in the COASST database and a series of environmental, socio-economic and personal predictors.

Response variables for retention

The first response variable used as a measure of retention, ***duration***, was calculated for each volunteer as the number of days between the first survey completed and the last survey completed. Of the 2411 individuals, 479 had neither a first or last survey date and were individuals who attended a training session but didn't participate in COASST surveys and so were excluded from this analysis. Furthermore, individuals who completed a survey within the last six months of 2014 (i.e. on or later than 06-01-15) were excluded (526 individuals) as this includes individuals who are, or may still be, actively collecting data. Of the remaining individuals, 350 had a ***duration*** of zero, and were individuals who had participated in a single survey. These individuals were excluded as they were predominantly the 3rd to 6th data collector and were likely friends/family who accompanied more experienced COASST data collectors on individual surveys. Furthermore, the home zip-code was missing for the majority (60%) of these individuals and so we were unable to assign census predictors to these individuals. After removing these individuals there were 1056 data points, representing 1056 unique volunteers.

A second binary response variable for retention, ***1-year***, was calculated for each volunteer and took a value of one if that volunteer had completed surveys for a year or more, and zero if they withdrew less than a year after their first survey. Individuals who attended a training session but didn't participate in COASST surveys were excluded. Individuals whose first survey was on or after 01-01-2014 were excluded (93 individuals) as this includes individuals who are active, but can't have completed one year of surveys due to when they started collecting data for COASST. Of the remaining individuals, 333 had a ***duration*** of zero, and were individuals who had participated in a single survey. These individuals were excluded as explained above. After removing these individuals there were 1506 data points, representing 1506 unique volunteers, of which 68% had completed surveys for a year or more.

Most COASST volunteers operate in pairs, as this aids in the process of measurement and data recording. Although many volunteers have had several data collection partners, there are many individuals who joined (and subsequently left) COASST as a couple (i.e. married couples, siblings, close friends), and as a result they have the same start date, end date, and set of predictors having completed all surveys together. As a result these individuals can be considered as pseudo-replicates, as the data points (in terms of response and predictors, with the exception of) are almost identical. This may have the effect of artificially inflating the proportion of variance explained and may mask important relationships. For this reason, these pairings were identified, by identifying which pairs had the same start and end dates, as well as survey locations and home-zip code. These groups of two or more people (sometimes up to four individuals shared the same response and predictor set) were labelled, and only one individual from each group (selected at random) was retained in the dataset. This resulted in datasets for ***duration*** that included 929 individuals and for ***1-year*** that included 1306 individuals.

Predictors

The response variables ***duration*** and ***1-year*** were combined with the volunteer-specific environmental, beach, socio-economic and personal information. The set of predictors described in **Table 1** were included as potential candidates for influencing volunteer retention.

Table 1. Names, description and types of the predictors used to model COASST volunteer retention.

Predictor	Description	Type
Personal		
Gender	Volunteer gender	factor
Trained.Age	Volunteer's age when trained	numeric
Involvement.1	Motivation for volunteer involvement	factor
Avg.Travel	Average travel time to get to and from survey location	numeric
Ave.Groupsize	Average size of data collection group	numeric
Ave.SurveyTime	Average survey duration	numeric
Primary Beach predictors		
Substrate.beach	Beach substrate type	factor
Length.beach	Beach length	numeric
Access.beach	Beach access type	factor
Width.beach	Beach width	factor
mean.primary.ER	Average encounter rate (birds / km)	numeric
mean.primary.PREV	Average prevalence (proportion of surveys with any birds)	numeric
primary.MAX	Maximum encounter rate (birds/km)	numeric
Weather		
MNTM.avg	Month-averaged mean temperature	numeric
DT90.sum	Number of days where maximum temperature > 90F	numeric
DT32.sum	Number of days where minimum temperature < 32F	numeric
DP01.sum	Number of days where rainfall > 0.1 inch	numeric
DP10.sum	Number of days where rainfall > 1 inch	numeric
TPCP.avg	Average total precipitation	numeric
Socio-economic predictors in home zip code		
edu_LessThanHighSchool	Proportion of population educated to less than High School standard	numeric
edu.University	Proportion of population educated to undergraduate level or above	numeric
laborforce_Unemployed	Proportion of population who are unemployed	numeric
Notlaborforce	Proportion of population not in labor force (retired, carers)	numeric
income_lt20G	Proportion of population that have an annual income < $20,000	numeric
income_gt100G	Proportion of population that have an annual income > $100,000	numeric
incomebelowpoverty_prop	Proportion of population whose income is below the poverty line	numeric

age_prop60plus	Proportion of population that are 60 or older	numeric

Statistical Analysis

We used machine learning methods to identify which factors are the best predictors for the two measures of retention. We decided to use machine learning methods primarily due to the large number of predictors and also the presence of missing values for some of the predictors (for example where volunteer zip-code was missing we were unable to assign any of the socio-economic predictors). We used the boosted regression tree analysis method in the "gbm" (Ridgeway *et al.* 2015) and "dismo" (Hijmans *et al.* 2015) packages in R version 3.2.1 (R core team 2015).

The boosted regression tree (BRT) method is based on the construction of decision trees to map a response onto a set of predictors. A decision tree consists of a sequence of binary partitions in the range of single or multiple predictor variables based on the identification of regions in predictor space that have the most homogenous response (**Figure 1**) (De'ath & Fabricius 2000; Hastie *et al.* 2001; Elith *et al.* 2008). When multiple explanatory variables are present each successive split can be implemented in the range of any of the explanatory variables, but being subject to splits higher in the tree (Elith *et al.* 2008). As a result, interactions between factors are modelled automatically and in a way that is simple to interpret. Boosted regression trees build on the decision tree framework but add several components to improve predictive capabilities (Elith *et al.* 2008). Boosting is a method that increases model accuracy by building and subsequently averaging many simple models in an iterative stagewise process (**Figure 1**). An initial decision tree is built that best reduces some loss function, such as deviance, that is usually a measure of predictive capability. The next tree is built on the residuals from the initial tree using the same loss criterion to identify the tree that best decreases the predictive deviance. The fitted values are then re-estimated due to the addition of the second tree and the residuals calculated. This process of building and adding trees continues in a stagewise fashion, until the final BRT model is a combination of all trees (usually ~ 1000's) (**Figure 1**).

