CAA 2009 Williamsburg

Application of data-mining technologies to predict Paleolithic site locations in the Zagros Mountains of Iran

Michael Märker[1] and Saman Heydari-Guran[2]

[1] Heidelberg Academy of Sciences and Humanities. Germany.
[2] Department of Early Prehistory and Quaternary Ecology, University of Tübingen, Tübingen 72070, Germany

Abstract

In this paper we present a concept for the storage, exchange, presentation and analysis of geospatial environmental and archaeological data to study and assess Paleolithic settlement dynamics and subsistence of the Zagros Mountains in Iran. Therefore, geographic information systems (GIS), database solutions and web-based technologies are used to handle and process archaeological and physiographic information. The project deals with a variety of variables and formats such as geology, geomorphology, land forms, and archaeology in vector and raster as well as text formats. The study presents a unique set of archaeological information sampled on the Zagros Mountains and its physiographic settings. To explain the spatial distribution of archaeological sites topographic indices were calculated and analyzed on a 90m resolution based on SRTM elevation data. These indices deliver i) information about erosion, transport and deposition processes of materials and sediments, ii) information about water availability and location as well as iii) information about site specific characteristics like aspect, radiation, elevation. We show that the distribution of the archaeological sites is strongly related to these parameters and proxies of ancient and current geomorphological, pedological and hydrological processes. Thus, this information was utilized to derive predictive models based on machine learning technologies. In this case we utilized a type of classification and regression tree, the random forest model. The application of the model on a test sample of the Dasht-e Rostam area showed the robustness of the analysis and the potential of the methodology

***Key words**: web based data base, predictive site modelling, Paleolithic of Zagros Mountains*

1 Introduction and Objectives

In the last decades spatial assessments to analyse the distribution and location of archaeological sites have been reported for a variety of different landscapes. These statistical modelling approaches use environmental variables like topography and its derivatives, geological settings, soil characteristics and hydrological features to predict the location of archaeological sites. The techniques well tested in other areas like ecology, pedology and geomorphology (e.g. Hengel & Reuter 2009, Mc Barathney 1992, Schröder 2008) were adopted within the archaeological context. Common for all these approaches is the hypothesis that the target variable, in this case the archaeological site, is strongly related to the environmental predictor variables (e.g. Kvamme 1988). However, in archaeology not only the spatial but also the temporal distribution of pattern plays an important role. Consequently one must define carefully the time frame the analysis is dedicated to. This time frame determines the paleo-environmental conditions that often are not known exactly. For example for certain time periods a lake level is reported for a specific region, whereas today this lake does not exist at all. Consequently, if we want to predict archaeological sites in the landscape we have to take into account that former conditions may have changed. Thus, remnants of these paleo-environmental conditions are very important to get predictions with good scoring results. Nevertheless, the present day landscape situation contains often conserved features especially if morphodynamics are low. Hence, proper analysis of the present day topography often reveals a lot of information useful in model generation. However, often problems arise due to a heterogeneous set of the target variable. In case of archaeological sites,

datasets are often assembled without a specific sampling design. Thus, a certain bias due to intrinsic and extrinsic factors occur: i) often only positive (present) cases are reported and not the negative (absent) ones ii) a spatial regular sampling design often does not exists, iii) often only surface founds are reported iv) buried material was not detected or v) only specific topographic areas were investigated.

In this paper we develop a model to predict the potential spatial distribution of archaeological sites for the Palaeolithic of the Zagros Mountains. Therefore, Geographic Information Systems (GIS), database solutions and web-based technologies are used to handle and process archaeological and physiographic information. Moreover a sophisticated classification and regression tree modelling approach based on Brainman's "bagging" and Ho's "random subspace method" was applied.

2 Study area

The study presents a unique set of archaeological information sampled in the Zagros Mountains. In this study we focus on the Dasht-e Rostam (Rostam Plain). Fig. 1 illustrates the location of the study area.

The Zagros Mountains stretch along the western and southern parts of Iran, forming a continuous range with numerous peaks over 3000 and 4000 meters above sea level. Structurally the Zagros Mountains consist of two parallel oriented geological zones: the highland and the folded zones (Heydari, 2007). The geology is essentially made up of sedimentary rocks. The main lithology consists of limestone.

