23rd International Congress on Modelling and Simulation, Canberra, ACT, Australia, 1 to 6 December 2019
mssanz.org.au/modsim2019

Where do we recreate? Comparison of different methods to determine importance of site characteristics for outdoor recreation

F. Willibald[a,b], **M.J. van Strien**[a] **and A. Grêt-Regamey**[a]

[a] *Planning of Landscape and Urban Systems, Institute for Spatial and Landscape Planning, ETH Zurich, Zurich, Switzerland*

[b] *Institute of Science, Technology and Policy, ETH Zurich, Zurich, Switzerland*

Email: fabian.willibald@istp.ethz.ch

Abstract: Residents of industrialized countries place increasingly more value on leisure time. For many alpine and remote municipalities, revenues from tourism and recreation belong to the most important sources of income. At the same time, these activities generate large carbon footprints through traffic. Identifying the drivers of demand for outdoor recreation is essential for a sustainable future transport and landscape planning. By using different types of regression models, we compared four different methods to determine the importance of variables quantifying landscape characteristics for explaining outdoor recreation day trips.

The regression models were:

- Generalized Linear Model of the Poisson family
- Random Forest regression
- Gradient Boosting regression

The applied variable importance measures were:

- Perturbation variable importance
- Hierarchical Partitioning
- Gini Index
- Gradient Boosting node impurity

The overall pattern of variable importance complies between the different regression models, but there are also some obvious differences. Two variables, namely population density and the number of land use counts dominate variable importance for all methods and are the by far most important predictors. Variables of medium importance can hardly be identified. There are only fractional differences between the importance indices of the low importance variables. Therefore, the overall ranking of low importance variables differs substantially between different models.

Keywords: *Random Forest, Gradient Boosting, Generalized Linear Models, variable importance*

1. INTRODUCTION

The demand for many outdoor recreation activities has steadily increased in the last decades (Cordell, 2012). A healthy work-life balance plays an increasing role in daily routines of residents in industrialized countries (Costanza et al., 2007). Further, population growth as well as demographic changes and wealth have significantly changed the demand for recreation (Lee and Schuett, 2014). While a large share of the European population lives in urban areas (United Nations, 2017), open land, mountainous areas and untouched nature are desired places for leisure and contact with nature (Ode and Fry, 2006). Revenues from recreational activities and tourism play a major role in the economic prosperity of many alpine and agricultural municipalities (Sen et al., 2011, Schägner et al., 2017), but they also put pressure on natural ecosystems and can cause severe environmental impacts (Buckley, 2004, McCullough et al., 2018). Furthermore, leisure activities are leading to large carbon footprints due to emissions, mainly from transportation (Druckman and Jackson, 2010). In Switzerland, for example, 50% of person-kilometers travelled are due to leisure activities (Federal Statistical Office, 2015).

For spatial- and transport planners, tourism dependent regions and policy makers, it is thus important to understand peoples' destination choice for outdoor recreation. Therefore, it is crucial to identify the drivers of demand and to understand which site characteristics are beneficial for recreational activities. Still, most regression models are not specifically suited to determine variable importance (Grömping, 2015). For that reason, many different methods have been suggested.

In this contribution, we compare four different methodologies to determine variable importance for the demand of outdoor recreation day trips. We present different variable importance measures for regression models, among them Generalized Linear Models (GLM) as well as machine learning algorithms like Random Forest regression (RF) and Gradient Boosting regression (GB).

2. METHODS AND DATA

2.1. Data on outdoor trips

The spatial scope of our study is Switzerland. The dependent variable in our models was the number of trips undertaken to Swiss municipalities for the purpose of outdoor recreation (i.e. count data), which were derived from revealed preference surveys: i.e. the Swiss National Microcensus on mobility and transport for the years 2010 and 2015 (Federal Statistical Office, 2012, Federal Statistical Office, 2017). The Microcensus is a Swiss-wide continuous survey, conducted at equal frequency all throughout the year, in which about one percent of the Swiss population is interviewed about their daily travel behavior. In an additional module, about one third of all respondents are asked about day trips. Day trips were defined as trips during which people leave their familiar surroundings for minimum three hours. Each respondent was asked to indicate the purposes and target municipalities of a maximum of three day trips that were undertaken during the last two weeks prior to the interview (Federal Statistical Office, 2017). We chose this dataset, because day trips, unlike over-night trips, are undertaken more frequently and therefore comprise a large share of CO_2 emissions. We included only destinations to which people went for activities, such as biking, hiking, sports activities and other outdoor activities.

