Session 1
Predictivity in Archaeology
Papers

1.1 Estimation of archeological potential with Page rank based predictive model: the MAPPA project results.

Francesca Anichini, Dario Bini, Nevio Dubbini, Fabio Fabiani, Gabriele Gattiglia, Francesco Ghizzani Marcìa, Francesca Grassini, M.Letizia Gualandi, Luca Parodi, Sergio Steffè
MAPPAproject - University of Pisa

One of the main research product of the MAPPA project (Methodologies Applied to Archaeological Potential Predictivity, www.mappaproject.org), involving the whole team formed by archaeologists, geologists and mathematicians, is a mathematical predictive model for estimating the archaeological potential in the areas with no available data, based on the data available from other areas. This was motivated by the extreme simplicity of models in literature, based mainly on map algebra or regression, which make those models not suitable for such a complex estimation. Moreover we considered the case of urban settlements, for which the literature is very poor. The case study was the urban area of Pisa, whose size, archaeological complexity and history, well represents a medium sized European city.

The archaeological potential represents the possibility that a more or less significant archaeological stratification is preserved. The following parameters were identified to properly define the archaeological potential: type of settlement, density of settlement, multi-layering of deposits, removable or non-removable nature of the archaeological deposit, degree of preservation of the deposit (Anichini *et alii* 2011). The identification of the relations that exist among finds is a key issue in the archaeological interpretation process, and this was also the key point in finding a suitable way to algorithmically determine the archaeological potential. In urban areas the spatial and the functional organization of the society, reflecting in the relations among finds, provides meaningful information for the automatic extraction of possible configurations of the parameters defining the potential. In other words, depending also on the archaeological period we are considering, it is possible to distinguish areas in which only some configurations of parameters that define the archaeological potential are feasible, or most probable. Such feasible or most probable configurations are given by relations among finds, that thus can *strengthen or weaken* the archaeological potential of the area itself.

We needed a number of datasets, in order to consider problem of the estimation of the archaological potential in all its aspects: archaeological data, building archaeological data, historical data, toponymic data, geomorphological data. The data model was developed to manage heterogeneous data, which draws the urban archaeological complexity. We worked with both topographical (e.g. geomorphologic, hydrographical, toponymic data, etc.) and urban data (e.g. archaeological stratifications, buildings, road network, hypotheses of historians and archaeologists, etc.). The archaeological data model combined raw data and interpreted data, and go from less synthetic data (i.e. the context level) to the more synthetic data. The key unit of the data model was the archaeological intervention, but the model included also: the filing of published data, of archive data and of data resulting from building archaeology, and data georeferencing and vectorisation in order to understand the urban fabric development and the level of architectural heritage preservation; the collection of written and published documentary sources with the aim to locate no longer existing place names, production activities, infrastructures and topographic structures; the computerised acquisition of historical mapping to trace urban transformation throughout the modern and contemporary ages.

After the creation of the data model the first step was the categorization of finds and assignment of the archeological potential value to every category. Each find was associated to a category with a proper level of generality, with two aims: to avoid the too particular information due to archaeological finds in order to obtain an efficient algorithmic procedure, and to allow for a spatial induction about archaeological potential. Then to every category was given a value of potential, which is computed on the basis of the types of archaeological information to which it is connected: those types of archaeological information were production, building techniques, trade, food, agriculture/breeding, worship, waste management, political/institutional aspects, social and gender aspects, physical anthropology, fauna/flora, geomorphology, viability/transport, health and hygiene, warfare, land management, leisure, tradition, water