There are four modelling parameters required by BRT models; tree complexity (*tc*), bag fraction (*bf*), learning rate (*lr*) and number of trees (*nt*). Each parameter controls an aspect of tree construction (*bf* and *tc*) or how they are combined (*nt* and *lr*) and can be altered to achieve optimal predictive performance. Tree complexity (*tc*) is equal to the number of nodes in each tree, with *tc* = 1 corresponding to trees consisting of single binary splits, and *tc* ≥ 2 corresponding to trees with multiple nodes, potentially allowing for interactions among parameters (*tc* = 2, can model two-way interactions between parameters) (**Figure 1**). Each individual tree constructed when fitting a BRT model is estimated based on a fraction of the data, known as the bag fraction (*bf*) (**Figure 1**). This introduces some variation among BRT models fitted to the same data, but provides benefits in the form of reduced over-fitting and improvements in model accuracy (Elith *et al.* 2008). Learning rate (*lr*) controls the contribution of each tree to the overall model. A low *lr* is required when introducing stochasticity into the modelling process (through bag fractions), to avoid overly large variation in predicted values between repeat modelling runs (Elith *et al.* 2008). However this also leads to a greater number of trees (*nt*) being required to achieve the lowest predictive deviance, as with a lower *lr* each individual tree explains less of the overall variation. The parameters *nt*, *lr* and *tc* are connected, with higher *tc* usually requiring lower *lr* and higher *nt* to achieve minimum predictive deviance. As with simple decision trees, over-fitting can be a problem, but in this case arises when too many trees are added. However, predictive accuracy can be optimised by constructing a model that overfits the data by fitting many more trees than are required to a training dataset. This model can then be used to make predictions for a test dataset (out-of-bag data not used to construct the model), but only using the first X_t trees fitted by the model. By incrementally increasing X_t from a minimum (e.g. the first 100 trees where predictive accuracy is low, but overfitting is unlikely to be an issue), through to the maximum number of trees and evaluating predictive accuracy at each value of X_t one can identify the point at which the model begins to model random noise, rather than signal. The number of trees identified using this method is then used to construct the final model, maximising predictive capacity without overfitting (**Figure 1**).

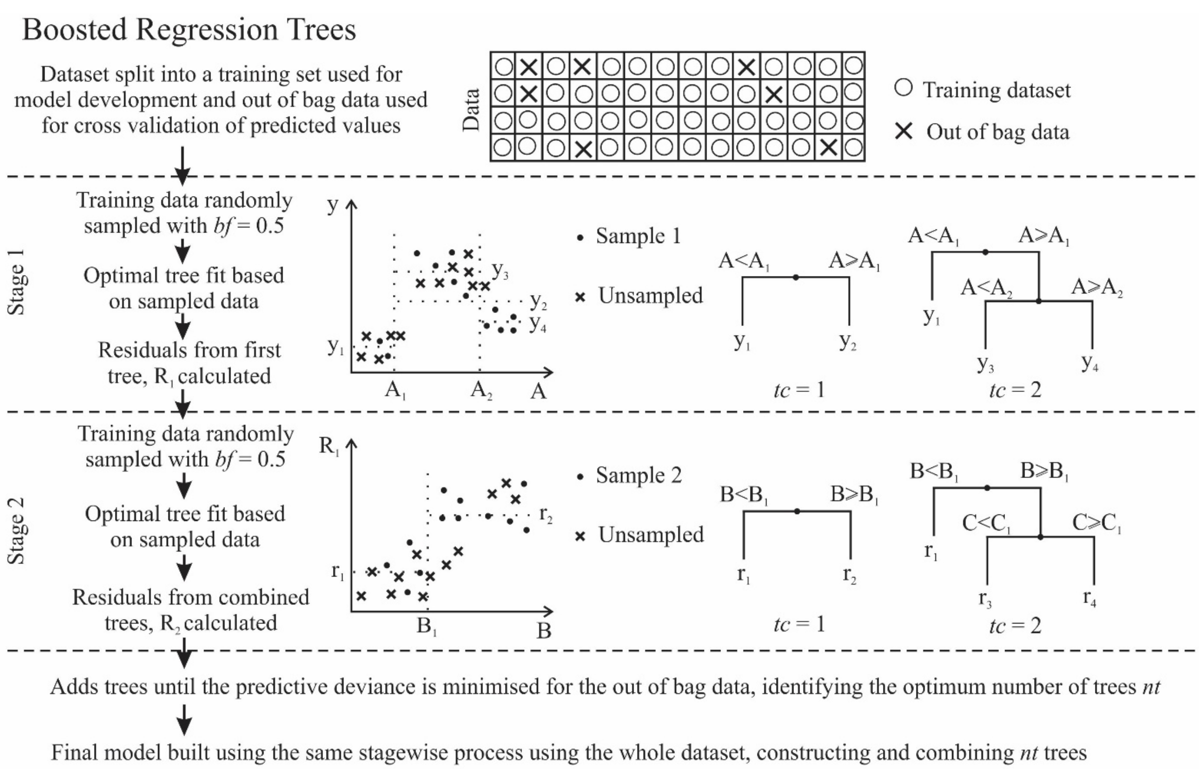

Figure 1. Illustrations of the processes involved in boosted regression tree analysis.

The response variables ***duration*** and ***1-year*** were modelled using BRT as a function of the predictors given in **Table 1**. The response ***duration*** was fourth-root transformed and modelled as a Gaussian response, whereas ***1-year*** was modelled as a Bernoulli (binomial) response. For each response variable every combination of *lr* (0.05, 0.01, 0.005, 0.002), *tc* (1, 2, 3, 4, 5) and *bf* (0.6, 0.65, 0.7, 0.75) was investigated. Each model was initially fitted with *nt* = 10,000. Model performance was assessed by splitting the dataset into training and test datasets. Training datasets consisted of 80% of the datapoints, which were used to construct the BRT models. The test dataset predictors were then fed into the BRT model to make predictions for the response variable of either ***duration*** or ***1-year***. Predicted values were then evaluated with respect to the observed values in the test dataset, and the proportion of deviance explained (D^2 = 1 – residual deviance/null deviance) was calculated. To optimise the models predictive capacity, for each combination of *lr*, *tc* and *bf* we evaluated D^2 by obtaining predictions from the BRT model at various numbers of trees (i.e. only using the first 1000 trees fitted by the model to make predictions). We did this for the first X_t number of trees, and varied X_t from 200 to 10,000 in increments of 50. This allowed us to identify the number of trees, *nt*, at which the BRT model begins overfitting relative to the test dataset, and hence identifies the point at which the model has the greatest predictive capacity. This procedure was evaluated on 30 different partitions of the dataset (i.e. 30 different training and test dataset combinations) for each combination of *lr*, *tc* and *bf* and D^2 values were averaged across the 30 repeats (Leathwick *et al.* 2006). The best combination of *lr*, *tc*, *bf* and *nt* was identified as the combination that had the highest D^2.

For the final model, the relative importance of each predictor (a measure of how often a predictor is chosen for splitting) is reported as well as the overall predictive capability (D^2) of the model. Partial dependency plots for the eight most important factors (determined by relative importance, but all partial dependency plots for each response variable are included in the Appendix) were plotted illustrating the response of retention across that predictors range, integrating across the response of all other predictors.