2.2. Selection of explanatory variables

We conducted a thorough literature review in scientific databases (e.g. *Web of Knowledge, ScienceDirect*) to select landscape variables that can explain outdoor recreation demand. We selected publications that focus on recreational activities and that address the role of site characteristics in influencing peoples' choice of nearby or holiday outdoor recreation destinations. Although multiple studies examine outdoor recreation demand, they mostly focus on nearby recreation or target specific nature areas like national parks (Schägner et al., 2016), forests (Agimass et al., 2018), or urban parks and greening (Caspersen and Olafsson, 2010). Most studies used between eight and twenty explanatory variables for their statistical analysis. From the literature analysis, we arranged the predictor variables into eight thematic groups, namely (1) settlement, (2) road network and accessibility, (3) roughness and aspect, (4) infrastructures for outdoor activities (hiking, biking, skiing), (5) streams and rivers, (6) lakes, (7) woodlands and (8) land use/-cover (Willibald et al., in review). In a next step, we eliminated highly positively or negatively correlated variables (absolute Pearson $r > 0.65$). Finally, we retained ten variables that we supposed to have significant influence on outdoor recreation demand. The data was processed for entire Switzerland and aggregated to the municipality scale. To account for the size differences between the municipalities (which can be substantial), we normalized, where appropriate, each variable to a mean per square-kilometer. By doing so variations between municipalities are well represented (Willibald et al., in review).

Table 1. List of final explanatory variables for the statistical analysis. For each group, identified from the literature review, at least one variable was chosen. Adapted from Willibald et al. (in review).

Groups	Variable	Data Source	Spatial resolution
Settlement	population density [count/km^2]	Federal Statistical Office Statpop 2015	100 m (raster)
Road network and accessibility	population accessibility	Federal Office for Spatial Development NPVM 2005	3-8 m (vector)
Roughness and aspect	absolute difference in altitude [m]	European Environmental Agency	25 m (digital elevation model)
Infrastructures for activities (hiking, biking, skiing)	density of hiking trails [m/km^2]	Federal Office of Topography Swisstopo	3-8 m (vector)
	length of ski slopes [m/km^2]	Federal Office of Topography Swisstopo	3-8 m (vector)
Streams and rivers	density of streams and rivers [m/km^2]	Federal Office of Topography Swisstopo	3-8 m (vector)
Lakes	lakeshore density [m/km^2]	Federal Office of Topography Swisstopo	3-8 m (vector)
Woodlands	share of forest [%]	Corine Landcover 2012	200 m (raster)
Land-use/-cover	number of land-cover classes per km^2 [count/km^2]	Corine Landcover 2012	200 m (raster)
	distance to protected areas [km]	Swisstopo	3-8 m (vector)

2.3. Model Selection and Goodness of Fit

We compared variable importance measures for three different kinds of regression models. In each case, the dependent variable consisted of count data. The applied regression models are a GLM from the Poisson family (negative binomial) (Cameron and Trivedi, 2013, Zeileis et al., 2008), the Random Forest algorithm (Breiman, 2001) and Gradient Boosting (Elith et al., 2008). For more details on the regression models we refer to the references cited above.

For each of the models, we calculated the goodness of fit. This is important, as variable importance measures of a specific model are deemed less meaningful, if its model fit is considerably poorer in comparison to other models. Finding a goodness of fit measure that is comparable for multiple different regression models is a challenging task. This is further complicated by the fact that many traditional goodness of fit measures like R-squared are not practical for count data. Finally, we used root mean square error (RMSE) and mean absolute error (MAE) to compare the fit of the three models.

2.4. Variable Importance Measures

Overall, we compared four different variable importance measures. Three of those methods were applied to one of the corresponding regression models, while the fourth measure was applied to all regression models.

Permutation based variable importance

To all three models, we applied a permutation based approach to assess variable importance. This approach can be applied to both, machine learning techniques like RF (Breiman, 2001), but also to any other regression model. The approach calculates variable importance by iteratively randomly shuffling all explanatory variables and comparing their impact on goodness of fit (in this study based on AIC for the negative binomial) or model error (MSE for RF and GB) before and after shuffling. A feature is considered important if shuffling its values decreases goodness of fit criteria (increases AIC) or increases model error compared to the unpermuted model (i.e. baseline model) (Wei et al., 2015).