system.
The next step was to express the relations between archaeological categories in the same chronological period and through different periods. We showed in (Bini *et alii* 2011; Bini *et alii* 2012) how a modification of the PageRank model can be used to assign archaeological potential, relying on the fact that the criteria used for attributing archaeological potential and the criteria used for assigning importance to web pages by search engines are both based on relations, and quite similar. A modified version of the PageRank model was applied to the heterogeneous datasets described above. The PageRank model needs as inputs a set of vertices (that play the role of web pages), and a set of weighted links among the vertices.
The vertices are obtained dividing the subsurface in a three-dimensional grid corresponding to the work area: the grid is composed of 7 layers of square cells with edges of 10 meters, one layer for each archaeological period under consideration. Every single cell plays the role of a web page, and has an initial value associated with this, due to the finds relative to that cell: the "importance" resulting from the application of the PageRank based model will be the archaeological potential of that cell. Indeed the archaeological potential of each cell should be more appropriately interpreted as the potential obtained when digging vertically from the surface down to that cell. For this reason, as the excavation goes deeper, the archaeological potential increases.
The set of weighted links, originating by cells with finds and "spreading" the potential, is represented by an $N \times N$ matrix H, where N is the total number of cells of the subsurface of the work area. The element H_{jk} of the matrix H is a number representing the part of importance that cell k transfers to cell j. The matrix H is obtained through the following steps:
- each cell containing a find is linked with the cells around it, within a certain distance, which is computed in the next step;
- a list of the possible functional areas, i.e. levels of spatial and functional organization (e.g. urban, suburban, rural areas) in which the urban space is organized, has been drew up. The distance of the previous step, giving the area of influence of each cell, is related to the functional area the cell is in. The principle behind this association is that the same find in an high-valued functional area (i.e. a "more important" one) has more probability of the presence of (more valuable) other finds in the surroundings;
- the weights of links of cells are not uniform inside the area of influence, but they are weighted by the geomorphological datum, since this datum constitutes a sort of "basic condition" for the development;
- the total amount of weights originated by each cell is related to an empirical estimation of the most probable finds next to the finds in the cell.
The application of PageRank based model consist in solving a linear, non-homogeneous and over-determined system, with more equations than unknowns. It must be treated, therefore, with least-squares techniques. The PageRank model gives as output the estimated potential value for each cell.
The map of archaeological potential is given so by the composition of the 7 layers, one for each archaeological period under consideration: Protohistory, Etruscan period, Roman period, Late Roman period, Early Medieval period, Late Medieval period, Modern Age, Contemporary Age. The final result has obtained after a validation of the results provided by a preliminary version, through 14 new cores, with which the algorithm was tested, in order to obtain a better fitting model. The results presented, including the archaeological potential map, are to be considered as the first steps towards an automatic, formally definable, and repeatable approach to the computation of archaeological potential. Of course no completely automated procedure would be possible in this and any task involving social and human behavior, so also in the proposed algorithm the procedure is controlled by the users (archaeologists), who can manage the whole process assigning values to every parameter. For these reasons, the map of archaeological potential should be always evaluated in conjunction with the interpreted archaeological data published in MappaGIS (www.mappaproject.org/webgis), and with the raw data released as open data in MOD (Mappa Open Data www.mappaproject.org/mod). In this way, the predictive map of archaeological potential is a useful and powerful tool both for land management and for archaeological research.

Bibliography

Anichini F., Bini M., Fabiani F., Gattiglia G., Giacomelli S., Gualandi M.L., Pappalardo M., Sarti G. 2011,*Definition of the parameters of the Archaeological Potential of an urban area*, in MapPapers 2en-I, pp.47-49

Bini D., Dubbini N., Steffè S. 2011,*Mathematical models for the determination of archaeological potential*, in MapPapers 4en-I, pp.77-85

Bini D., Dubbini N., Steffè S. 2012, *On the two main issues about the application of page rank for the determination of archaeological potential*, in MapPapers 2en-II, pp.45-50

Langville A. N., Meyer C. D. 2006, *Google page rank and beyond Princeton University Press*

1.2 Predicting settlement location through cost surfaces: a case study

Carlo Citter
Università di Siena . MediArG *

Antonia Arnoldus-Huyzendveld
Università di Siena, Digister srl**

Chiara Maccani
Università di Siena . MediArG*

Predictive modelling in archaeology is a long-lasting debated topic among scholars, with a consolidated set of tools (see Verhagen, Whitley 2011). This paper introduces a general procedure, which doesn't require high-level GIS or mathematical skills. It uses cumulative cost surfaces not only to calculate least cost paths, but also to predict settlement location and to evaluate resource exploitation. But the researcher needs to know the basics of landscape development in historical times, and the meaning of landscape features for practical purposes. In addition, one must be aware of the uncertainties created by the weighing procedures and by using proxy input data.

