A PageRank based predictive model for the estimation of the archaeological potential of an urban area

Nevio Dubbini, Gabriele Gattiglia

***Abstract*—We present the analysis of multi-faceted, GIS managed data for determining the archaeological potential, i.e. a measure of the possibility that a more or less significant archaeological stratification is preserved. We used a sizable number of datasets, in order to consider the problem of estimation of archaeological potential in all of its aspects: archaeological data, building archaeological data, historical data, toponymic data, geomorphological data. As the identification of relations among finds is a key issue for the data mining in archaeological interpretation process, we applied a modified version of the PageRank model, because the criteria for assigning importance to web pages by search engines are similar and based on relations, also. The procedure included a categorization archaeological data, the assignment of initial values of potential to the available data through an automatic procedure, the creation of geomorphological *facies* maps, the definition of functional areas (i.e. the levels of spatial and functional organization: urban, suburban and rural areas), and the application of the PageRank based algorithm. The model has been applied on the urban area of Pisa, and tested through the data of 14 new cores. The map of archaeological potential consists of 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 results, 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.**

Keywords*—predictive modelling, archaeological potential, PageRank, archaeological GIS, geomorphology.

I. Introduction

This paper studies the problem of computation of archaeological potential, the assumptions made to solve it, the mathematical model used, the software implementation, and the test of the algorithm in the case study of the urban area of Pisa. We based the mathematical model on PageRank, because there is an analogy between the criteria used for attributing archaeological potential and the criteria used for assigning importance to web pages in search engine algorithms. The key issue of the computation of archaeological potential, from an abstract viewpoint, is the identification of relations among finds: the presence of a particular find near another could strengthen or weaken the probability that they will form a more complex structure, and so strengthen or weaken the archaeological potential of the area. This is exactly the criterion upon which page ranking algorithms are based, whereby each web page attributes importance to the web pages it points to (via a link) and receives importance from the web pages it receives a link from. The reader can refer to [4] for further explanations about the choice of the mathematical model, and to [8] for a general mathematical introduction to PageRank models. In the following we will consider all the archaeological data as categorised, having assigned each find to a category in order to characterize its salient features, to effectively implement the algorithm, and to make the results general enough to be applied also in different contexts (pp. 89-99, [2]).

N. Dubbini is with University of Pisa, Mathematics department, Pisa, Italy, nevio.dubbini@gmail.com.

G. Gattiglia is with University of Pisa, Archaeology Department, Pisa, Italy.

II. Definition of Archaeological Potential

The archaeological potential of an area represents the probability that a more or less significant archaeological stratification is preserved. It is computed by analysing a series of historical, archaeological and paleo-environmental data, with a degree of approximation that may vary according to the quantity and quality of the data provided. The archaeological potential of an area is independent of any other following intervention that is carried out, which must be regarded as a contingent risk factor. The process of defining overall urban archaeological potential consists in drawing up a series of predictive maps relative to historical periods. The general criterion was to reconstruct stratigraphic intervals, and integrate this information with both archaeological and geomorphological data: geological maps define stratigraphic units and sedimentary bodies, geomorphological and paleogeographical maps show relief forms and define the geomorphic processes responsible for their genesis, in addition to recent modifications. Generally speaking, each morphological unit (or morphotype) can be more or less suitable for settlements. Subsequently the diachronic evolution of the forms has been characterised. In archaeological terms, the following parameters were taken into consideration for the predictive definition of the city throughout its historical periods: typology of finds, inferred on the basis of the interpretation of the archaeological records [7]; quality and quantity of the archaeographic data; spatial and typological relations among the finds, which allow identification in probabilistic terms of the presence of further finds in areas that have not been archaeologically investigated; expert judgment; land use, including traces that are not strictly connected to constructions or settlements, such as agricultural and/or farming practices; historical data from written sources and maps. Finally, we indentified the following overall parameters that best determine urban archaeological potential: type of settlement, i.e. the presence of settlement structures and their different typology; density of settlement; multi-layering of deposits; removable or non-removable nature of the archaeological deposit; degree of preservation of the deposit, calculated according to the presence of anthropic and natural removals [1].

978-1-4799-3170-5/13/$31.00 ©2013 IEEE

For managing all the heterogeneous data which draw the urban archaeological complexity, we realised a data model [3] capable of working with both topographical (geomorphologic, hydrographical, toponymic data, etc.) and urban data (archaeological stratifications, buildings, road network, hypotheses of historians and archaeologists, etc.), combining inter-site analysis and archaeological excavation GIS resources.