Variable importance from hierarchical partitioning

The theorem of hierarchical partitioning was first introduced by Chevan and Sutherland (1991): hierarchical partitioning estimates relative importance of an explanatory variable by computing goodness of fit of all models containing a particular variable to the fit of all nested models lacking that variable. The variable's contribution to the dependent variable is averaged over all possible combinations of explanatory variables to determine the relative importance (Murray and Conner, 2009). As goodness of fit measure, we used root mean square prediction error (RMSPE). Hierarchical partitioning was only applied to the GLM.

Variable importance from Gini importance

For the Random Forest algorithm developed by Breiman (2001), a second measure to compute variable importance was the Gini impurity index. A variable is considered important, if using this variable for splitting a tree leads to a large decrease in node impurity. The impurity decrease is averaged over all nodes where a variable was used for splitting the tree to determine the relative variable importance. Impurity is measured by the Gini Index (Nembrini et al., 2018).

Variable Importance from node impurity

For Gradient Boosting regression, the relative importance for a single tree is based on the number of times a variable is selected for splitting at a node, weighted by the squared improvement to the model as a result of each split. This is averaged over all trees to learn the overall contribution of each variable. Since these measures are relative, the contribution of each variable can be scaled, so that the influence of all variables adds up to 100 (Elith et al., 2008, Friedman and Meulman, 2003).

3. RESULTS

3.1. Goodness of Fit

A comparison of the goodness of fit shows that the negative binomial GLM and RF model perform similarly well. The GB model has a significantly lower error compared to those models (Table 2).

Table 2. RMSE and MAE of the three different regression models

	GLM (neg. bin)	RF	GB
RMSE	2.35	2.2	1.2
MAE	1.03	1	0.66

3.2. Variable Importance

Figures 1 and 2 visualize the results of variable importance for the different methodologies. While Figure 1 shows the absolute values of the variable importance indices, Figure 2 compares the overall rankings of the different landscape variables. The overall pattern of variable importance is quite similar for all applied methods and we can roughly divide the variables into two groups of importance: high and low. All methods have in common that population density, the number of land use classes and the length of ski slopes are the three most important predictors. However, the order of these three variables differs per method. While population density is the most important variable for the GLM and GB permutation based variable importance, the number of land use classes is the most important variable for the remaining variable importance measures. Length of ski slopes is for five of six methods ranked as the third most important variable. Only for hierarchical partitioning it is found to be the second most important variable.

In five of the six variable importance rankings, the remaining variables can be considered of low importance. Only for the RF permutation variable importance we can identify a group of medium importance. Surface Roughness and population accessibility can be assigned to this group of medium importance. We can observe that the importance indices of the low importance variables vary only very minor. For that reason, we see quite large deviations in the ranking of the low importance variables over different methodologies (Fig. 2). E.g. lakeshore density ranks between four and seven.

To conclude, we can summarize that there exist three variables that dominate variable importance over all methods, while most of the remaining variables are only of minor importance and follow no strict hierarchical ranking. Population density and the number of land use counts are the main predictors to identify hotspots of outdoor recreation demand for day trips, while other variables that are considered of large importance in other studies, among them accessibility or the share of forest, are found to be of low importance.