The first part of the procedure is deductive (see Deeben and others 2002, Canning 2005, Kamermans and others 2009). It starts with a critical evaluation of the reliability of the available environmental data. We produce several weighted cost surfaces related to moving or to taking advantage of the landscape in any other way, like in agricultural production and urban or farmstead settling. This means to create a series of raster surfaces to evaluate landscape features independently. They can be either attractors, detractors or "repellers", facilitators, and obstacles. The difference is that the first two act at a distance, the latter "under our feet". Examples of each type for moving are: springs, active volcanoes, level areas, steep slopes or main rivers to cross. Instead, a river to follow by boat or an existing road can be considered as "facilitators". Examples for agricultural land use are: presence or absence of water resources, good or stony soils. Examples for urban settlement are: nearness of good soils or forests, nearness of marshy areas, higher ridges in a plain, low-lying areas.

We produce the weighted cost surfaces by evaluating and classifying each factor for the use considered on a scale from 1 or 0 (high advantage) to 100 (no advantage, disadvantage). Often, we can combine the two opposed factors of a group in a single cost surface, for instance good and stony soils. Next, we combine all cost surfaces by weighing them against one another. The result is a map - or several maps - that express for each area or cell the degree of profitability for a specific category of settlement or use, for a certain period in a historical landscape with known characteristics. The result is not a statistical evaluation, but a qualitative estimation.

Often, we need to use proxy input data. For instance: when we have no soil maps, we consider flatlands synonymous of highly productive soils, which is not always the case, but indeed often. This affects the uncertainty of the results. Especially in coastal areas, where expansion of the shoreline occurred, we must consider the development of the landscape in historical times. The same is true for highly mobile alluvial areas, where river courses have shifted laterally in time.

Then, we turn to the archaeological features in general terms: how do known sites of a given category of a given period relate to the environment? Where could a route run in that landscape? Which were, and how were the natural resources of a given context exploited? After that, the procedure becomes inductive: which were the central places and their territories in that period? What do we know about the historical mobility- and transport network? In this stage we upload these data to the GIS platform.

We chose this method for three reasons. First, we will never know the exact number of archaeological settlements of a given period in a territory. Therefore, we cannot calibrate the procedure to match the sampled population. Second, predictive models rarely evaluate the environmental sustainability for the population. For instance, a dozen of single farmsteads found during a survey could well represent the maximum population which that region could sustain at the time. Therefore, a predictive model may be misleading, since it could suggest more sites, according to certain unrealistic parameters (unless goods were imported from outside the region). Thus, before evaluating, we should ask how many people could be sustained within a specific context. In addition, we should not weigh the parameters in the same way if we wish to predict the location of a tomb, a farmstead or a shepherd's hut.

Third, we think quantitative approaches to predictivity do not encourage archaeologists to use them in their daily workflow. A high knowledge of GIS technology and algorithms is assumed (Van Leusen and others 2009), a background most archaeologists do not have. Instead, working with cumulative weighted cost surfaces has a lower learning threshold, since it needs only the use of the raster calculator on a GIS platform. It is a flexible tool for risk management and archaeological research. We can increase or reduce the weight of a certain parameter according to the evaluation we are running. For instance, drainage can be more relevant than slope in evaluating the cultivable land. It can be the opposite when evaluating a potential route. It is a good practice to declare which parameters are introduced, their relative weights applied, and the overall procedure followed (see Citter, Arnoldus-Huyzendveld 2011). As said, this procedure implicates other risks given by the uncertainty in the use of proxy data and the weighing process. However, not even a quantitative approach avoids uncertainty, nor does the expert judgement.