III. The problem: estimation of the archaeological potential

The algorithm we tried to set up was intended not only to provide an estimation of the overall (i.e. summing up all archaeological periods) archaeological potential, but also to give a period specific estimation. Therefore the subsurface under study is divided in 3-dimensional cells forming layers covering each one the work area. The number of layers is equal to the number of archaeological periods we used, because in this way a natural distinction between the relations going inside the same archaeological period and the relations going through periods can be achieved. The 7 periods the archaeological team has defined are: Protohistory, Etruscan period, Roman period, Late Roman period, Early Medieval Age, Late Medieval Age, Modern Age, Contemporary Age. Hence, the subsurface of the work area has been divided into $n = n1 \times n2$ cells for each one of the 7 layers.

First, we assigned to each of the $7n$ cells an initial value of the archaeological potential on the basis of the available data. The archaeological potential represents the possibility that a more or less significant archaeological stratification is preserved. For the purposes of this paper, we distinguish between absolute archaeological potential, and estimated archaeological potential. The *absolute potential* is defined to give a potential value to cells on the basis of available data, while the *estimated archaeological potential* is the estimate of the informative value of archaeological stratification (given by the algorithm), and is divided into 5 levels, from the level 1 (almost no importance), to the level 5 (very important stratification). The value of absolute potential of finds has been obtained by summing two different components:

The type of archaeological information. To compute the value of this parameter a list of 19 areas of interest has been drawn up, corresponding to the main informative fields on which the finds can provide information. The identified areas of interest are: 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. Finds were so assigned a value of absolute potential by summing, for each area of interest, the value 1 if the category provided information on that informative field, and 0 otherwise. For example the category 'domus' was given the value 12 since a domus can provide information about the 12 areas of interest: production, building techniques, trade, food, worship, waste management, social and gender aspects, health and hygiene, land management, leisure, tradition, water system;

The removable or non-removable nature of archaeological finds. This parameter is connected to the consistent/transient nature of each archaeological trace. It is assumed that more consistent archaeological remains (e.g. a stone building) corresponds to a higher absolute potential value, since there is a greater chance of identifying the remains themselves. The values assigned to this parameter are 1, in the case of archaeological findings achieved in masonry (e.g. 'insula', 'tower house', 'domus', 'palace', 'theater', 'church', 'prison', 'forum', 'church'); 0.5, in the case of archaeological remains which provide a masonry but which may be less consistence (e.g. an adobe building) (e.g. 'river bank', 'enclosure', 'henhouse', 'pigsty', 'stable', 'hearth'); 0, in the case of finds with a transient nature such as a wooden hut (e.g.,' canal', 'reclamation', 'trench', 'agricultural land', 'clearing', 'camp', 'waste dump').

The values of absolute potential were assigned to cells on the basis of the available data, and, depending on the features of data, different input data were created, for each archaeological period:

- *Certain geolocation data* were defined as the data with known spatial 2-dimensional coordinates (dating can be uncertain). So, under this class there are data from excavations with known spatial 2-dimensional coordinates, data from aerial photography anomalies, etc. Certain geolocation data were organized in a $n1 \times n2$ matrix, whose element i, j represents the absolute potential of that cell given by the certain geolocation data inside the cell;

- *Uncertain geolocation data* were defined as the data with unknown spatial coordinates, i.e. data for which we only knew that they are located in a certain region of the space. So, under this class there are data from excavations with uncertain spatial 2-dimensional coordinates, data from medieval written sources, etc. Uncertain geolocation data were organized in a $n1 \times n2$ matrix, whose element i, j represents the absolute potential of that cell given by the uncertain geolocation data inside the cell.

- *Shapes*: from certain geolocation data, the archaeological team has tried to give a shape to the find, e.g. to outline a house from a wall or a floor, or the continuation of a street from a piece of it, and so on. Each shape was given an empirical precision value varying from 1 to 6 (1 is the value of maximum precision), expressing how much the size and the orientation of the shape can be deduced with precision, on the basis of the finds in the nearby and on the geomorphological datum, for instance. The values of absolute potential of shapes is computed dividing the absolute potential of the category of the find the shape belongs to, by the precision value of the shape. Shapes data were organized in a $n1 \times n2$ matrix, whose element i, j represents the absolute potential of that cell given by the shapes;

- *Geomorphological data*: these data were deduced from geological surveys, in order to identify, for each archaeological period, the diverse geomorphological *facies*, which were distinguished in river, floodplain, wetland, marshy area, elevation. Also the geomorphological datum was given an absolute potential value, by summing the absolute potential of all the categories of finds that can be present in a *facies*.

Where this datum is not available it was replaced with the mean of geomorphological *facies* values.

In addition to these inputs, the algorithm makes use of *functional areas*, i.e. levels of spatial and functional organization in which the urban space is organized, and of their values of absolute potential. Each urban centre, in each archaeological period, is surrounded by a suburban area, and, more externally, by a rural area. The identification of the functional areas is based on many elements, since it depends on the different settlement types, the relationships among them and the environmental context. To limit the subjectivity in defining the functional areas we used an automatic procedure, described in details in [2] (pp. 89-99).