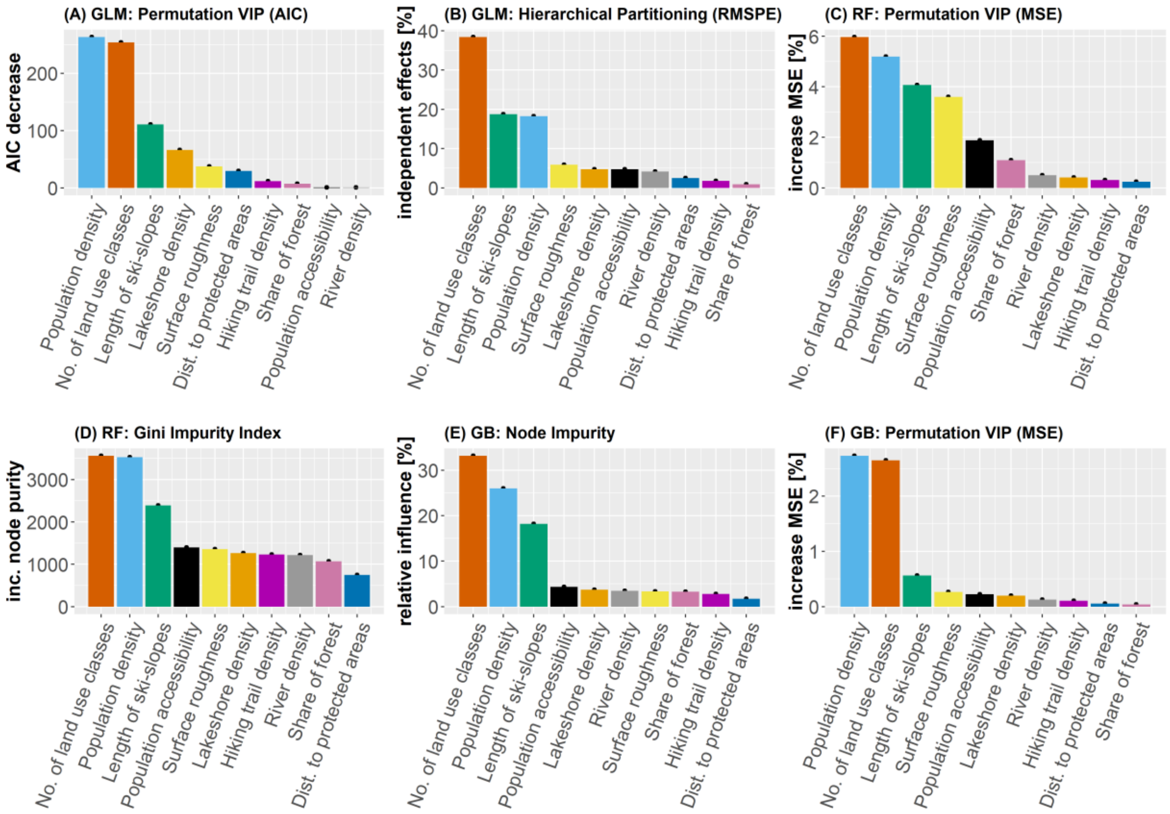

Figure 1. predictor variable importance from (A) GLM permutation VIP, (B) GLM hierarchical partitioning (C) Random Forest MSE increase, (D) Random Forest Gini Impurity, (E) Gradient Boosting Node Impurity, (F) Gradient Boosting permutation VIP.

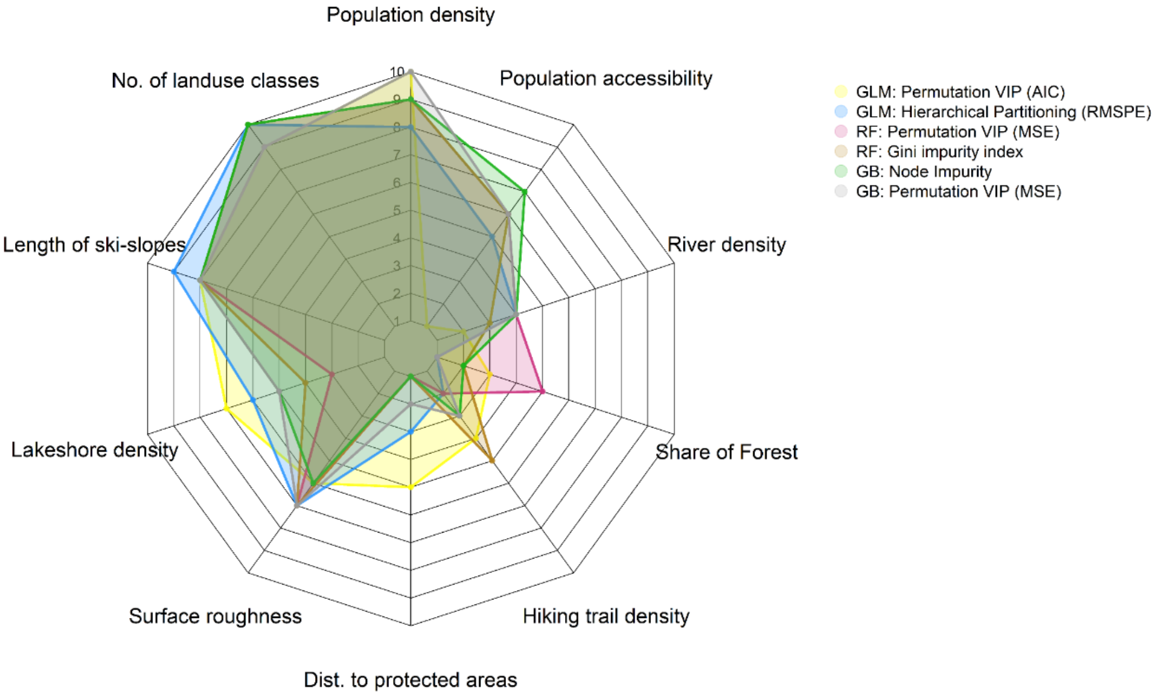

Figure 2. spider plot of predictor variable importance rankings (10: most important, 1: least important)