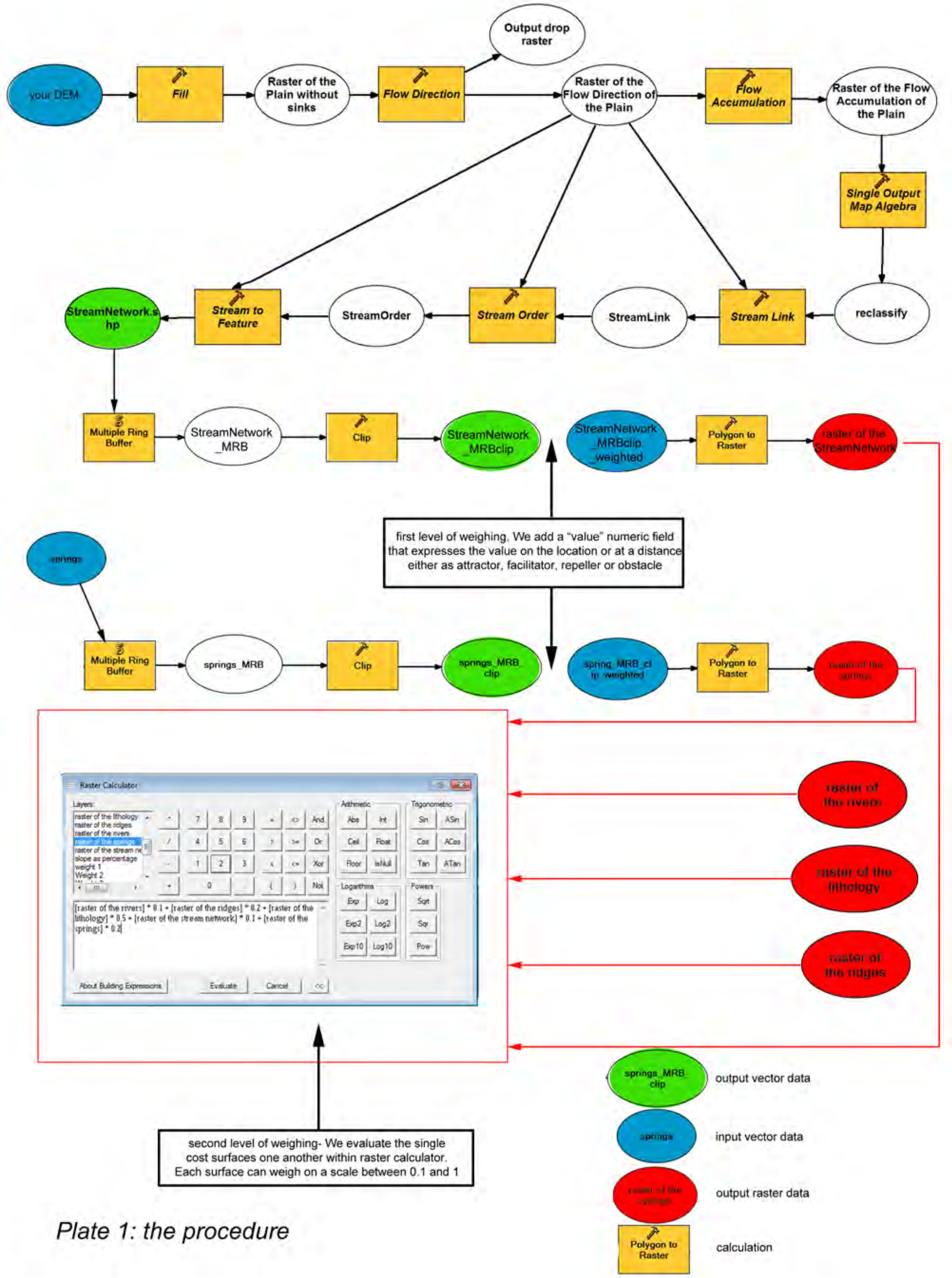
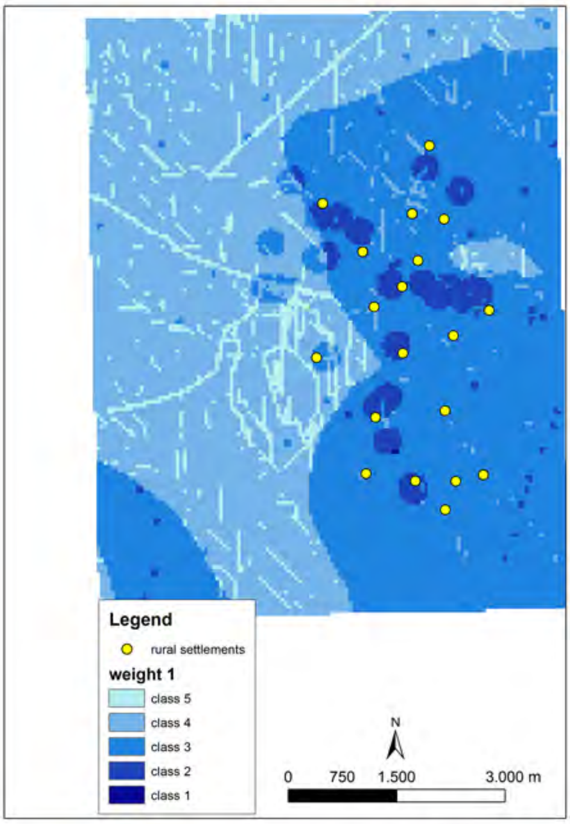
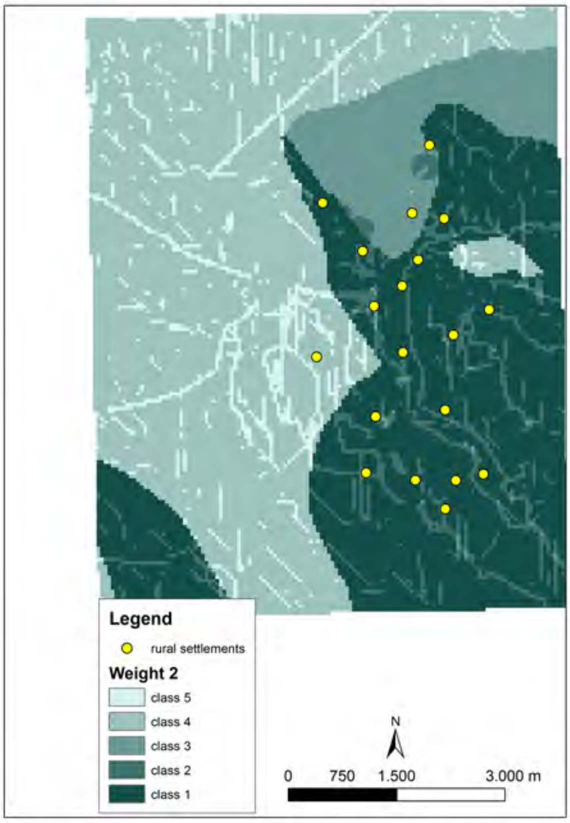