Finally, another two input files have been created, expressing synchronic and diachronic associations among finds. Those associations were used in the algorithm to enhance the spread of archaeological potential on the basis of the probable presence of further finds in the surroundings.

- As for the *synchronic associations*, for all categories of finds, depending on the functional areas in which they are located, each category was associated to the most probable categories founded at a distance < 50 metres. With the expression "most probable" we mean at least in the 75% of the attestations. This input files, for each cell, contains the sum of the potential values of the categories associated to the finds located in that cell;
- As for the *diachronic associations*, for all categories of finds, depending on the functional areas in which they are located, each category was associated to the most probable categories founded at a distance < 50 metres (in the 2-dimensional spatial coordinates), in the chronologically previous or following archaeological period. This input files, for each cell, contains the sum of the potential values of the categories associated to the find located in the cell.

The size of (square) cells was chosen to be $10 \times 10\ m^2$: this size was the outcome of different factors taken into account. On one hand, the archaeological data could be located with precision, but the geomorphological data could be given a precision no more than $10 \times 10\ m^2$, due to the number of elevation points available for the creation of each historical DEM; on the other hand there is a trade-off between the size of the cells and the total number of cells covering the work area, so that the smaller is the size of the cells, the higher is the total number of cells covering the area. Therefore the smaller is the size of the cells, the more are the cells for which the archaeological potential has to be estimated. For example, at the limit case for which we would like to estimate the archaeological potential of each point of the work area with arbitrary precision, we would have a finite number of input data, but we should estimate the potential of an infinite number of points. To solve this problem we searched (numerically) the maximum of a function representing the difference between the amount of information given by the only presence of available data, and the area of cells divided by the total number of cells. The amount of information given by the presence of available data is computed as the Shannon entropy of a binary matrix where each cell containing input data is given the value 1, and the other cells are given a 0 value. The resulting formula is

$$\max_{l \in \mathbb{N}} \quad w \cdot [-p_1 \log(p_1) - p_0 \log(p_0)] - (1-w) \cdot (l^2/n),$$

where p_1 and p_0 are defined as the relative frequency of cells with data and with no data, respectively, in the work area, l is the edge of the cells, and w is a parameter. The maximization, for different values of w near $1/2$, yield the "optimal" size of cells between 10 and 14 metres, so we chose the size of cells to be 10 metres, also for practical reasons: since the first tries, data were given for 10 metres, and so no other smoothing is needed to adapt data for cells of other sizes. In order to perform computations on values of archaeological potential "per unity of area", the value of absolute potentials of cells was divided by its area, so making computations as independent as possible of the cell size.

The problem is that of estimating the archaeological potential in every cell.

IV. THE MATHEMATICAL MODEL

Some mathematical models have already been applied to the prediction of the archaeological potential, basically divided into two groups: models based on map algebra, which are easy to implement, but provide only on/off (e.g. presence/absence) results; models based on the application of linear (or logistic) regression, that has the benefit of using variables to predict further variables, but does not take into account the great complexity that must be considered when determining archaeological potential (This is so true that current models based on regression are often not preferred to those based on map algebra).

We applied a PageRank based model to the input data described in the previous section. The model needed the matrix of weights and the vector of absolute archaeological potential, representing what is known by available data. A general introduction to the PageRank model can be found in [8], while the application of PageRank based techniques to the estimation of the archaeological potential can be found in [4]. It is important to note that the algorithm was applied to each period separately, so that we have a potential map for each archaeological period, which we can "sum" as a last step. Before describing how the model works, we show how the vector D representing available data, and the matrix of weights S, are constructed. The vector D is a matrix of dimensions $n1 \times n2$, reshaped to a column vector of length n, and is obtained by the sum of the matrices of absolute potential of certain data, uncertain data and shapes. The matrix of weights S is a $n \times n$ matrix whose element i, j represents the weight (value) of the link between the cell j and the cell i. The matrix S is computed in the following way:

- Each cell with a find distributes its importance to a square mask of cells centered in the cell itself. The edge of this square is set to be proportional to the absolute potential value of the functional area the cell belongs to. We used the values of functional areas in this way since their absolute potentials

were assigned on the basis of the finds that can be present in the functional areas: therefore, in higher valued functional areas it is justified that the region where to spread the potential is larger, and viceversa;

- When a mask as in the previous step is constructed, the total weight (i.e. the sum of weights distributed by a cell) inside the mask is given by the value of the functional area plus a value proportional to the sum of synchronic and diachronic associations. We used this value to address the total weight because the "quantity of potential" spread by a cell is influenced as well by the probability of finding high or low valued finds in the nearby;

- The distribution of weights in the mask around cells is given by the uniform distribution weighted by the geomorphological values of the mask around the cell, and weighted by the functional areas values. This is because the geomorphological datum constitutes a basic influence on the spread or archaeological potential, and the functional areas values - since they are proportional to the total potential of the finds you can find in the area - can be used to weight the diffusion of potential.