4. DISCUSSION AND CONCLUSION

We compared four different variable importance measures for three different regression models. Despite the quite large differences between the types of regression models and the differences between the types of variable importance measures, we found a striking similarity between the variable importance of the three most important variables. We also observe a clear dichotomy of our variables into the three most important variables population density, number of land use classes and length of ski slopes and the remaining variables of low importance, which are ranking quite differently between the different methods.

Despite the large consistency between the different methods, there does not exist a universally generic way of calculating variable importance. While permutation based variable importance is applicable over all different models it is also the most disputed method, because of its sensitivity towards multicollinearity of predictors (Murray and Conner, 2009). Nevertheless, also the other variable importance measures are widely disputed among scientists. Hierarchical partitioning that solves the problem of multicollinearity is criticized for not detecting the presence of spurious variables (Murray and Conner, 2009). The RF Gini impurity index and GB node impurity are criticized for favoring variables with many possible split points (Wright et al., 2017).

In spite of the mentioned drawbacks, for our case study, we can conclude that all applied methods provided relatively similar results and we cannot identify a single superior methodology. Therefore, if the goal of a research question is to identify the most important variables in a regression, it is advisable to apply different methods of variable importance to be able to draw robust conclusions.

Knowledge about drivers of outdoor recreation demand can help planning authorities develop more integrated transport, spatial and landscape planning decisions across scales. While municipalities are often dependent on revenues from outdoor recreation (Schägner et al., 2017), a thorough understanding of recreational demand can support long-term planning and management of these regions, avoiding negative environmental impacts (Schirpke et al., 2018). We can only provide results for the country of Switzerland, which is comparably small and has a very heterogeneous landscape. While we expect that our results are also valid for other alpine countries such as Austria, leisure behavior and mobility is very much dependent on the leisure time budget (Schlich et al., 2004) and social norms and contacts (Guidon et al., 2018). For that reason, more research is needed to indicate how generalizable our results are. For that purpose, it would be of interest to repeat and validate our study for other countries and regions.

REFERENCES

AGIMASS, F., LUNDHEDE, T., PANDURO, T. E. & JACOBSEN, J. B. 2018. The choice of forest site for recreation: A revealed preference analysis using spatial data. *Ecosystem Services,* 31**,** 445-454.

BREIMAN, L. 2001. Random Forests. *Machine Learning,* 45**,** 5-32.

BUCKLEY, R. 2004. *Environmental Impacts of Ecotourism,* Oxford, CABI Publishing.

CAMERON, A. C. & TRIVEDI, P. K. 2013. *Regression Analysis of Count Data,* Cambridge, Cambridge University Press.

CASPERSEN, O. H. & OLAFSSON, A. S. 2010. Recreational mapping and planning for enlargement of the green structure in greater Copenhagen. *Urban Forestry & Urban Greening,* 9**,** 101-112.

CHEVAN, A. & SUTHERLAND, M. 1991. Hierarchical Partitioning. *The American Statistician,* 45**,** 90-96.

CORDELL, H. K. 2012. Outdoor Recreation Trends and Futures - A Technical Document Supporting the Forest Service 2010 RPA Assessment. Asheville, NC.

COSTANZA, R., FISHER, B., ALI, S., BEER, C., BOND, L., BOUMANS, R., DANIGELIS, N. L., DICKINSON, J., ELLIOTT, C., FARLEY, J., GAYER, D. E., GLENN, L. M., HUDSPETH, T., MAHONEY, D., MCCAHILL, L., MCINTOSH, B., REED, B., RIZVI, S. A. T., RIZZO, D. M., SIMPATICO, T. & SNAPP, R. 2007. Quality of life: An approach integrating opportunities, human needs, and subjective well-being. *Ecological Economics,* 61**,** 267-276.

DRUCKMAN, A. & JACKSON, T. 2010. An exploration into the carbon footprint of UK households. *RESOLVE Working Paper Series.* Guildford: University of Surrey.

ELITH, J., LEATHWICK, J. R. & HASTIE, T. 2008. A working guide to boosted regression trees. *J Anim Ecol,* 77**,** 802-13.