Historically the Zagros Mountains are well-known to archaeologists because of their numerous caves and rockshelters associated with Paleolithic archaeological remains (Heydari-Guran, et al 2009).

The Dasht-e Rostam with elevations between 790-850 meters a.s.l. is located in the northwestern part of the Fars Province, and is bordered on all sides by limestone formations that rise up to 700 meters above the plain. The basin is divided into eastern and western plains, which today are under cultivation (Conard et al, 2007).The Dasht-e Rostam Region is characterized by synclinal and anticlinal structures and most of the folds are incised by tributary rivers (Oberlander, 1965, Heydari, 2007). The basin is drained by three major rivers; the Solak, Fahlian and Shiv Rivers, which flow into the Persian Gulf. More than 50 large springs are also originating in this region. The Dasht-e Rostam basin is a tectonically active zone, being located close to Kazerun-Qatar Fault, which continue to produce earthquakes and fractures (Roustaei et al, 2006).

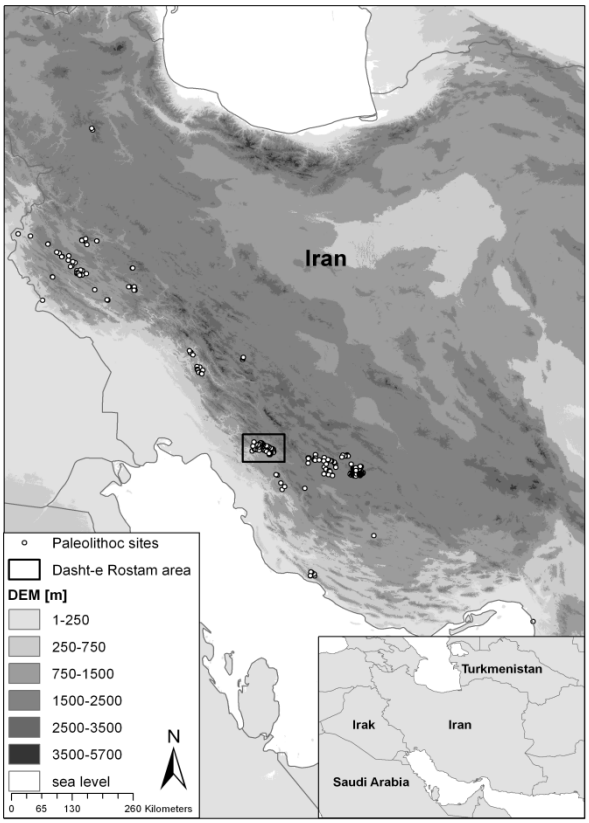

Figure 1. Study area showing Palaeolithic sites in the Zagors Mountains of Iran and the Dasht-e Rostam study area

The Zagros Mountains have a classical Mediterranean climate and can be divided into different regional climatic zones which are depending on altitude and longitude showing distinct vegetation types (Bobek, 1968).

From 2005 to 2007 researchers from the University of Tübingen and the Iranian Center for Archaeological Research (TISARP) discovered 139 sites in the Dasht-e Rostam area associated

with chipped stone tools on the surfaces (Conard et al, 2007). In this study we use the site information for a predictive archaeological site modelling, based on topographic functions.

Besides own archaeological field work we also collected available literature information on Paleolithic sites for the TISARP database. This database will cover the entire Zagros Mountain region. (Heydari-Guran in prep.).