Plate 1: the procedure





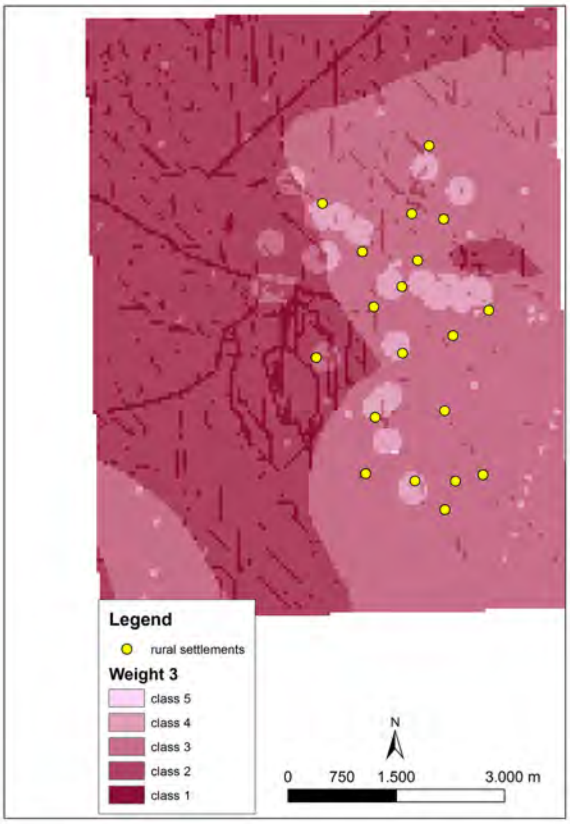


Plate 2: three different cumulative cost surfaces

Weight 1: [rivers] * 0.2 + [ridges] * 0.2 + [lithology] * 0.3 + [stream-network] * 0.1 + [springs] * 0.2

Weight 2: [rivers] * 0.1 + [ridges] * 0.1 + [lithology] * 0.5 + [stream-network] * 0.1 + [springs] * 0.2

Weight 3: [rivers] * 0.1 + [ridges] * 0.3 + [lithology] * 0.2 + [streamnetwork] * 0.1 + [springs] * 0.3