The PageRank based model uses the data described above, to estimate the archaeological potential. The algorithm consists of a basic procedure, applied repeatedly.

Procedure 1:

1) The vector D representing available data and the matrix of weights S are generated as described above;
2) The following iterations are performed

for $i = 1, \ldots, 1000$

$$A = S \cdot x + \begin{bmatrix} 1/n & \cdots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \cdots & 1/n \end{bmatrix} \cdot x$$

$$A = (1 - yield) \cdot A + yield \cdot x$$

$$u = [1, 1, \ldots, 1]$$

$$y = rel \cdot A + (1 - rel) \cdot [D \cdot (uT)]$$

$$y = \frac{y}{\sum_{i=1}^{n} y_i}$$

$$x = y$$

end

$$D = speed\_up \cdot x + (1 - speed\_up) \cdot D.$$

$\diamondsuit$

In these formulas x is a stochastic (i.e. the sum of its components equals 1) random column vector of dimension n, used as an initial condition for the application of the iteration described in the "for" cycle (the result of these iterations are independent of the initial condition). The following are the tunable parameters used in the algorithm:

- $maxit \in \mathbb{N}$ is the number of times the Procedure 1 is executed. Each time we applied it the algorithm makes a step in the prediction of archaeological potential, and after each step the result is taken as the new starting point for the next step. So, the greater is $maxit$, the more the prediction "turns from" the original data;

- $speed\_up \in [0,1]$ is the weight expressing the part of the new potential due to the results of the application of Procedure 1, and the part due to the absolute potential of the previous step. So, the more $speed\_up$ approaches 1, the less the new computation is due to the data from the previous step;

- $rel \in [0,1]$ is the parameter ruling how much the potential given by the weight matrix S (relations) is taken into account, with respect to how much the potential given by the absolute values is taken into account. So, the nearer rel is to 1, the less the absolute values of potential are taken in consideration, and the more the matrix of weights S (i.e. the relations) is preeminent in determining the archaeological potential;

- $yield \in [0,1]$ is the amount of potential each cell keeps for itself, with respect to the rest, which is distributed on the basis of the weight matrix S. So the nearer yield is to 1, the more each cell keeps potential for itself.

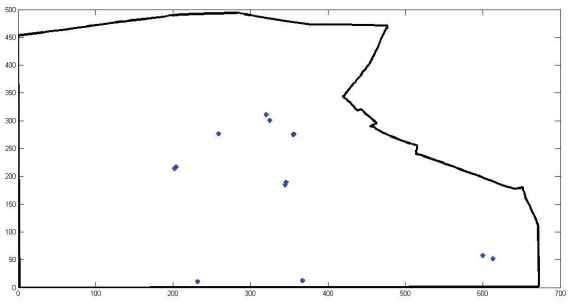

Figure 1. New cores locations. In Gauss Boaga coordinates
V1 (1615949.323, 4839553,933);
V2 (1616075.662, 4839500.355);
V3 (1613622.696, 4839094.774);
V4 (1612270.018 4839084.766);
V5 (1613507.696 4841737.042);
V6 (1613500.508 4841726.018);
V7 (1613149.328 4842085.415);
V8 (1613193.465 4841997.189);
V9 (1613194.700 4841990.480);
V10 (1613407.183 4840864.885);
V11 (1613400.304 4840824.139);
V12 (1611973.376 4841113.678);
V13 (1611986.606 4841150.059);
V14 (1612541.350 4841744.557).

V. THE TESTING

The model has been applied on the urban area of Pisa, and has been obtained after the tuning of the parameters of the algorithm, on the basis of the data of 14 new cores, by which the algorithm was tested, in order to optimize the fit of the proposed model ([6]). The new cores were executed in the places indicated by the Figure 1.

The data resulting from the cores, depending on the nature itself of the method of investigation, couldn't be included in

a specific archaeological period, apart from some exceptions, which are more precisely described in [2]. Therefore the validation of the results provided by the algorithm was performed on the overall archaeological potential. Even if some finds could be dated back to their own archaeological period, they cannot be used for a quantitative test for single archaeological periods, since the greatest part of the data is uncertain and it could be completely misleading for a quantitative approach.