Results

Response variable – ***duration***

The BRT models fitted to the ***duration*** response variable achieved a maximum deviance explained, D^2, of 0.203, with model factors of *lr* = 0.005, *bf* = 0.75, *tc* = 5 and *nt* = 3027 (determined across 30 repeat runs) (**Table 2**). However, all models with *tc* ≥ 2 performed similarly with D^2 values ranging from 0.16 to 0.203, whereas models with *tc* = 1 had lower predictive ability, achieving D^2 values ranging from 0.13 to 0.141. For each of the ten best models there was a range in predictive ability from $D^2 \sim 0.10$ to 0.26 depending on how the data was partitioned, but in general models performed similarly with respect to predictive ability.

Table 2. Boosted regression tree model factors for the response variable ***duration*** along with the average proportion of deviance explained (D^2 = 1 – residual deviance/null deviance). Presented are the top ten models ranked by average D^2.

lr	*bf*	*tc*	*nt*	D^2
0.005	0.75	5	3027	0.203
0.01	0.75	5	4387	0.203
0.005	0.75	4	2440	0.203
0.01	0.75	4	5087	0.203
0.02	0.65	4	3947	0.202
0.002	0.75	5	4487	0.202
0.02	0.75	3	5707	0.202
0.002	0.75	4	2780	0.202
0.01	0.7	4	2327	0.202
0.01	0.7	5	4993	0.202

Of the model predictors, personal and survey related factors including average group size, average travel time, and the age at which the individual was trained came out as the most influential based on relative influence scores (a measure of how often a variable is chosen to be split) (**Table 3**). Factors related to aspects of the beach, including beach length, access type, mean encounter rate and prevalence were also relatively influential, but less so than survey related factors (**Table 3**). Predictors related to average climate were relatively less influential, with average temperature (MNTM.avg) and the number of days receiving a minimum of 0.1 inches of rain (DP01.sum) being the most influential in explaining ***duration*** of the six weather predictors trialled (**Table 3**). Socio-economic factors were also relatively uninfluential, with the exception of the proportion of individuals aged 60 or more in the home zip code, which was the 4th most influential predictor of ***duration*** (**Table 3**).

Table 3. Summary of the relative influence of predictors for BRT models fitted to the response variable ***duration***. Relative influence scores are presented as the mean and the range (minimum and maximum) across the 30 data partitions the models were evaluated on.

Variable		Relative Influence	
Name	Type	Mean	Range
Ave.Groupsize	Personal/Survey	16.9	13.2 - 22
Avg.Travel	Personal/Survey	9.9	8 - 11.7
Trained.Age	Personal	8.9	6.2 - 12.6
age_prop60plus	Socio-economic	6.8	4.1 - 10.7
Length.beach	Beach	6.2	3.3 - 8.2
Access.beach	Beach	4.8	2.8 - 6.4
Ave.SurveyTime	Personal/Survey	4.5	3.1 - 6.5
mean.primary.ER	Beach	4.0	2.4 - 5.5
Involvement.1	Personal	3.1	1.5 - 4.4

mean.primary.PREV	Beach	3.1	1.4 - 4.3
MNTM.avg	Weather	3.1	1.8 - 4.9
income_lt20G	Socio-economic	2.8	1.7 - 5
primary.MAX	Beach	2.6	1.5 - 3.9
DP01.sum	Weather	2.4	1.4 - 4
income_gt100G	Socio-economic	2.3	1.2 - 3.3
edu_LessThanHighSchool	Socio-economic	2.2	0.9 - 3.1
TPCP.avg	Weather	2.0	1 - 3.5
laborforce_Unemployed	Socio-economic	2.0	0.9 - 3
DP10.sum	Weather	1.9	0.8 - 2.7
Notlaborforce	Socio-economic	1.8	0.8 - 2.9
edu.University	Socio-economic	1.7	0.9 - 2.4
DT32.sum	Weather	1.7	0.9 - 2.8
DT90.sum	Weather	1.6	0.8 - 2.6
Width.beach	Beach	1.3	0.4 - 2.2
Gender	Personal	1.1	0.3 - 4.1
incomebelowpoverty_prop	Socio-economic	0.9	0.2 - 1.6
Substrate.beach	Beach	0.6	0.2 - 1.6

The response profile for average group-size shows that individuals that always perform surveys alone (group-size = 1) have a lower retention than individuals who have at any point performed surveys with others (group-size > 1) (**Figure 2**). However, individuals performing surveys with group-sizes ~ 2 have on average the highest retention, with a sharp decrease in retention for individuals who participate in surveys in group-sizes > 2 (**Figure 2**). The response for average travel time is complex, but shows that individuals who travel for less than 30 minutes have a higher retention than individuals who travel for 30-60 minutes (**Figure 2**). Beyond a travel time of 60 minutes there is an increase in retention up to a peak at ~ 160 minutes, and then decreases from 160-230 minutes (**Figure 2**). This second peak at ~160 minutes is perhaps suggestive of a group of volunteers that are very dedicated and are likely to travel great distances, and are likely to remain as volunteers for longer (**Figure 2**). The response of age when trained shows that younger individuals (age < 30) have a lower average retention than individuals aged 30 +, with a general increase in retention as age increases (**Figure 2**). Similarly, the proportion of the population aged 60+ has a positive effect on retention, perhaps due to a larger community of active volunteers leading to a greater feeling of engagement within the community (**Figure 2**). The response of beach length with respect to retention show that individuals surveying shorter beaches have lower retention than those surveying longer beaches (**Figure 2**). The effect of beach access illustrates that individuals that have to drive (responses of drive and walk/drive), to get to their beach have a lower retention on average than individuals using any other mode of transport (**Figure 2**). The response for average survey time shows that individuals carrying out shorter surveys (< 60 minutes) have a lower retention than individuals carrying out surveys averaging 60 – 110 minutes, with a slight reduction in retention for individuals carrying out surveys averaging > 110 minutes (**Figure 2**).

Finally, individuals surveying beaches with lower bird encounter rates (0 – 2 birds/km) have higher retention than individuals surveying beaches with encounter rates > 2 birds/km (**Figure 2**).