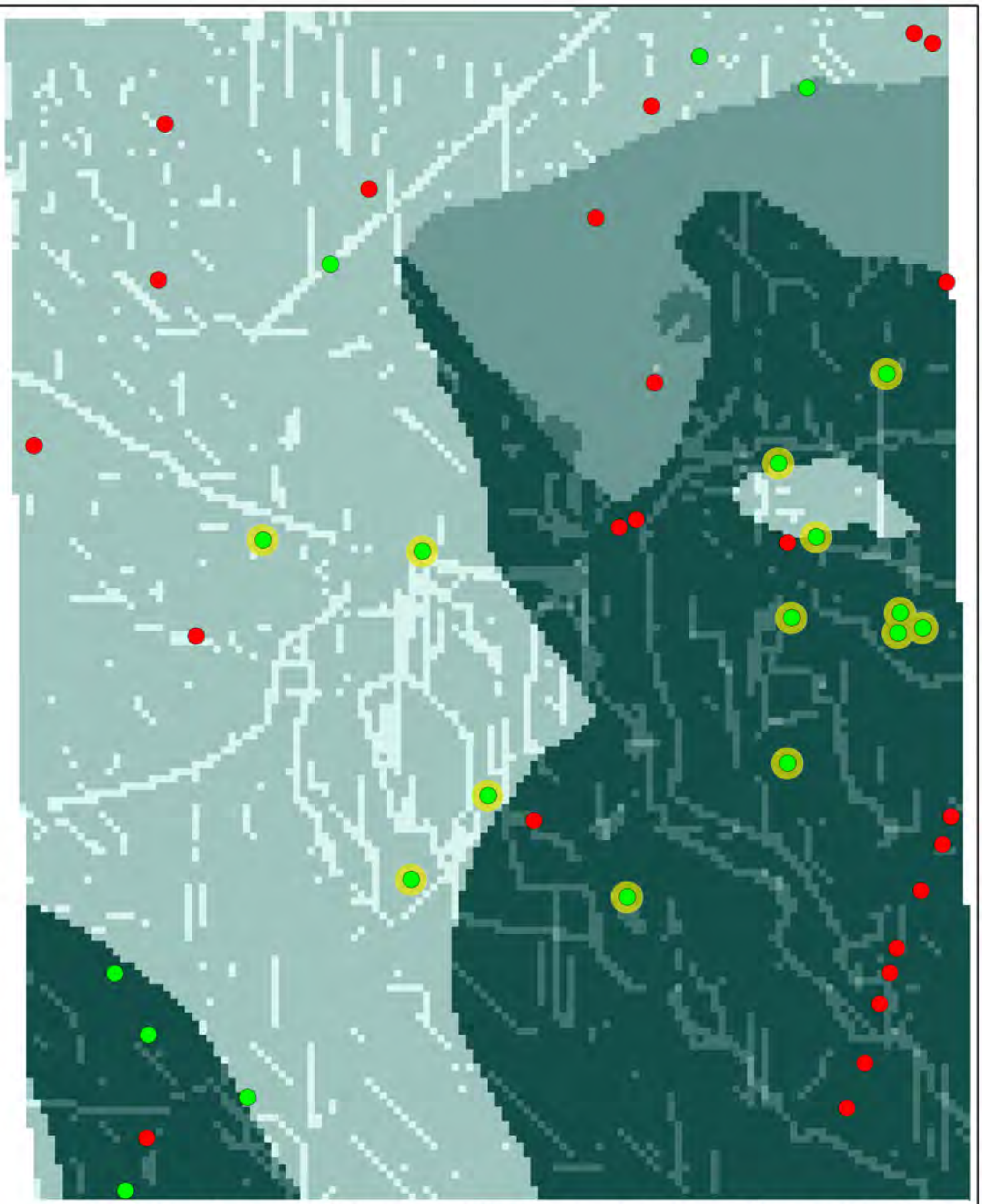


Plate 3: the ridges in relationship with the second cumulative cost surface.

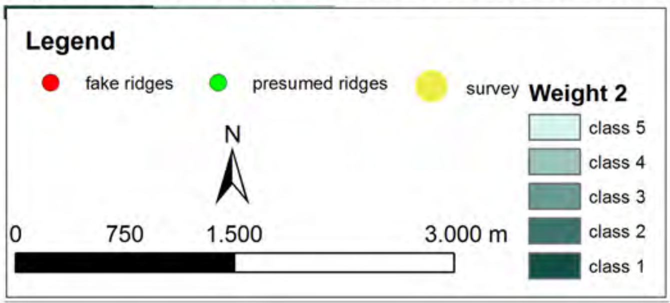


Plate 4: the survey on 2 of the ridges that produced evidence of a settlement.

We developed and tested the procedure for predicting resource exploitation and road networks in an area along the West coast of central Italy (Citter, Arnoldus-Huyzendveld 2011). Next, we applied it for predicting site location in a sample area of 6.5 x 8 km in the alluvial Po plain in northern Italy (the territory of Povegliano Veronese - 8 km SE of Verona). The area is located upon the alignment between the high and low glacial Alpine outwash plains, rich in springs. The steps of this procedure are detailed hereon (plate 1).
The available input data were: rivers, springs, lithology, soils, corine land cover, and a 10 m cell size DEM. We considered springs as attractors for settlement, fine textured and humid soils as facilitators, and gravelly soils and streams as obstacles. We tried out both the lithological and the soil map for productivity evaluation. We produced from the DEM a 50 cm interval contour map to identify the presence of small reliefs (at least 50 cm higher than the surroundings), which could have been suitable for settlement in an alluvial plain. These were considered as attractors. Next, we derived from it a potential stream network, i.e. where the water flows during the rainy season, since to settle on an (intermittent) small stream is likely less favourable.
The next step is to assign to each parameter a weight in relation to the distance from it or to the local value. Then, we turned the vector files into rasters, and added them in the raster calculator. We fixed the final amount to 1, thus weighing each raster as a fraction. This second weighing step allows to increase and decrease the value of each single parameter in the final cumulative cost surface. We can do it either for evaluating the site location, the route network, the resource exploitation, etc. In this case study, we produced three different combined cost surfaces to evaluate the most profitable areas