Another important issue about the testing method concerns the way of weighting the different archaeological periods in the computation of the overall potential. This problem arises because the potential of each archaeological period is given the sum equal to 1, for technical reasons, so the total archaeological potential is the same for every period. This means that those archaeological periods with less data available get higher values of potential, since the potential "concentrates" in a smaller number of cells and the sum is 1, at the same time. To (partially) compensate this problem, we multiplied the matrix of each archaeological period for a weight equal to the maximum absolute value of that period divided by the maximum of archaeological potential for that period, so that the archaeological potential of each period has a range of values going from zero to the values of its maximum absolute potential. Having done this, we can sum the values of estimated archaeological potential in the same way as we sum the values of absolute potential in the new cores. In details, the test was performed as follows:

- The matrix of the overall absolute potential was computed summing the absolute potentials of each period, with no weighting. For the overall absolute potential it was the only possibility, since the sum of absolute potential of available data has to be comparable with the sum of absolute potential of new cores, and it is computed with no weights;
- The overall estimated archaeological potential was computed with a weighted sum: the potential of each archaeological period was divided for its maximum and multiplied for the maximum absolute potential of that period. In this way the archaeological potential of each period has a range of values going from zero to the values of its maximum absolute potential, and the sum is comparable with the one of absolute potentials in the new cores;
- The absolute potential of cells where the new cores took place was computed by adding to the absolute potential of those cells the additional information given by the cores. Due to the uncertain dating of many of the finds, the cores have a minimum and a maximum value of absolute potential: indeed, an uncertainly dated find across three archaeological periods, for example, could have been present in one, two, or three periods, so that the overall absolute potential varies consequently. We considered the mean of these two values;
- The overall estimated archaeological potential was divided into five different levels, going from the minimum value 1 (almost no archaeological importance), to the maximum value 5 (very important stratification), according to the following criteria: the first level goes from 0 to the minimum value of the estimated archaeological potential, and the remaining interval is divided into 4 equal parts. We chose this particular division to distinguish the cells where the estimation of the archaeological potential equals the minimum value (the "zero" of the estimation), from the others;
- The overall absolute potential was divided into five levels, with the same threshold as the estimated archaeological potential;
- The comparison was performed computing the difference between the level of the overall estimated archaeological potential and the level of the overall absolute potential.

Table I.

Cores	Estimated Potential	Estimated Potential Level	Absolute Potential	Absolute Potential Level	Comparison
V1	0,000468	1	0,249423	1	0
V2	0,000467	1	0,249423	1	0
V3	0,000525	1	0,302870	1	0
V4	0,000507	1	0,249423	1	0
V5	2,534638	4	2,351700	3	1
V6	2,589417	4	2,057738	3	1
V7	2,987109	4	1,413396	3	1
V8	2,948501	4	1,009568	2	2
V9	2,943188	4	2,675356	4	0
V10	2,196394	3	1,416365	3	0
V11	2,121690	3	2,129001	3	0
V12	1,111461	2	0,605741	2	0
V13	1,195012	2	0,463214	2	0
V14	3,5171420	4	2,004290	3	1

The obtained results are shown in Table I. The proportion of exactly estimated potential levels is $9/13 = 69.2\%$, while the maximum error is 1, with an average error of 0.3077. Without referring to levels of potential, the maximum error is 1.5737 (on the core V7), while the averaged error is 0.5495, and the error variance is 0.2421. It is worthwhile noting, moreover, that when the estimation disagrees with the absolute potential of the new cores, it is always greater. This agrees with the fact that the area of the cores was about 35 cm^2 centimetres, while we are estimating the potential of a 100 mt^2 cells: a core is less informative, since it does not occupy the whole cell. Finally, we report that the core V8 has not been used for testing because it was blocked by some unidentified structures, and it did not provide the whole stratification information.

VI. Results

The final outcome of the MAPPA algorithm is an archaeological potential map for each archaeological period, but as "ongoing" products we got maps showing the absolute potentials, the geo-morphological data, and the functional areas. We put here, for all archaeological period, all these maps together with the estimated archaeological potential in order to make some technical consideration, leaving the comments of more archaeological and geological nature to [2]. For every archaeological period, we have $maxit = 3$, i.e. the PageRank was iterated three times. Indeed lower

values of maxit could be ”compensated” by higher values of *speed_up*, but in archaeological periods with very few data we needed in any case at least 3 iterations (with an high value of *speed_up*) to let the potential “spread”. In the same way, we set *rel* and *yield* to about 0.5 for each period, with some oscillations depending on the quantity of data, so that “half” of the potential is given by the relations, and “half” by the absolute potential values. The value of *speed_up*, in general, instead, is inversely correlated to the quantity of available data.

The Protohistoric period (Figure 2) is characterized by a very small amount of available data, because on one hand very few finds were discovered, and on the other hand the geo-morphological datum is not known for the majority of the cells. The reader can observe that both the functional areas and the estimated archaeological potential are affected by this scarcity of data, and indeed no real relation-based mechanism comes into play: the prediction is mainly based on the functional areas definition and on the geo-morphological datum.