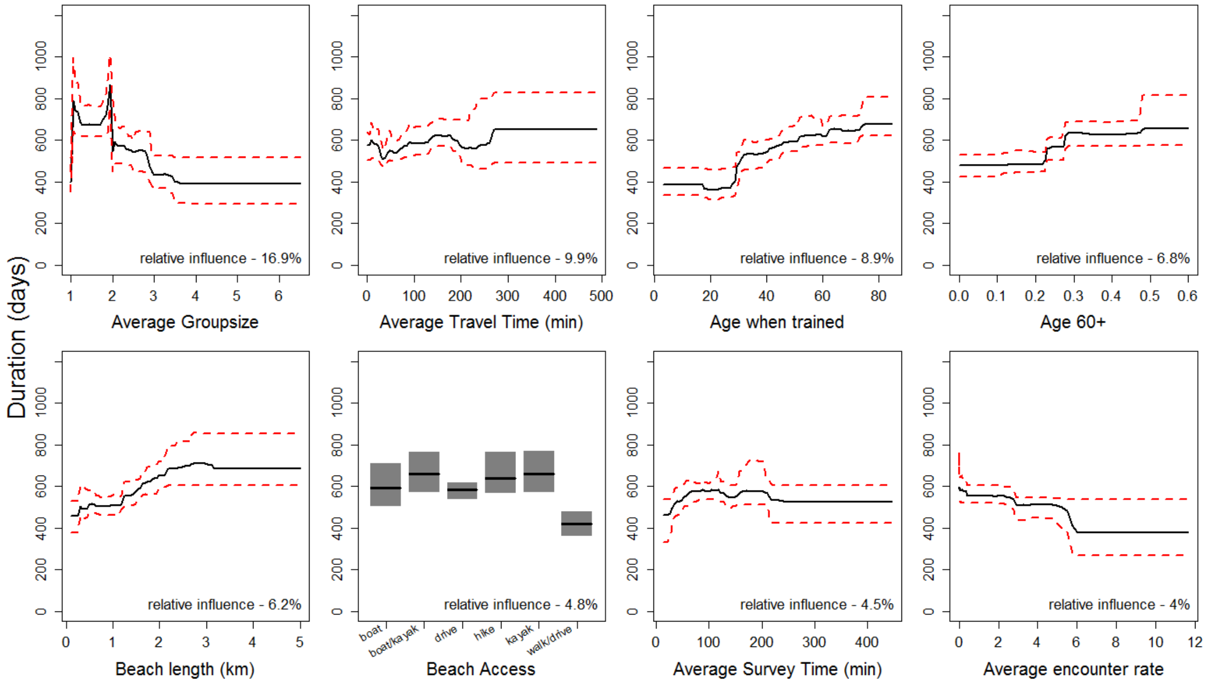

Figure 2. Response profiles for the eight most influential predictors for the BRT model fitted to the response variable, ***duration***. Response profiles represent the average response for each variable integrated across the response of all other predictors. The black line illustrates the average response profile across the 30 BRT models fitted to each of the 30 data partitions, and red dotted lines, and grey boxes for factorial variables, show the maximum and minimum predicted responses.

Response variable – ***1-year***

The BRT models fitted to the ***1-year*** response variable achieved a maximum deviance explained, D^2, of 0.139, with model factors of lr = 0.002, bf = 0.65, tc = 5 and nt = 2612 (determined across 30 repeat runs) (**Table 4**). However, all models with $tc \geq 2$ performed similarly with D^2 values ranging from 0.135 to 0.139, whereas models with tc = 1 had lower predictive ability, achieving D^2 values ranging from 0.125 to 0.128.

Table 4. Boosted regression tree model factors for the response variable ***1-year*** along with the average proportion of deviance explained (D^2 = 1 – residual deviance/null deviance). Presented are the top ten models ranked by average D^2.

lr	*bf*	*tc*	*nt*	D^2
0.002	0.65	5	2612	0.139
0.002	0.65	4	1272	0.139
0.01	0.7	4	3078	0.139
0.002	0.7	5	2538	0.139
0.005	0.75	5	1565	0.139
0.005	0.7	5	667	0.139

0.002	0.75	4	1200	0.139
0.01	0.75	4	620	0.139
0.005	0.75	4	2910	0.139
0.005	0.65	4	2438	0.139

The relative influence of predictors were similar for the response variable ***1-year*** as they were for the response variable ***duration***. The exceptions are that age when trained is now the most influential predictor and the number of days receiving 0.1 inches of rainfall is the 5th most influential predictor of ***1-year*** compared to 14th most influential predictor of ***duration***. In addition, beach access goes from 6th most influential predictor of duration to 24th most influential predictor of ***1-year*** (**Table 5**).

Table 5. Summary of the relative influence of predictors for BRT models fitted to the response variable ***1-year***. Relative influence scores are presented as the mean and the range (minimum and maximum) across the 30 data partitions the models were evaluated on.

Variable		Relative Influence	
Name	Type	Mean	Range
Trained.Age	Personal	13.9	11.7 - 19.7
Ave.Groupsize	Personal/Survey	11.5	8.8 - 13.6
Avg.Travel	Personal/Survey	11.1	9.7 - 13.1
Length.beach	Beach	6.0	4.1 - 8
DP01.sum	Weather	5.9	3.9 - 9.9
age_prop60plus	Socio-economic	4.5	2.3 - 8.3
Ave.SurveyTime	Personal/Survey	3.9	1.7 - 5.3
mean.primary.ER	Beach	3.3	1.9 - 4.4
MNTM.avg	Weather	3.1	2.3 - 4.8
Involvement.1	Personal	3.0	1.5 - 4.4
edu_LessThanHighSchool	Socio-economic	3.0	0.9 - 4.2
Notlaborforce	Socio-economic	3.0	2.1 - 3.9
laborforce_Unemployed	Socio-economic	2.7	1.3 - 5
income_lt20G	Socio-economic	2.5	1.4 - 3.3
primary.MAX	Beach	2.4	1.6 - 3.9
mean.primary.PREV	Beach	2.3	1.6 - 3.7
DT90.sum	Weather	2.3	1.5 - 3.7
DT32.sum	Weather	2.3	1.6 - 3.8
Width.beach	Beach	1.9	1.1 - 2.7
Substrate.beach	Beach	1.9	0.3 - 3.2
edu.University	Socio-economic	1.7	1 - 2.6
income_gt100G	Socio-economic	1.5	0.7 - 2.3
incomebelowpoverty_prop	Socio-economic	1.5	0.6 - 2.3
Access.beach	Beach	1.3	0.5 - 2.3
Gender	Personal	1.3	0.2 - 3.2

TPCP.avg	Weather	1.1	0.4 - 1.6
DP10.sum	Weather	1.1	0.7 - 1.7

Response profiles for age when trained, average group-size, beach length, proportion of community aged 60 or more and average encounter rate were similar between duration and 1-year (**Figure 3**). The response profile for average travel time is similar between ***1-year*** and ***duration***, with the exception that individuals that on average travel for > 180 minutes are less likely to volunteer for a year or more, relative to those that have a lower average travel time (**Figure 3**). The response profile for average survey time is also similar between 1-year and duration, with the exception that individuals performing surveys that take an average of 200 minutes or more being less likely to volunteer for a year or more, relative to those that have shorter survey times (**Figure 3**). Finally, with respect to rainfall days (days with > 0.1 inch of rain), individuals carrying out surveys at locations with a low number of rainfall days are less likely to volunteer for a year or more, relative to individuals surveying beaches that receive more rainfall days (**Figure 3**).