2 Methods and Materials

The data, we utilized in this specific study, and the entire TISARP project, were imported and managed in a larger database developed for "The Role of culture in early expansions of humans (ROCEEH)-project". This database was implemented in PostgreSQL with a PostGIS extension and a UMN-mapserver application. It is an interdisciplinary georelational database with WEBGIS functions. The platform guarantees a simple exchange and assessment of the data with basic GIS functionalities. (see: www.roceeh.net; Märker et al. this volume)

Table 1: Derived Topographic Indices

Predictor variables	Method/Reference
Elevation	Preprocessed in ArcGIS 9.2
Altitude above channel network	(Olaya & Conrad, 2006)
Aspect	(Zevenberg & Thorn, 1987)
Catchment area	(Olaya & Conrad, 2006)
Channel network	(Olaya & Conrad, 2006)
Channel network base level	(Olaya & Conrad, 2006)
Convergence index	(Köthe & Lehmeier, 1996)
Curvature	(Zevenberg & Thorn, 1987)
Curvature classification	(Dikau, 1988)
Plan curvature	(Zevenberg & Thorn, 1987)
Profile curvature	(Zevenberg & Thorn, 1987)
LS-factor	(Olaya & Conrad, 2006)
Slope	(Zevenberg & Thorn, 1987)
Stream power	(Olaya & Conrad, 2006)
Watershed subbasins	(Olaya & Conrad, 2006)
Wetness index	(Olaya & Conrad, 2006)

The spatial analysis of the topographic relations and processes is based on Shuttle Radar Topography Mission (SRTM) data with a 90m resolution. The data was pre-processed, corrected and projected using ARCGIS 9.2 (ESRI 2009). The procedure encloses the elimination of no data values, the setting of the correct value range and the extent of the area. Moreover we applied a low pass filter to extract artefacts like local noise and terraces. The pre-processes DEM was imported into SAGA GIS 2.0.3 (Conrad 2006). Furthermore the DEM was corrected for hydrological applications using the methodology proposed by Planchon & Darboux (2001). Subsequently we performed a Terrain Analysis on the pre-processed DEM (Hengl & Reuter 2009). For this study we derived a set of 15 topographic indices. Tab. 1 shows these indices and the respective method applied for the delineation procedure. The delineated topographic indices deliver information about: i) erosion, transport and deposition processes of materials and sediments, ii) water availability and location as well as iii) site specific characteristics like aspect, radiation and elevation.

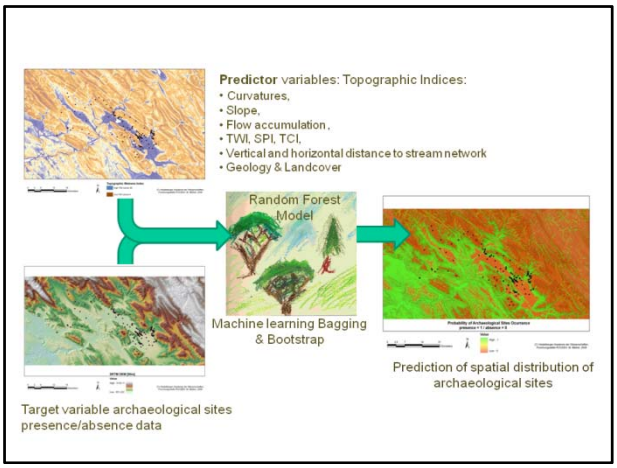

Figure 2. General modelling concept using Random Forest

In the following step we delineate a dataset containing all topographic indices and the information about the presence or absence of the archaeological sites. We obtained the information by archaeological fieldwork, screening the central area of the Dasht-e-Rosdam zone (Heydari-Guran in prep.). This site specific information was utilized to generate a spatial model that is subsequently applied to predict the potential distribution of archaeological sites of the entire region. Fig. 2 shows the general concept of the spatial modelling concept applied in this study.

The method utilized is based on a Random Forest approach (Breiman 1996, 2001). This method combines Breiman's "bagging" idea and Ho's "random subspace method" to construct a collection of decision trees with controlled variations (Ho 1995). The advantages of Random Forest are: i) that predictors are selected automatically, ii) that the method is more accurate than a single tree approach, iii) that data do not have to be rescaled or transformed, iv) that it is resistant to overtraining and v) that it provides an internal cross validation using "out of bag (OOB)" data (Breimann1996, 2001).

3 Results

The first analyses support our hypothesis that archaeological sites in the study area are related to topographic characteristics and processes. Especially the curvatures as well as transport capacities (LS-Factor) of runoff and the distribution of soil humidity (Topographic Wetness Index) show a high correlation with the archaeological sites. Fig. 3 illustrates the variable importance for the archaeological site modelling. In Fig. 4 the receiver operator curve (ROC) indicates a very high sensitivity by a low 1-specificity value. This means that the model generated with the Random Forest approach is highly robust. The calculated value for the ROC integral of 0,995 is therefore very high.