Also the Etruscan period (Figure 3) is characterized by a quite small amount of available data, but, unlike the Proto-history the available data seems to be sufficient to identify a nucleus, and indeed both the functional areas and the estimation of the archaeological potential develop around a “centre”, where the relations between the geology, the functional areas, and the links, clearly interact. The reader should note that the finds outside the central area are (almost) not able to give rise to some potential around, since they are quite isolated, in a rural area, and in a “disadvantageous” geo-morphological situation.

The Roman period (Figure 4) is characterized by a certain structure in the archaeological data and a certain scarcity in geo-morphological data. Anyway the archaeological data are sufficient to identify the diverse functional areas, and to “draw” the picture of the Roman settlement. It is worthwhile noting the difference between the estimated archaeological potential and the functional areas: the suburban areas in the bottom and on the left of the picture do not correspond immediately to the same areas in the estimate of the archaeological potential. This is because the definition of functional areas is “static”, and does not take into account the interaction of finds among themselves and with the geo-morphology.

The Early Medieval Age (Figure 5), though chronologically more recent than the Roman one, is characterized by few archaeological data, even if the geo-morphological data are almost complete. Also here, like the Roman period, the data seem to be sufficient to identify a “centre”, and there is a difference between the values of the functional areas and the areas where the estimated archaeological potential concentrates.

From the Late Medieval Age (Figure 6) to the Contemporary age the geo-morphological data are complete for the work area. This period is characterized by a clear role in enclosing the urban areas of the walls, which act as a border. Anyway, both in the functional area picture, and in the estimated archaeological potential a suburban area can be noted outside the walls.

The Modern and Contemporary ages (Figure 7, Figure 8) are characterized by a relative abundance of data, and a development of the city that reflects the current structure.

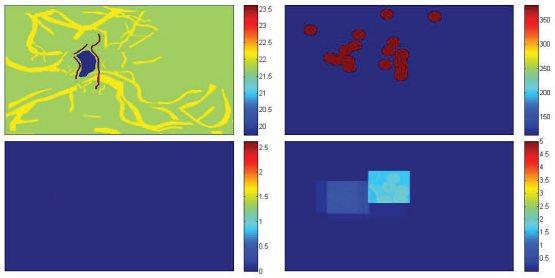

Figure 2. The Protohistoric period. From top left, clockwise a) the geomorphological datum: the rivers are in red, the floodplain in blue, the mean geomorphological datum (where data are not available) in green, and the uncertain paleochannel in yellow; b) the functional areas: in red the urban areas, in blue the rural areas; c) the estimated archaeological potential; d) the absolute potential. The values of the parameters are: $speed\_up = 0.95, rel = 0.3, yield = 0.4$: the high value of *speed_up* and the low value of rel was due to the scarce quantity of data.

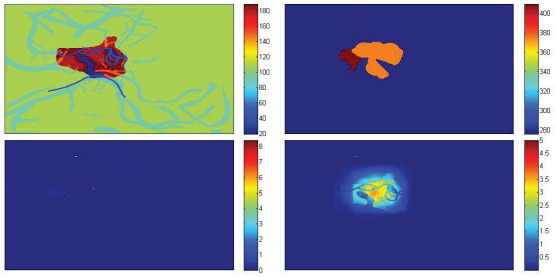

Figure 3. The Etruscan period. From top left, clockwise a) the geomorphological datum: the river and wetland are in blue, the floodplain and elevation are in red, the mean geomorphological datum (where data are not available) in green, and the uncertain paleochannel are displayed in sky blue inside the green, and in orange inside the red; b) the functional areas: in orange the urban areas, in red the suburban areas, in blue the rural areas; c) the estimated archaeological potential; d) the absolute potential. The values of the parameters are: $speed\_up = 0.8, rel = 0.5, yield = 0.5$.

The overall estimated potential map, obtained as described in the testing Section, is depicted in Figure 9, while its subdivision in the 5 levels of potential is shown in Figure 10.

VII. Concluding remarks and further developments

This section contains some conclusions and considerations of quite general type. We think that these are the most important achievements of the model

1) Datasets of different nature (archaeological, geomorphological, historical, and so on), and of the whole chronological range, were put together to produce a unique output: the estimation of archaeological potential;

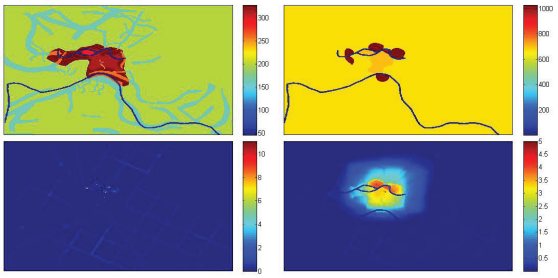

Figure 4. The Roman period. From top left, clockwise a) the geomorphological datum: the rivers are in blue, the elevation and floodplain are in red and dark red respectively, the mean geomorphological datum (where data are not available) in green, and the uncertain paleochannel are displayed in sky blue inside the green, and in orange inside the red; b) the functional areas: in dark yellow the urban areas, in red the suburban areas, in yellow the rural areas; c) the estimated archaeological potential; d) the absolute potential. The values of the parameters are: $speed\_up = 0.75, rel = 0.5, yield = 0.5$.