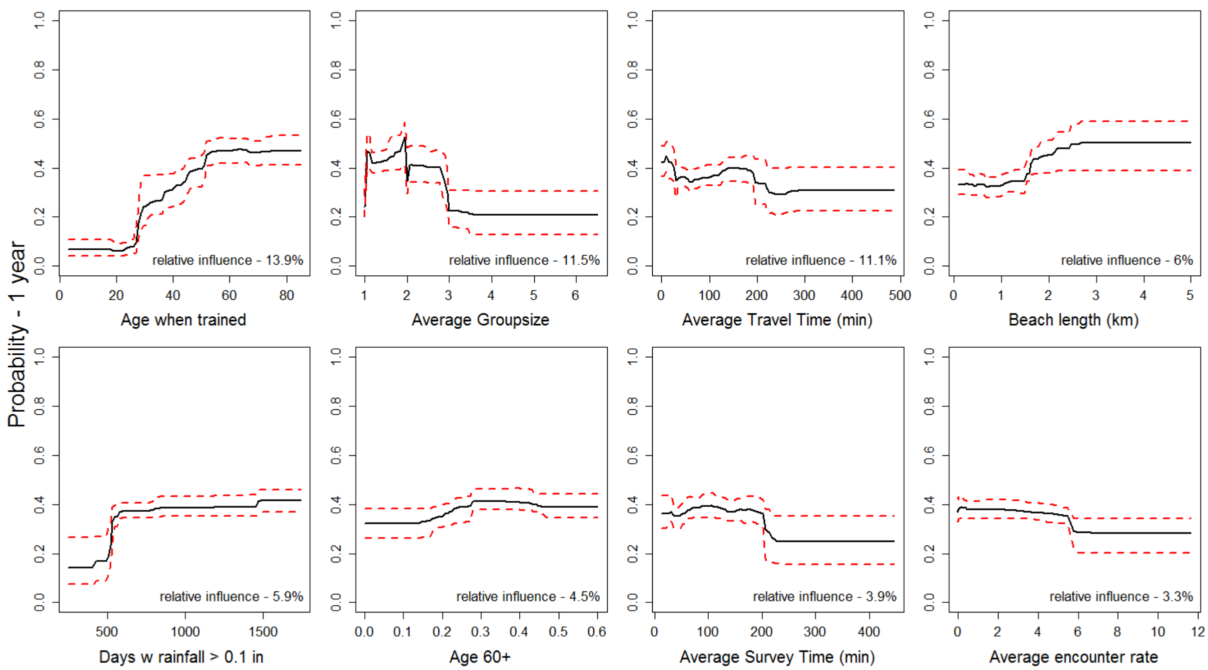

Figure 3. Response profiles for the eight most influential predictors for the BRT model fitted to the response variable, ***1-year***. Response profiles represent the average response, (in terms of the probability that individuals volunteer for a year or more) for each variable integrated across the response of all other predictors. The black line illustrates the average response profile across the 30 BRT models fitted to each of the 30 data partitions, and red dotted lines show the maximum and minimum predicted responses.

References

De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, **81**, 3178-3192.

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802-813.

Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning: data mining, inference and prediction. Springer-Verlag, New York.

Hijmans RJ, Phillips S, Leathwick J, Elith J (2015). dismo: Species Distribution Modeling. R package version 1.0-12. http://CRAN.R-project.org/package=dismo

Leathwick JR, Elith J, Francis MP, Hastie T, Taylor P (2006) Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series*, **321**, 267-281.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ridgeway G, et al. (2015). gbm: Generalized Boosted Regression Models. R package version 2.1.1. http://CRAN.R-project.org/package=gbm.

Appendix

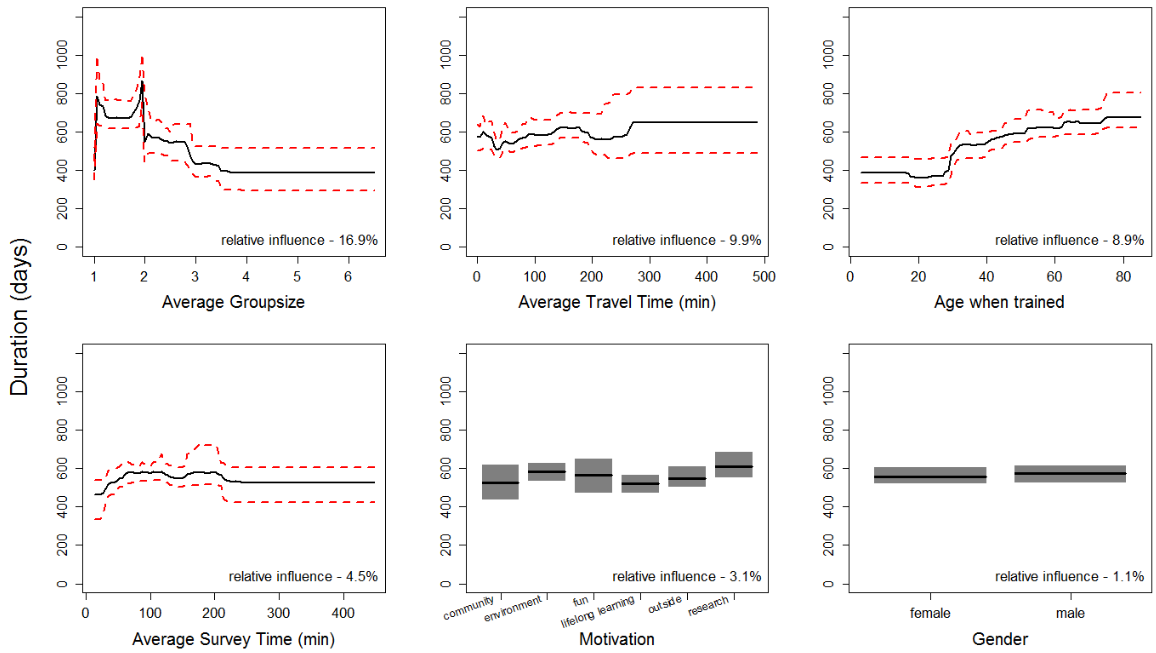

Figure A1. Response profiles for the personal/survey related predictors for the BRT model fitted to the response variable, ***duration***. Response profiles represent the average response for each variable integrated across the response of all other predictors. The black line illustrates the average response profile across the 30 BRT models fitted to each of the 30 data partitions, and red dotted lines, and grey boxes for factorial variables, show the maximum and minimum predicted responses.