FEDERAL STATISTICAL OFFICE 2012. Mobilität in der Schweiz: Ergebnisse des Mikrozensus Mobilität und Verkehr 2010. Neuchâtel and Bern.

FEDERAL STATISTICAL OFFICE. 2015. *Verkehrsverhalten der Bevölkerung, Kenngrössen - Schweiz* [Online]. Available: https://www.bfs.admin.ch/bfs/de/home/statistiken/mobilitaet-verkehr/personenverkehr/verkehrsverhalten.assetdetail.2004970.html [Accessed 23.05. 2017].

FEDERAL STATISTICAL OFFICE 2017. Verkehrsverhalten der Bevölkerung. Ergebnisse des Mikrozensus Mobilität und Verkehr 2015. Neuchâtel und Bern.

FRIEDMAN, J. H. & MEULMAN, J. J. 2003. Multiple additive regression trees with application in epidemiology. *Stat Med,* 22**,** 1365-81.

GRÖMPING, U. 2015. Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics,* 7**,** 137-152.

GUIDON, S., WICKI, M., BERNAUER, T. & AXHAUSEN, K. W. 2018. Explaining socially motivated travel with social network analysis: survey method and results from a study in Zurich, Switzerland. *Transportation Research Procedia,* 32**,** 99-109.

LEE, K. H. & SCHUETT, M. A. 2014. Exploring spatial variations in the relationships between residents' recreation demand and associated factors: A case study in Texas. *Applied Geography,* 53**,** 213-222.

MCCULLOUGH, B., BERGSGARD, N. A., COLLINS, A., MUHAR, A. & TYRVÄLNEN, L. 2018. The Impact of Sport and Outdoor Recreation (Friluftsliv) on the Natural Environment. MISTRA The Swedish Foundation for Strategic Environmental Research.

MURRAY, K. & CONNER, M. M. 2009. Methods to quantify variable importance: implications for the analysis of noisy ecological data. *Ecology,* 90**,** 348-355.

NEMBRINI, S., KONIG, I. R. & WRIGHT, M. N. 2018. The revival of the Gini importance? *Bioinformatics,* 34**,** 3711-3718.

ODE, Å. & FRY, G. 2006. A model for quantifying and predicting urban pressure on woodland. *Landscape and Urban Planning,* 77**,** 17-27.

SCHÄGNER, J. P., BRANDER, L., MAES, J., PARACCHINI, M. L. & HARTJE, V. 2016. Mapping recreational visits and values of European National Parks by combining statistical modelling and unit value transfer. *Journal for Nature Conservation,* 31**,** 71-84.

SCHÄGNER, J. P., MAES, J., BRANDER, L., PARACCHINI, M.-L., HARTJE, V. & DUBOIS, G. 2017. Monitoring recreation across European nature areas: A geo-database of visitor counts, a review of literature and a call for a visitor counting reporting standard. *Journal of Outdoor Recreation and Tourism,* 18**,** 44-55.

SCHIRPKE, U., MEISCH, C., MARSONER, T. & TAPPEINER, U. 2018. Revealing spatial and temporal patterns of outdoor recreation in the European Alps and their surroundings. *Ecosystem Services,* 31**,** 336-350.

SCHLICH, R., SCHÖNFELDER, S., HANSON, S. & AXHAUSEN, K. W. 2004. Structures of Leisure Travel: Temporal and Spatial Variability. *Transport Reviews,* 24**,** 219-237.

SEN, A., DARNELL, A., CROWE, A., BATEMAN, I., MUNDAY, P. & FODEN, J. 2011. Economic Assessment of the Recreational Value of Ecosystems in Great Britain. *Report to the Economics Team of the UK National Ecosystem Assessment.* CSERGE.

UNITED NATIONS 2017. 2017 Demographic Yearbook. *In:* DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS (ed.). New York: United Nations.

WEI, P., LU, Z. & SONG, J. 2015. Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety,* 142**,** 399-432.

WILLIBALD, F., VAN STRIEN, M. J. & GRÊT-REGAMEY, A. in review. Predicting outdoor recreation on a national scale - the case of Switzerland. *Applied Geography*.

WRIGHT, M. N., DANKOWSKI, T. & ZIEGLER, A. 2017. Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statistics in Medicine,* 36**,** 1272-1284.

ZEILEIS, A., KLEIBER, C. & JACKMAN, S. 2008. Regression Models for Count Data in R. *Journal of Statistical Software,* 27.