for the location of rural settlements focused on crop production. After, we reclassified the cells in five qualitative classes (plate 2). Class 5 means no site is likely to be found, and class 1 means "it is very likely". This allowed to experiment with several weight combinations to check which results fitted best with the archaeological knowledge. One of the maps matched the 65% of the known settlements within the class 1 cells. We consider this result only an intermediate step. Also the a-posteriori confrontation with the corine land use map gave positive results.
Finally, we did spot surveys on the most and least potential areas to get more insight in the procedure's reliability. In particular, we chose to evaluate the 44 small ridges we derived from the DEM. We surveyed 20 of them (plate 3), being the others probably recent artificial mounds, as was suggested by satellite images. None of them lay on a class 5, 3 and 2 of the final map; 8 lay on class 4, and 12 on class 1. We could not verify 3 ridges on class 5 because they were cultivated at the moment of the survey, while the other 5 produced no evidence of settlement. Instead, 8 out of 12 on class 1 produced the most interesting data. In two cases the small heights were close to known archaeological sites; the others revealed 4 areas of potsherds spanning from the Bronze Age to the Early Middle Ages, while the other 2 returned generic Protohistoric and Roman material (plate 4). We could not investigate the remaining 4, because they lay on cultivated fields.
Although only standard GIS techniques are involved, we think this method deserves to be promoted in all its aspects. In particular, we stress the critical and open handling of the landscape data and of the weights applied in the cost surfaces. In addition, we think it is crucial the accompanying field survey. We plan to develop it further, especially for practical use by the students of archaeology, through the use of open source software like QGIS. Thus, we could overcome automatic handling of spatial data or, even worse, handling archaeological data as if they were not distributed in a real physical landscape.