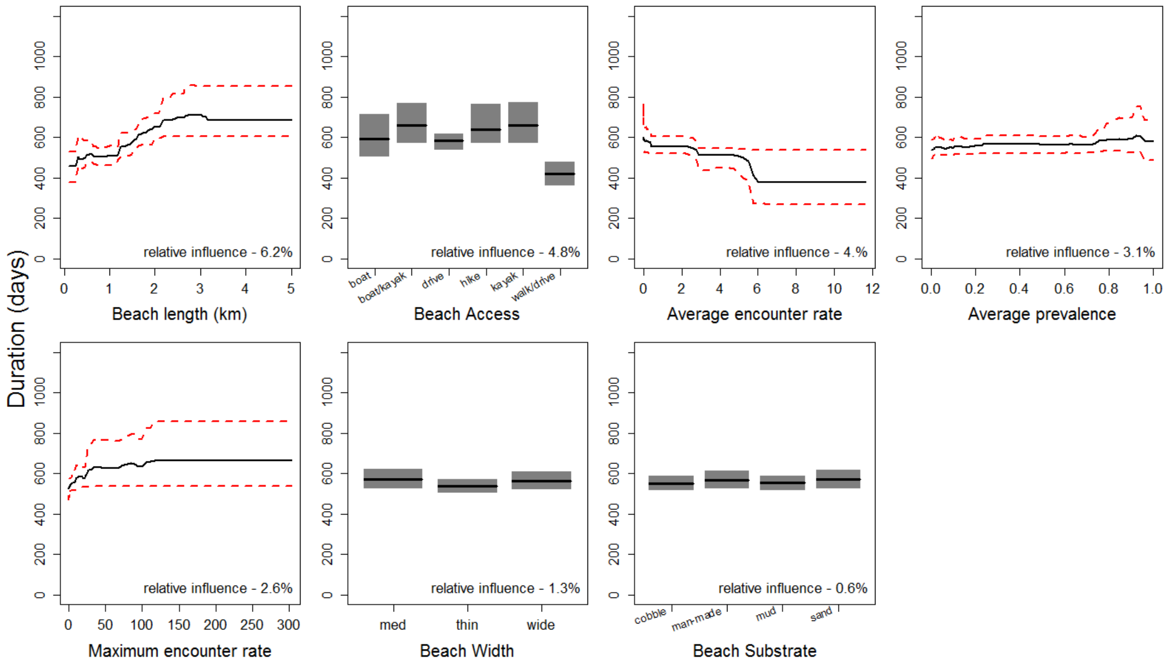

Figure A2. Response profiles for the beach related predictors for the BRT model fitted to the response variable, ***duration***. Response profiles represent the average response for each variable integrated across the response of all other predictors. The black line illustrates the average response profile across the 30 BRT models fitted to each of the 30 data partitions, and red dotted lines, and grey boxes for factorial variables, show the maximum and minimum predicted responses.

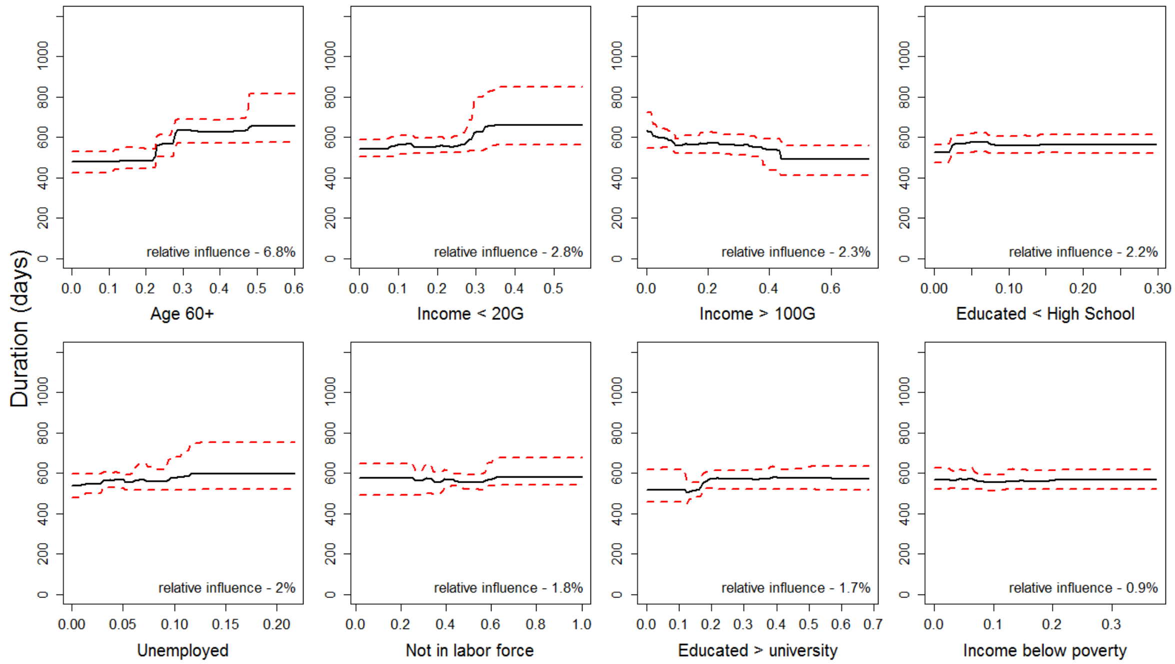

Figure A3. Response profiles for the socio-economic predictors for the BRT model fitted to the response variable, ***duration***. Response profiles represent the average response for each variable integrated across the response of all other predictors. The black line illustrates the average response profile across the 30 BRT models fitted to each of the 30 data partitions, and red dotted lines, and grey boxes for factorial variables, show the maximum and minimum predicted responses.