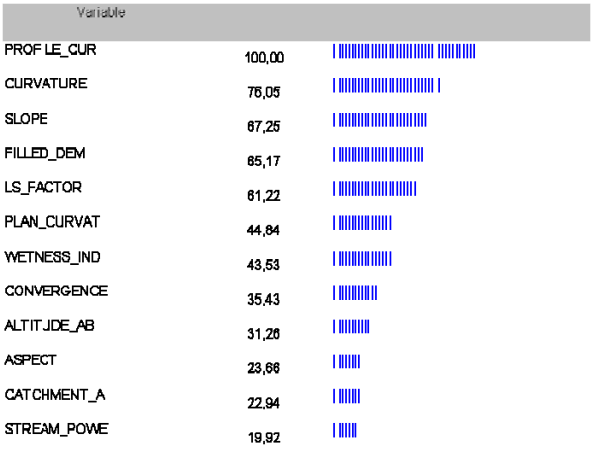

Figure 3. Importance of Predictor variables in cumulative percentages

The prediction success for the OOB sample, which is a kind of internal validation of the model, yielded values of 97% of correctly predicted "absence" of sites and 94% of correctly predicted "presence" of sites.

Fig 5 shows the regionalization for the entire study area. It is clearly visible, that especially the border of plains as well as the lower hillslopes close to river networks demonstrates high probabilities of occurrence. The physiographic settings of the arachaeological sites in terms of erosion, transport and deposition processes, micro climate, hydrological and edaphic conditions indicate areas with rectilinear or concave diverging landforms which means that the sites are well drained and thus, dry.

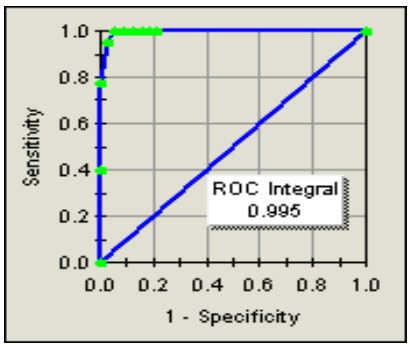

Figure 4. Receiver over operator curve

Moreover, the specific distribution of archaeological sites points towards a topographic boundary that might correlate with a former lake boundary. This is also stressed by a specific elevation range of the sites of 5 m to 85 m above the respective plain base level. The plain may have been flooded only seasonally. In the dry periods animals may have frequented the plain areas for browsing. This hypothesis is sustained by the fact that a concentration of Palaeolithic sites is located in a bottleneck situation between the two major plains. This indicates a specific suitable area for hunting when animals browse on the plain or pass from one plain to the other.

The regionalized model information (Fig. 5) shows in red colour areas with a high probability of occurrence of Palaeolithic sites, whereas the green areas are those where no Palaeolithic sites are predicted. In the South-western part of the study area a high density of sites is displayed. In these areas it is likely that the model can be improved taking into account other information like geological, landuse and/or spectral information form remotely sensed data. Nevertheless, it is obvious that Palaeolithic sites are distributed

around the Dasht-e Rosdam area following specific topographic characteristics.

4 Conclusions

We show that the distribution of the archaeological sites is strongly related to terrain characteristics and processes that can be assessed by topographic indices. Moreover we demonstrated that the Random Forest model yielded relations that can be regionalized to derive spatial archaeological site probabilities. The model already yields a lot of interesting information to develop hypothesis that can be tested in further model applications. Moreover, the results will help to define a sampling design for future field work.

In the next project step, we will assess the archaeological site data utilizing a DEM with a 25m resolution (ASTER).
Thus, we will enhance the DEM resolution to assess more detailed slope processes. Therefore the ROAD database is very helpful to store and exchange large datasets.

Acknowledgements

We would like to thank Andrew Kandel for reviewing our English manuscript.

