MapPapers 2en-II, 2012, pp.45-50
doi:10.4456/MAPPA.2012.04

On the two main issues about the application of page rank models for the determination of archaeological potential

Bini D., Dubbini N., Steffè S.

This report is about the mathematical modeling of the two main implementative issues about applying the page rank algorithms to the determination of the archaeological potential: the effective determination of categories of finds, and the way of constructing the matrix of weights. As to the categories, we define them starting from the assignment of four labels for each find, deducing then the right level of generality for the definition of categories. As to the matrix of weights, we construct it based on a coding of spatial properties and functional relations of finds, mathematically properly defined. Future work will involve the practical implementation of this information to produce one matrix of weights, and some trials to fix all the different possibilities for the page rank algorithm.

Keywords: page rank, shape theory, categorization of finds, functional relations among finds

1. Introdution

In the previous report (Bini, 2011) the authors identified in page rank models the suitable properties to determine the archaeological potential of a urban area. They are good models, above all because they let the relations among the various finds, both in spatial terms (i.e., dealing with the location in space) and in functional terms (i.e., about which is or could be the function the finds are useful for), be conveniently codified in mathematical terms. These relationships are exactly the key point on which the practical determination of potential is carried on by the archaeologists.

When archaeological information is available, two kind of quantities are associated in the page rank model to each find (or more precisely to each 3-dimensional cell by which we model the subsurface): a value representing its importance, and a set of values representing the 'strength' of the influence on the archaeological potential of the cell on the other cells. The general idea was to categorize finds, and for each category to assign some parameters that provide the importance of the find and the geometry of the distribution of that importance to other finds. Finally the page rank model will put together all this information and compute the archaeological potential in each cell. So a fundamental part of the implementation of the page rank model deals with a reliable categorization of finds, and an appropriate set of parameters that can describe, for each category, the distribution of its weights.

In this report we propose a categorization based on practical considerations for the particular case of the urban area of Pisa and on previous work on the field. Moreover we will give a set of parameters that, for each category, can describe the geometry and the distribution of its weights.

The report is organized as follows: in section 2 we present the effective way of categorizing finds, while section 3 is about the way of constructing the matrix of weights, through a mathematical model of spatial properties and functional relations among finds. In

section 4 there are conclusions and future investigations, and in the appendix a brief introduction to the mathematical theory of shapes is presented, useful to specify spatial properties of categories of finds.

2. Determination of categories

A categorization for the finds is essential for the creation of an algorithm, because there's too much variability if we consider every find's properties and features. The categorization will then help to properly implement the algorithm, and specify the distributions of the archaeological potential between finds. However, in the choice of the categories, two opposite needs has to be taken into account: first, a categorization should be general enough to be applied also in different contexts (and also in different archaeological periods, as we will show), and not only in the particular case of our analysis; on the opposite, a categorization should be detailed enough to assure that the geometry, the distribution of weights, and parameters, describe the peculiarities of each category.

Just to illustrate these opposite needs in the simplest possible way, let us suppose that the categorization is the most detailed possible, i.e., each find has its own category. In this case we would assign, for each find, parameters that describe the importance and the potential 'spread' by that find. This could be also an outstanding work for the case under study, but it is completely useless for any other different case, since every category should be again completely defined from the beginning.

In order to define the categories, and to do it with a proper generality, we follow the procedure already adopted by authors for the cataloguing of archaeological data in (Anichini, 2004). This method resulted from various discussions with the archaeological team, in which their need (the particularity) of giving importance to each find was contrasted by the need of an efficient algorithmic implementation. In that procedure each find was given a certain number of labels (four), describing in an increasing order of generality the 'structure' where the find was. While the fourth label, for instance, describes what the find is, or its immediate interpretation/function (e.g. the find 'window'), the other labels describe the structures in which the object is embedded, at different levels of complexity, or of functionality, that is the same: the find 'windows' is in a '*domus*' (the third label), which in turn is an 'housing edifice' (the second label), which is a particular 'private use area' (the first label). The database we have created contains about 250 different labels of the fourth level, about 200 different labels of the third level, about 40 different labels of the second level, and about 10 different labels of the first level, the most general one. See (Anichini, 2012) for further details.