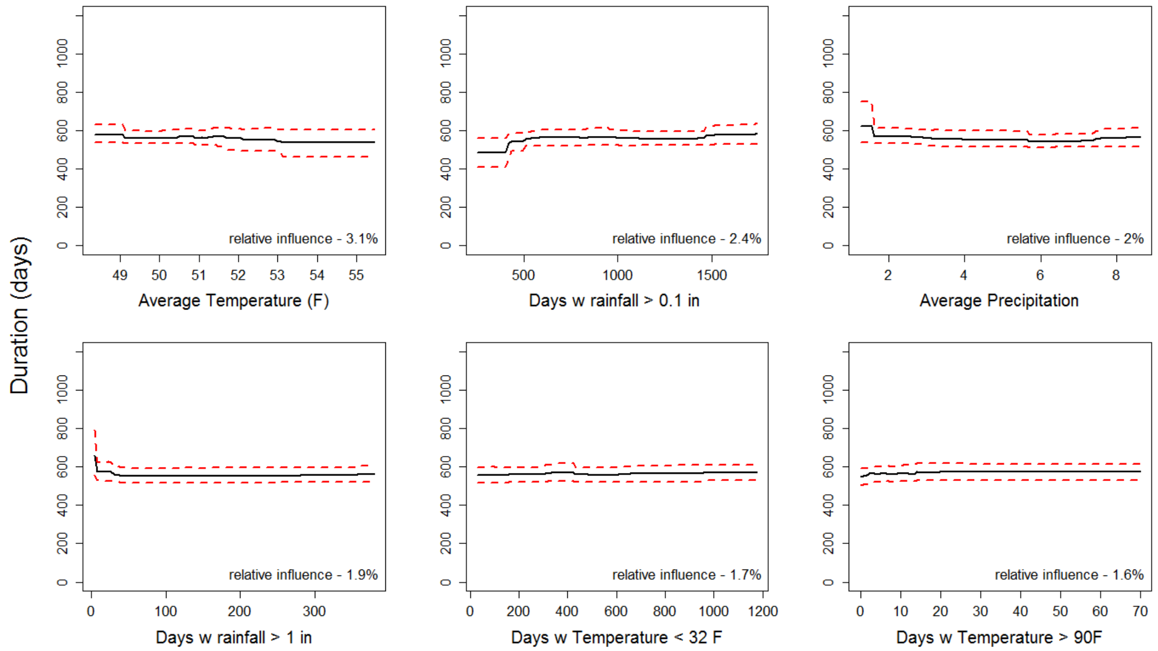

Figure A4. Response profiles for the weather related predictors for the BRT model fitted to the response variable, ***duration***. Response profiles represent the average response for each variable integrated across the response of all other predictors. The black line illustrates the average response profile across the 30 BRT models fitted to each of the 30 data partitions, and red dotted lines, and grey boxes for factorial variables, show the maximum and minimum predicted responses.

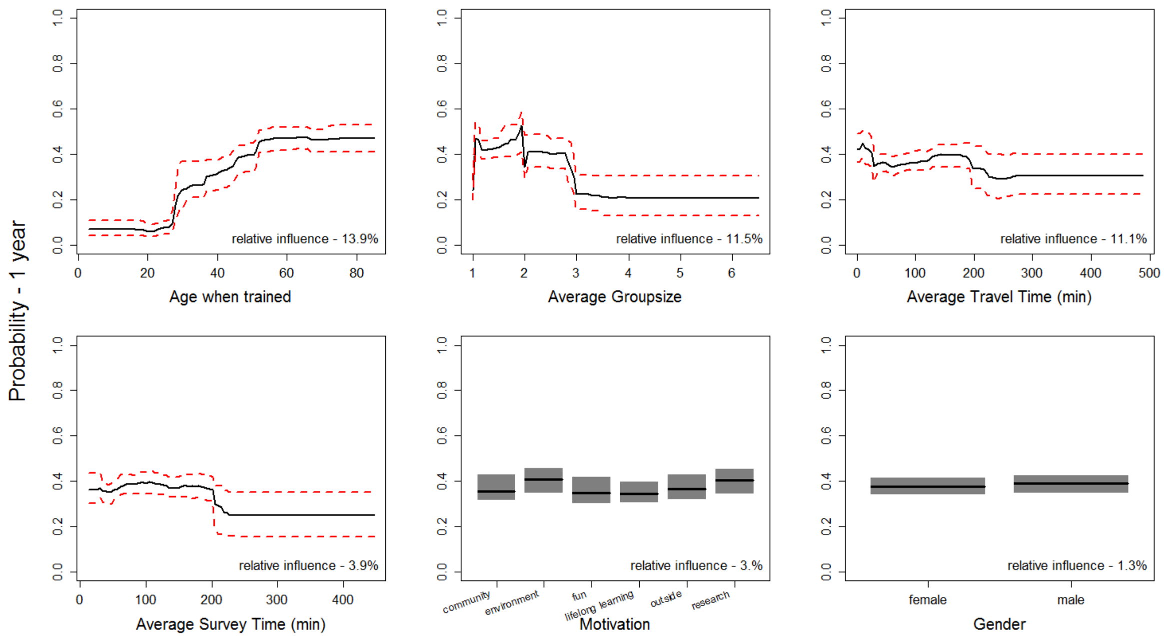

Figure B1. Response profiles for the personal/survey related predictors for the BRT model fitted to the response variable, ***1-year***. Response profiles represent the average response, (in terms of the probability that individuals volunteer for a year or more) for each variable integrated across the response of all other predictors. The black line illustrates the average response profile across the 30 BRT models fitted to each of the 30 data partitions, and red dotted lines, and grey boxes for factorial variables, show the maximum and minimum predicted responses.

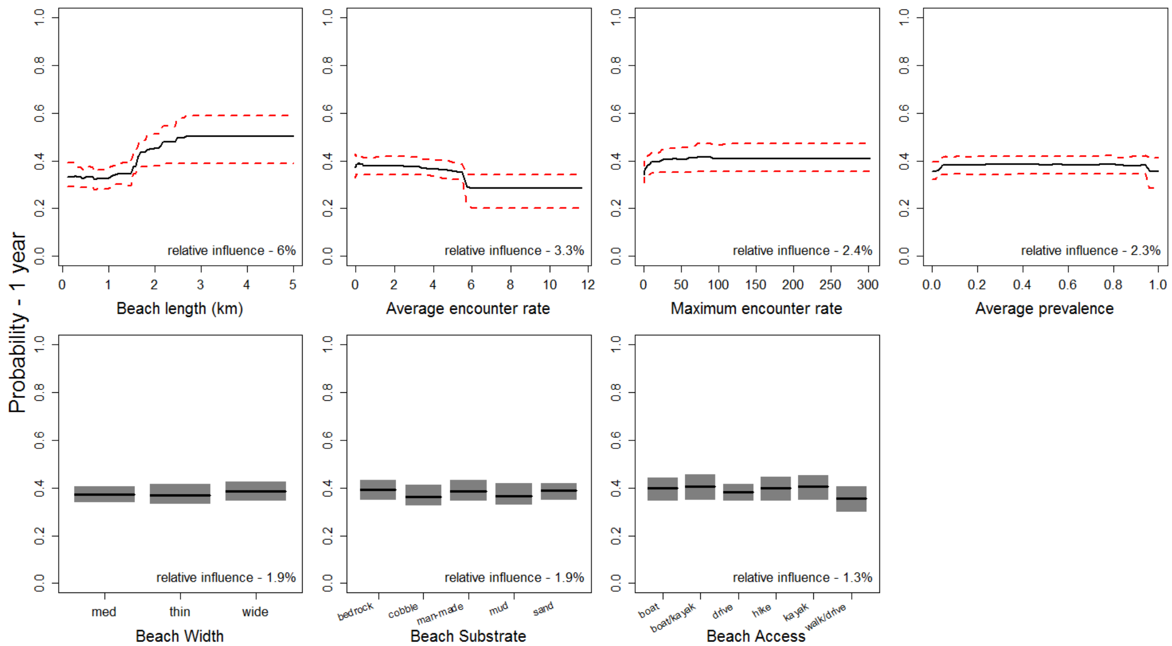

Figure B2. Response profiles for the beach related predictors for the BRT model fitted to the response variable, ***1-year***. Response profiles represent the average response, (in terms of the probability that individuals volunteer for a year or more) for each variable integrated across the response of all other predictors. The black line illustrates the average response profile across the 30 BRT models fitted to each of the 30 data partitions, and red dotted lines, and grey boxes for factorial variables, show the maximum and minimum predicted responses.