*www.archeogr.unisi.it/CCGBA/laboratori/lam/

**www.digiter.it/geoarcheologia/

Bibliography

Canning S. 2005, *'Belief' in the Past: Dempster-Shafer Theory, GIS and Archaeological Predictive Modelling*, in «*Australian Archaeology*», 60, pp. 6-15.

Citter, Arnoldus-Huyzendveld 2011, *Uso del suolo e sfruttamento delle risorse nella pianura grossetana nel medioevo verso una storia del parcellario e del paesaggio agrario*, Confronti, 1, Roma.

Kamermans H., Van Leusen M., Verhagen P., (eds.) 2009, *Archaeological Prediction and Risk Management*, Leiden.

LeusenM. Van, Millard A.R., Dicke B. 2009, *Dealing With Uncertainty in Archaeological Prediction*, in Kamermans H. and others eds.: pp. 123-160.

Verhagen P., Witley T. G. 2011, *Integrating Archaeological Theory and Predictive Modeling: a Live Report from the Scene*, in «Journal of Archaeological Method and Theory», february 2011, no pp.

1.3 Archaeological Predictive modelling: a proposal for the CRM of the Veneto region.

Anita Casarotto, Hans Kamermans

University of Leiden

The title of this contribution is taken from my (AC) Specialization thesis which exposes the work plan and the outcomes of a five-month research project. Such a project was implemented through an Erasmus agreement between University of Padova and Leiden University and concerned with the methodological study of predictive modelling. The thesis encompasses study history, epistemological issues, limits and successful aspects of predictive modelling in both CRM and research environments, and a comparison between the Dutch practice with examples from others European countries. It aims at coming up with a proposal for the CRM of the Veneto region by especially analysing what has been already conducted for AHM-oriented predictive modelling in the Netherlands and referring to it as the main applicative instance throughout Europe. For the Veneto Region, currently engaged in updating the P.T.R.C. (Piano Territoriale Regionale di Coordinamento), this methodology may be helpful to improve the monitoring of the archaeological resources in the territory and to assess the archaeological risk involved. The practical target of our proposal will be the implementation of a supposed working model to be adopted by the regional CRM authority, that is presently addressing the predictive/preventive issue as the top priority of its agenda. A predictive model has been developed for the case-study of eastern Lessini area, in the provinces of Verona and Vicenza (Casarotto, De Guio, Ferrarese, Leonardi 2011). Such a model could be revised, improved and afterwards used as a test-area for the Veneto region-wide target. *We need to predict the past in order to have a role in spatial planning* (Kamermans 2011: 15), as a matter of fact predictive modelling would be a valuable tool in CRM for assessing the archaeological potential of a region, and it allows policy makers to more consciously scale the protective actions as to the territory. A predictive model will be always a subjective interpretation of cultural processes occurred in the past, but differently from others approaches it uses objective operators during the analysis, indeed it exploits mathematical algorithms and statistical methods for producing probability maps. For this reason predictive modelling could become a shared platform for the standardised and controlled representation of the archaeological potential in a Region or, even better, in an entire country.

Nevertheless we have come to the conclusion that