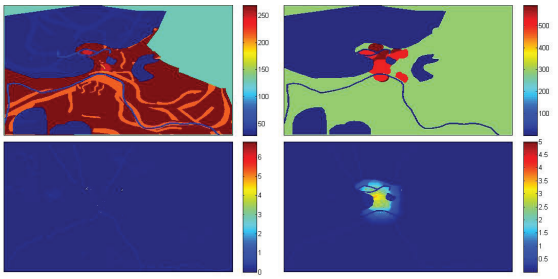

Figure 5. The Early Medieval Age. From top left, clockwise a) the geomorphological datum: the river and wetland are in blue, the floodplain and elevation are in red, the mean geomorphological datum (where data are not available) in green, and the uncertain paleochannel are displayed in sky blue inside the green, and in orange inside the red; b) the functional areas: in orange the urban areas, in red the suburban areas, in green the rural areas; c) the estimated archaeological potential; d) the absolute potential. The values of the parameters are: $speed\_up = 0.7, rel = 0.5, yield = 0.5$.

2) The whole process of estimating the archaeological potential is algorithmic, and formally defined. Of course the archaeological interpretation process comes into play, but once the parameters have been changed according to other interpretation philosophies, the algorithmic procedure works as well;
3) All the data are treated as if they are in a complex network. The archaeological potential rises through the interactions between finds, functional areas, geomorphology.

The PageRank model is a mathematical model capable of assigning archaeological potential to cells, ranking them on the basis of their interactions, and it turned out to be a good

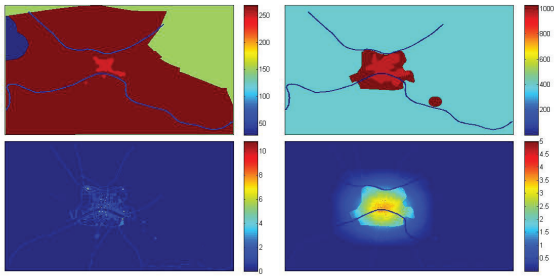

Figure 6. The Late Medieval Age. From top left, clockwise a) the geomorphological datum: the river and wetland are in blue, the floodplain and elevation are in red, the cells outside the work area are in green; b) the functional areas: in light red the urban areas, in dark red the suburban areas, in sky blue the rural areas; c) the estimated archaeological potential; d) the absolute potential. The values of the parameters are: $speed\_up = 0.7, rel = 0.5, yield = 0.5$.

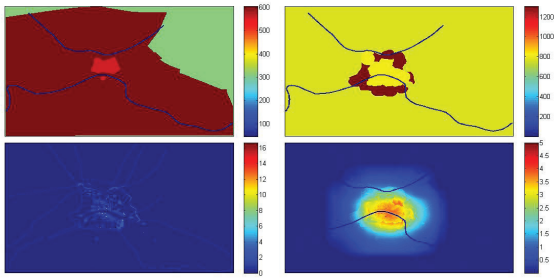

Figure 7. The Modern Age. From top left, clockwise a) the geomorphological datum: the river and wetland are in blue, the floodplain and elevation are in red, the cells outside the work area are in green; b) the functional areas: in light yellow the urban areas, in red the suburban areas, in dark yellow the rural areas; c) the estimated archaeological potential; d) the absolute potential. The values of the parameters are: $speed\_up = 0.75, rel = 0.5, yield = 0.5$.

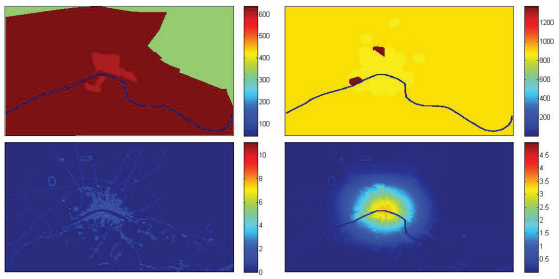

Figure 8. The Contemporary Age. From top left, clockwise a) the geomorphological datum: the river is in blue, the floodplain and elevation are in red, the cells outside the work area are in green; b) the functional areas: in light yellow the urban areas, in red the suburban areas, in dark yellow the rural areas; c) the estimated archaeological potential; d) the absolute potential. The values of the parameters are: $speed\_up = 0.6, rel = 0.5, yield = 0.5$.