We chose the third level labels to be the base to create our categorization of finds, for the following reasons:

- the third level labels are the only one which give information not only about the place where the find is located, but also about its neighborhood. The other labels are either too specific, or too general to give information about the surrounding cells. In other words the third level labels allow for a spatial induction on archaeological potential;
- the generality of third level labels is such that at this level of complexity of structures, the complexity being due to the human work, the archaeological potential 'rises'.

It is fundamental that categories are given in such a way that each one of them provides information about the neighborhood, both for the good performance of the algorithm and for a careful determination of the archaeological potential. Since the value of the archaeological potential will be given for each cell, the possibility of providing information for neighboring cells depends also on how big the cells are.

The size of cells is one of the parameters we have discussed, and a parameter that should be discussed for each analysis, no generality here: the 'right' size depends on the density of finds, on the vastity of the area, on the richness of structures and archaeological ages, etc. So it cannot be decided a priori. In our analysis we distinguished two different reasoning, one for the determination of the surface dimensions (length and width), and one for the determination of the depth. The latter is quite immediate, because archaeological periods naturally identify a way of partitioning the subsoil. On the other hand it is nonsense having in one cell finds of two different periods. So the obvious choice is to set depth of cells in such a way that each cell is in one and only one archaeological period. Observe that in this way the cells depth can be different from point to point, but this is not a problem from the algorithmic point of view. As to the surface dimensions, we have to face two opposite tendencies. If the cells are too small, then too many cells do not contain finds, and we should assign the potential of many cells with few known values: a critical issue, algorithmically speaking. On the other hand, if the cells are too big, we can have different finds (categories) in a single cell, and the 'spread' of archaeological potential by means of weights is reduced. In the case under analysis, the urban area of Pisa, it covers a surface of 26 square kilometers, and a maximum depth of 10 meters. Considering the presence of rivers and other conditions by which some cells will have 0 potential, we have about 130.000.000 square meters of subsurface to which the archaeological potential has to be assigned. The approximate number of (categorized) finds is 2000, while their approximate dimension can vary between less than one meter and tens of meters. Taking into account all these indicators, and the considerations before,

we evaluated that a good (square) dimension of each cell can be approximately between 1 and 2 meters.

Another point to be discussed is whether the categories should be the same for each archaeological period or not. This could be another issue to be taken into account in choosing the generality of the categorization for finds. It is useful to have the same set of categories for each archaeological period, for the following reasons at least: this allows for a greater 'economy' in the algorithmic procedures, since categories and their defining parameters have to be defined once for all; the relationships among categories, also in different archaeological periods, are easier to define; this gives a push toward a sufficient level of abstraction. In our approach we chose to define the same set of categories for each archaeological period. However it must be said that we needed a slightly greater set of categories, so that we included in the set some categories that are 'present' only in some periods, and not in others (e.g. churches, not present before the christian age). Anyway, these are few in number, so we can certainly consider to have a unique overall set of categories.

3. The geometry and the distribution of weights for each category

Once the categories have been determined, and each find has been assigned to its correspondent category, the matrix of weights has to be determined. According to our first report (Bini, 2011) each category has its own 'absolute' value of the potential, and its associated set of weights toward other cells. With the absolute values, and the assignment of weights the page rank algorithm can be implemented, providing the archaeological potential as the output. This section explains the ideas behind the construction of the weight matrix.

Two main types of properties among categories have influences on the archaeological potential, which we named spatial properties and functional relationships. Spatial properties have to do with the displacement in the space of the categories, so that finds in a cell implies the presence of the object indicated by the corresponding category, but the particular distribution in the space can vary in dimension and orientation, depending on the particular case. Functional relationships among categories refer to the 'role' that links some categories to some other: e.g. a house should have a well or a garden in the nearby, and this clearly have effects on the determination of the archaeological potential.

3.1 Spatial properties of find categories

Spatial properties of each category should describe the probable displacement of the category in the subsoil. Since the archaeologists deduce the presence of each category in a particular place from the finds, and since each category can have 'realizations' of different shapes, dimensions, and orientation in the subsoil, each category should be identified through its peculiar features, while the variation of its features depending on the particular 'realization' could be codified by means of some parameters, to be chosen from time to time.