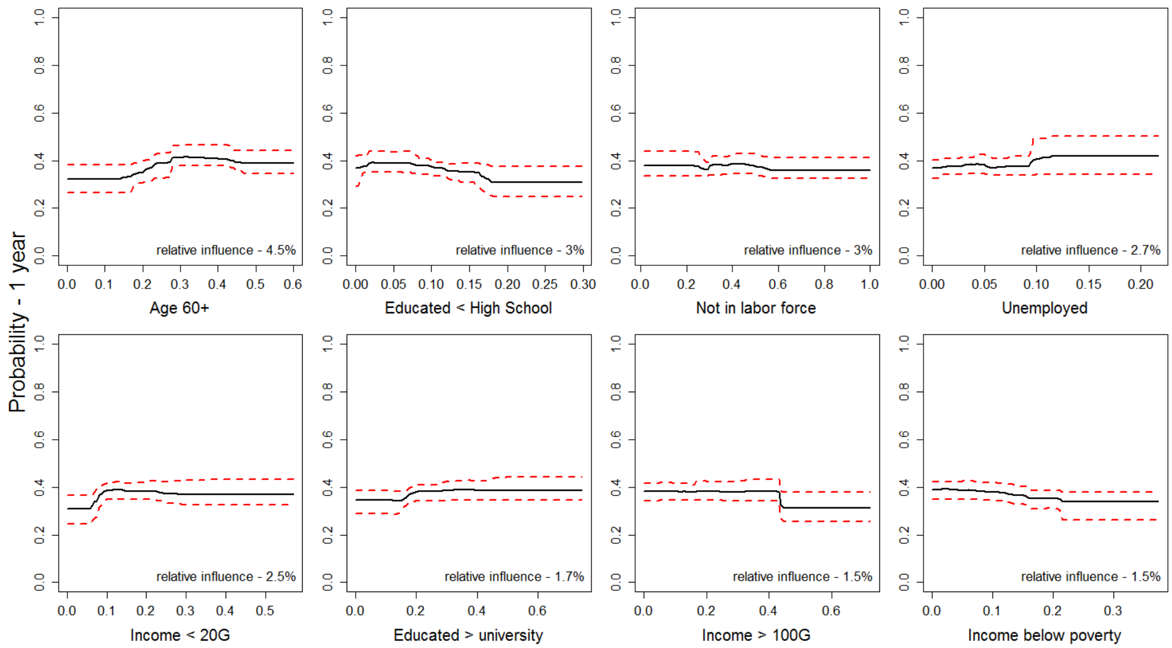

Figure B3. Response profiles for the socio-economic predictors for the BRT model fitted to the response variable, ***1-year***. Response profiles represent the average response, (in terms of the probability that individuals volunteer for a year or more) for each variable integrated across the response of all other predictors. The black line illustrates the average response profile across the 30 BRT models fitted to each of the 30 data partitions, and red dotted lines, and grey boxes for factorial variables, show the maximum and minimum predicted responses.

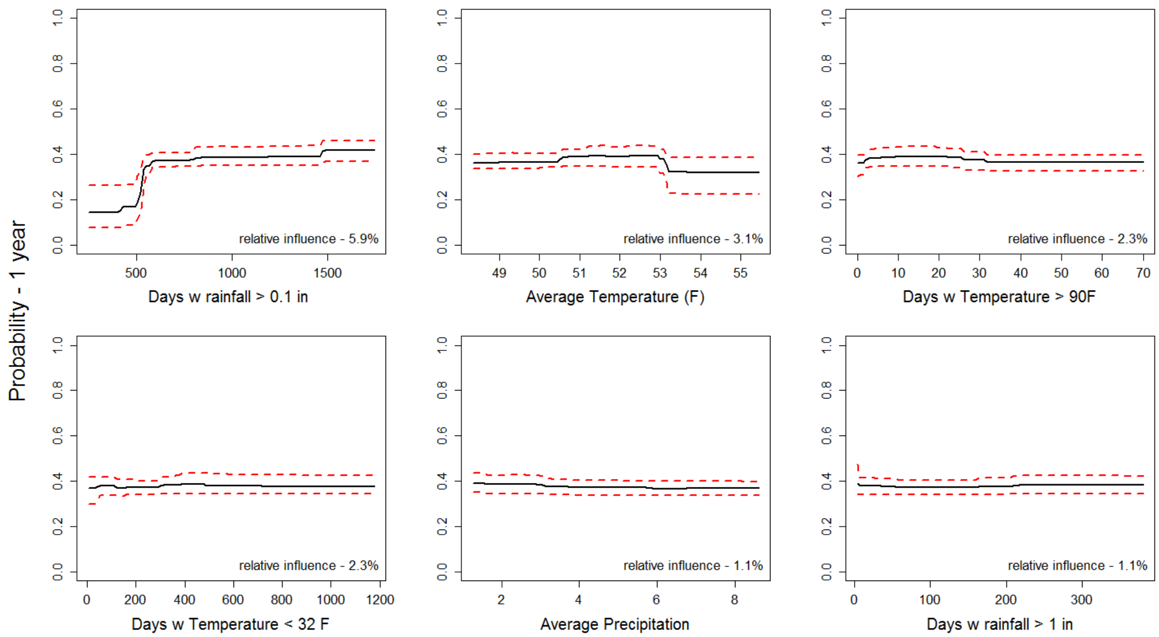

Figure B2. Response profiles for the weather related predictors for the BRT model fitted to the response variable, ***1-year***. Response profiles represent the average response, (in terms of the probability that individuals volunteer for a year or more) for each variable integrated across the response of all other predictors. The black line illustrates the average response profile across the 30 BRT models fitted to each of the 30 data partitions, and red dotted lines, and grey boxes for factorial variables, show the maximum and minimum predicted responses.