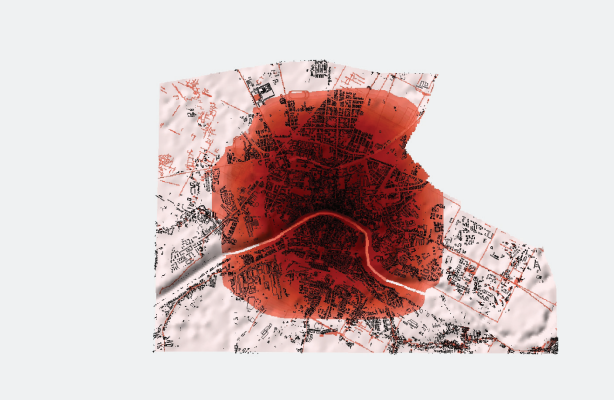

Figure 9. Overall estimated potential, computed as described in the testing section.

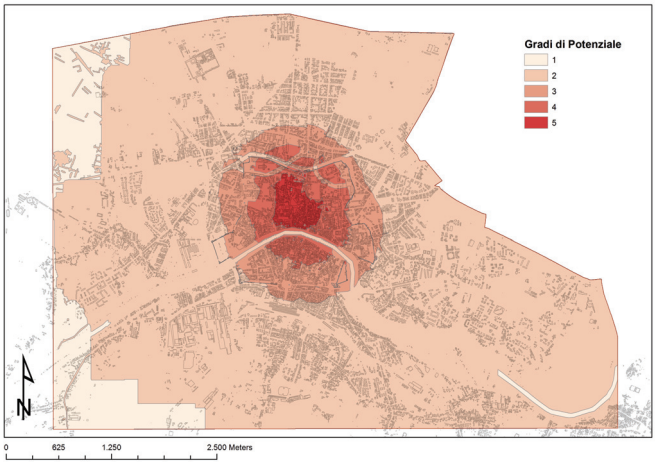

Figure 10. Overall estimated potential subdivided into the 5 levels, as described in the testing section.

choice in estimating the archaeological potential. However we point out here some possible improvements for future studies.

- The PageRank model seems to work better for those data which concentrate in an area, than for "polycentric" data. This not only has to do with the PageRank modelling itself, but also with the way we draw functional areas, because in the "mono-centric" case only the borders of functional areas should be traced, while in the "polycentric" case (like e.g. in the countryside) it is needed to guess where some functional area is. We didn't try to do such an inference, and probably a PageRank based model is not the right tool to try to guess new built up area, "independent" of the others. This will be a part of the future research, to be implemented with other models;

- How can we estimated the general goodness of the prediction provided by the PageRank model? We test the prediction with the new cores, but we have no general method, e.g. a minimum amount of data or relations, to decide whether a PageRank based model could be appropriate, for instance with respect to a more standard one. This is related also to how much data are representative of the situation, and how to measure it;

- It should be noted that we worked only on "positive" data, since no evidence of cells where data for some archaeological period is missing were available. The knowledge of areas where archaeological data are not present are so important as the cells where data are present, because they contribute to fix the absolute values of potential of cells as well. For further developments it is important to have also "negative" data, and probably to develop a model for their "spread", too, to be used in conjunction with the present 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 parameters. For these reasons, the map of archaeological potential should be always evaluated in conjunction with interpreted and raw archaeological data. In this way, the predictive map of archaeological potential is a useful and powerful tool both for land management and for archaeological research.

References

[1] 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.

[2] Anichini F., Dubbini N., Fabiani F., Gattiglia G., Gualandi M.L., (2013), *MAPPA, Metodologie Applicate alla Predittività del Potenziale Archeologico*, Volume II, Roma.

[3] Anichini F., Gattiglia G. (2012), *Urban Archaeological Information System. Considerations and critical aspects*, in Anichini F., Fabiani F., Gattiglia G., Gualandi M.L. , Mappa. Methodology Applied to Archaeological Potential Predicitivity, Volume I, Rome.

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

[5] 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 2enII, pp.45-50.

[6] Capitani M., Fabiani F., Sciuto C., Tarantino G. (2013), *Carotaggi per la verifica del potenziale archeologico* (Report), Pisa, MAPPAProject - Università di Pisa, doi: 10.4456/MAPPA.2013.21.

[7] Fabiani F., Gattiglia G., (2012), *The digital archiving structure*, in Anichini F., Fabiani F., Gattiglia G., Gualandi M.L. , Mappa. Methodology Applied to Archaeological Potential Predicitivity. Volume I. Roma.

[8] Langville A.N. and Meyer C.D., (2006), *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press.