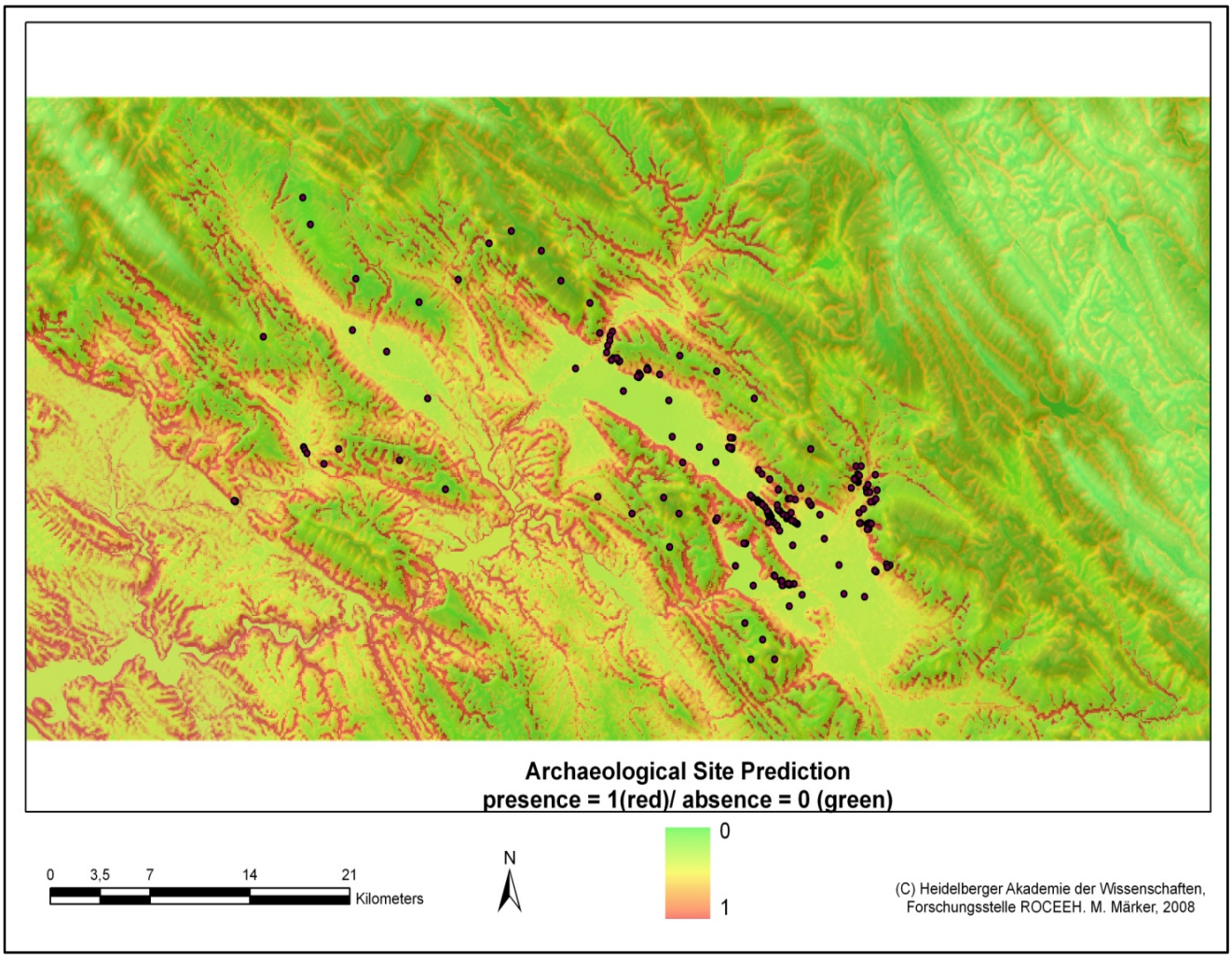

Figure 5. Predicted probabilities of Palaeolithic site occurrence. Red values indicate high probabilities of occurrence, green values indicate high probabilities of absence of Palaeolithic sites.

Bibliography

Bobek, H., 1968. Vegetation, In: *The Cambridge History of Iran* (Ed. Fisher, W.B.), Volume 1: The Land of Iran, Cambridge University Press, Cambridge:280-293.