We believe that the spatial peculiar feature of each category can be identified with its shape, meant as a geometrical description of the part of that space occupied by the object, abstracting from location, orientation, and size. There exists a mathematical theory that allows for formal definitions and properties: the so called shape theory (Kendall, 1989), of which we give a brief introduction in the appendix. Anyway here the important thing is that the shape of each category, in the sense of the definition above, can be codified in a proper way as a feature of the category itself. Moreover the shape theory allows also to consider some uncertainty in the shape. This is useful, because often in the archaeological practice the shape is not completely determined or known by means of finds.

Once the shape has been assigned to every category, the particular realization of the category has to be placed in the subsoil by giving an estimation of its location, orientation and size. Sketchily, the location corresponds to a numerical parameter of translation, the orientation to a numerical parameter of rotation, and the size to a stretching numerical parameter. This set of parameters lets the particular realization of each category be 'embedded' in the subsoil. Further details on the way of assigning and codifying shapes is given in the appendix section.

3.2 Functional relationships among categories

Functional relationships among categories should estimate the probability of each category to have another category in the space nearby, because of their common occurrence, due to their function. A simple example of a functional relationship is the contemporary presence of a house, together with a street, a well, a garden, which are very often near a house, because their function contributes to the same aim any house is built for. So what we called functional relationships are given by categories that often occur together because they are built to reach a common goal (like the above house, well, street, and so on), that can be practical, spiritual (like a church together with bell tower, or with a cemetery), or of other kind.

Observe that functional relationships among categories can involve quite often relationships across different archaeological period, that are 'functional' in a further sense. For instance, often on the ruins of some finds included in a particular category, in the next archaeological periods, the presence of some categories is much more probable than others. This may be given by a 'continuation' in different periods of the same function of the category (e.g. churches), or by the fact that the same materials are used to build something else, or so. In this case the functional

relationship, among categories and among archaeological periods, is a sort of property of inertia of each category.

Given the previous considerations, there are two main points about functional relationships among categories. The first one is to estimate the probability of the presence of another category, give the presence of a first one. This will be done by creating a 4-dimensional array $M \in [0,1]^{N\times T\times N\times T}$, where N is the total number of categories used for the analysis, and T is the number of archaeological periods. The entry $M_{i,j,k,l}$ will represent the probability of the presence of the category i in the archaeological period j, given the presence of the category k in the archaeological period l. This will be estimated with different methods, depending on the particular case. For example, such a probability could be given by 'standard' procedures of that particular archaeological period or of that place, or by estimating it with available (historical or other kind of) data.

The second main point of functional relationships is about the place where each category is located. Once we have guessed that the presence of a category implies in some way the presence of another category, in the same archaeological period or not, an estimation of the place (i.e. of the cells) occupied by this latter category should be performed. That's why the functional relationships have to be considered after the spatial relationships. Each category should at this point already have its shape (in the sense we specified before) and its parameters describing the displacement in the space. In this way, we will assign to cells nearby probabilities contributing to the weight matrix of the page rank algorithm depending on the categories.

Appendix

We give in this appendix the basics of the mathematical shape theory, as first developed by (Kendall, 1999), which will be used to assign spatial properties to categories. Maybe it is worthwhile to mention that one of the practical problems from which the mathematical shape theory originated was an archaeological problem. Quoting (Kendall, 1989): "Thus the set of 52 standing stones near Land's End, Cornwall, studied by (Broadbent, 1980), yields $\binom{52}{3} = 22.100$ triplets of stones, and there are those who say vaguely that 'too many' of these are 'too nearly' collinear, and to attribute this to deliberate planning, whereas others dismiss such claims as ridiculous. Who is right?". So the problem here was how to quantify the property of being 'too nearly collinear', for stones of that particular archaeological site.

The first question in the mathematical shape theory was how to mathematically define the shape of a set of *k* not totally coincident points in the space. For what concern this report, we consider only the case of a 2 or 3-dimensional spaces, despite the shape theory treats spaces of any dimension. As we already observed, the idea is that of filtering out the effects of translations, change of scales, and rotations. We'll now describe step by step how the mathematical definition of shape is obtained:

1. Consider a set of k points in the 3-dimensional space, given as a matrix $X \in R^{3\times k}$, where each column is the vector containing the coordinates of one of the *k* points;
2. Take the barycenter of those points to the origin, so that the translation parameter will be then the vector of the coordinates of the barycenter of the *k* points. In formulas, for every column vector X_i, i=1,...,k;
$$X_i \to X_i - \frac{X_1 + \ldots + X_k}{k};$$
3. To eliminate the effect of size in defining the shape, consider the quantity $L = \sqrt{\sum_{i=1}^{k} \|X_i\|^2}$, i.e. the norm of the vector made by the distances between each point X_i and the origin. Divide each column vector X_i by *L* so that *L* equals then *1*:
$$X \to \frac{1}{L} X = \frac{1}{\sqrt{\sum_{i=1}^{k} \|X_i\|^2}} X;$$
4. Now the rank of the matrix *X* is at most *2*, so we can multiply *X* by that element of the orthogonal group *O(k)* that sends *(0,...,0,1)* to the element of R^3 with all coordinates equal to $1/\sqrt{k}$. *X* now has the last column equal to *0*, and so we consider $X \in R^{k\times 2}$;
5. Now that *X* is made of *k-1* vectors which sum up to a unity vector, we can identify *X* with a point of the sphere of unit radius and dimension *2k-1*. This sphere is called the preshapes space, and is denoted by S_k^3;
6. The preshape sphere is then identified with the space of matrices of $R^{k\times 2}$ on which the special orthogonal group *SO(k)* acts from the left. We now define Σ_k^3, i.e., the shape space of *k* points in *3* dimensions, to be the quotient space of the preshape sphere by the special orthogonal group:
$$\Sigma_k^3 = S_k^3 / SO(k).$$
In this way also the effect of rotations is filtered out, and the definition of the shape space in obtained.

Having defined the shape space, with each category of findings will be associated its own shape, decided on the basis of available data of each archaeological period on the shape of the find. Note that shapes are allowed to be specified with some uncertainty: this is very useful in the archaeological practice, because the shape of categories of finds are not known with precision. Mathematically speaking, it is possible since on the shape space a (many) distance could be defined, in such a way that we can assign an 'approximate' shape: we choose a point in the shape space, and the points near that represent shapes that are approximately the one we chose. Here the term 'near' means within a certain small distance to the original point on the shape space, the distance being that one defined on the shape space.

4. Conclusions and future work

In this report we handled some implementative problems, relative to the application of the page rank algorithm to the determination of the archaeological potential. The report is a continuation of the first one (Bini, 2011), where the authors proposed the application of the page rank algorithm, which is mostly used to classify and give 'mportance' to web pages based on the links they send and receive. The use of the page rank was motivated by the fact that, from an abstract viewpoint, relationships among finds are the most important element that contributes to the archaeological potential, and these relationships show properties similar to the links in page rank algorithms. Of course, some modifications have to be done in order to adapt the page rank algorithm to the computation of archaeological potential. This report was about the main such modifications:

- the determination of categories of finds, which is necessary for an algorithmic storage of archaeological data and an algorithmic computation of the potential;
- the way of constructing the matrix of weights in the page rank algorithm, which we proposed to be done exploiting spatial and functional relationships among (categories of) finds.

We proposed, for spatial properties, to store the shape, understood in a mathematical sense, of each category of finds, i.e., geometrical description of the part of the space occupied, abstracting from location, orientation, and size. After that each particular realization of a category can be specified through some parameters that describe the location, the orientation, and the size. For functional relationships, we proposed to construct a matrix whose entries indicate the probability of the presence of a category, given another one.

As for future work, an important step will be the way of storing and using all those information - i.e., the shape and the displacement of each category, the functional relationships among categories, the geological information - to produce *one* matrix of weights, to implement the page rank algorithm. Many trials have to be done to fix all the different possibility for the page rank algorithms, and to give the best possible values to weights.

Bibliografia

Anichini F. 2005, *Tutela, ricerca, valorizzazione del patrimonio archeologico: progetto per il G.I.S. della città di Pisa*, t.d.l. Università di Pisa.

Anichini F., Fabiani F., Gattiglia G., Gualandi M.L. 2012, A *database for archaeological data recording and analysis*, in MapPapers 1en-II, pp.21-38.

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

Broadbent S.R. 1980, *Simulating the ley-hunter (with discussion)*, Journal of Royal Statistical Society Ser. A, 143, pp. 109-140.

Kendall D.J. 1989, *A survey of the statistical theory of shape,* Statistical Science 4(2), pp. 87-120.

Kendall D.J., Barden D., Carne T.K., Le H. 1999, *Shape and shape theory, Wiley series in probability and statistics*.

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