Breiman, L., 1996. *Bagging predictors*. Machine Learning. 24: 123-140.

Breiman, L., 2001. *Random forests.* Machine Learning 45, 5-32.

Conard, N. J., E. Ghasidian, S. Heydari and Zeidee, M., 2006. *Report on the 2005 survey of the Tübingen-Iranian Stone Age Research Project in the provinces of Esfahan,* Fars and Kohgiluyeh-Boyerahmad. Archaeological Reports 5: 9-34.

Conrad, O., 2006. *SAGA. Entwurf, Funktionsumfang und Anwendung eines Systems für Automatisierte Geowissenschaftliche Analysen.* PhD Thesis , Unievrsity of Göttingen, Germany.

Dikau, R., 1988. *Entwurf einer geomorphographisch-analytischen Systematik von Reliefeinheiten.* Heidelberger Geographische Bausteine 5, 1-45.

Hengl, T. and Reuter, I.H., 2009. *Geomorphometry. Concepts, Software, Applications. Developments.* In: Soil Science 33, Amsterdam, Oxford, 765 pp.

Heydari, S.,2007. The *Impact of Geology and Geomorphology on Cave and Rockshelter Archaeological Site Formation, Preservation, and Distribution in the Zagros Mountains of Iran*. In: geoarchaeology: an International Journal, vol. 22, no.6.

Heydari-Guran, S., in prep. *Paleolithic landscapes of Iran: Landscape functional analysis for hunter-gatherer settlement patterns and modeling to predict habitat areas.* PhD thesis, University of Tübingen.

Heydari-Guran, S., Märker, M., and Conard, N.J., 2009. Geoarchaeological *predictive models for Paleolithic sites in the Zagros Mountains of Iran.* In: poster session of International metting Geoarchaeology in Central Europe. Dresden 31.04.-02-05.2009 .

Ho, T. K., 1995. *Random Decision Forest.* Proc. of the 3rd Int'l Conf. on Document Analysis and Recognition, Montreal, Canada, August 14-18, 1995, 278-282.

Kvamme, K.L. 1988. *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modeling*. In: Judge W. J. and Sebastian, L, eds, U.S. Government Printing Office, Washington, D.C., pp.325-428.

Köthe, R. & Lehmeier, F. 1993. *SAGA – Ein Programmsystem zur Automatischen Relief-Analyse.* Zeitschrift für Angewandte Geographie, 4/1993: 11-21.

Maerker, M., Hochschild, V. & Z. Kanaeva 2009. *Multidisciplinary Integrative Georelational Database for Spatio-Temporal Analysis of Expansion Dynamics of Early Humans.* CAA, Williamsburg, Virginia, USA. 18.-22.03.2009.

McBratney A.B., 1992. *On variation, uncertainty and informatics in environmental soil management.* Australian Journal of Soil Research 30, 913-935.

Olava, V. and Conrad, O., 2006. *Geomorphometry in SAGA.* In: Hengl, T. & Reuter, H.I. (Eds.). Geomorphometry: Concepts, Software, Applications.

Oberlander, T.M., 1965. *The Zagros stream. Syracuse Geographical Series 1.* Syracuse, NY: Syracuse University Press.

Planchon, O. and Darboux, F., 2001. *A fast, Simple and Versatile Algorithm to Fill the Depressions of Digital Elevation Models.* Catena, 46: 159-176.

Roustaei K., Alamdari K., and Petrie C.A., 2006. *Landscape and environment in the Mamasani District.* In: D.T. Potts and K. Roustaei (eds.), *The Mamasani archaeological project stage one: A report on the first two seasons of the ICAR-University of Sydney expedition to the Mamasani District, Fars Province, Iran,* pp. 17-30. Tehran: Iranian Center for Archaeological Research.

Schröder, B., 2008. *Challenges of species distribution modelling belowground*. Journal of Plant Nutrition and Soil Science 171(3): 325-337.

Zevenbergen, L.W. and Thorne, C.R., 1987. *Quantitative Analysis of Land Surface Topography.* Earth Surface Processes and Landforms 12, 47-56.