Dealing with Incomplete and Uncertain Context Data in Geographic Information Systems

Jonathan Frez
School of Informatics & Telecommunication Engineering
Universidad Diego Portales
Santiago, Chile
jonathan.frez@mail.udp.cl

Nelson Baloian
Department of Computer Science
University of Chile
Santiago, Chile
nbaloian@dcc.uchile.cl

Gustavo Zurita
Information Systems and Management Department
University of Chile
Santiago, Chile
gzurita@fen.uchile.cl

José A. Pino
Department of Computer Science
University of Chile
Santiago, Chile
jpino@dcc.uchile.cl

***Abstract*—There are currently a growing number of people using smartphones or tablets, thus being potentially online at every moment. There are many useful applications using people's context data, to provide services to mobile telephony/internet subscribers. Location data is particularly interesting. These applications use location data assuming it is correct, which is sometimes not the case. In this work we propose a methodology for using incomplete/uncertain information to answer questions which include uncertainty like: "Which is the probability of finding exactly N persons within the geographic area A from time T1 to time T2?", or "Which is the probability of having a traffic jam on street S between times T1 and T2?". We also consider some logical constraints on the data. For instance: "Exclude counting people on the subway or inside buildings because the advertising will be on screens at open air". Our approach uses Dempster-Shafer theory combined with an ontological definition of variable types sharing similar probabilistic behavior. The whole process and the results are explained using an example case based in one of the busiest areas of the world (the Shibuya Station in Tokyo, Japan), consisting of underground train lines, surface transportation, large avenues and shopping centers. A language to describe the fuzzy scenarios is also introduced along with an application which allows users to generate and visualize 2D and 3D suitability maps using this language.**

Keywords—mobile data; probabilistic scenarios; Dempster-Shafer theory; location information

I. INTRODUCTION

During the last few years we have witnessed a fast development of mobile computing. There is currently a growing proportion of the population living in big cities having a smartphone or a tablet PC. This means, they are potentially online almost at every moment. This scenario has made location related information important and interesting at the same time. There are already many useful applications using context information to provide services to mobile telephony/internet subscribers, most of them based on location information. For example, there are applications to easily find the restaurants near the subscriber's current geographical location. Another example is an application to find out if there is currently a traffic jam in a certain street or not. These applications use location data obtained directly from the sources (GPS, triangulation of cell antennas) and use it assuming it is correct, which is sometimes not the case. On the other hand, these applications generally do not use other probabilistic context data to answer interesting questions involving the probability concept. For example, when marketing targeted advertisements on digital screens located on streets in various parts of the city, it might be useful to answer questions dealing with information uncertainty such as "Find a place in the city with number of people greater than X with a given profile with probability Z". For matching services offered to the public we could try to answer the following question: "Which is the probability of finding people with certain characteristics X in the area delimited by the points [x1,x2] x [y2,y2] between times T1 and T2?". Or "Which is the probability of having a traffic jam on street S between times T1 and T2?".

In order to answer the aforementioned questions, we need to know the location and/or the probability of being at a given place for each person during a certain time. There are several methods to collect this data; perhaps the most popular one is using mobile phones location services. However we maybe want to use various sources of information to get a more accurate picture about where has been a certain person and which is the probability of this data. In order to combine more than one source of information we must deal with many details, for example, the cartographic scale of the source, the methodology error, or even the confidence level of the samples. We must also consider some logical constraints on the data. For example: "Exclude counting people who are on the subway or inside buildings because the advertising will be on screens at open air".

978-1-4799-3775-2/14/$31.00 ©2014 IEEE

In this work we present a method which combines and processes data in order to answer questions like those mentioned above. One of the assumptionS is that we already have the multiples sources, and the data is represented using a mean value and a standard deviation. For example: on a certain location we may have a mean density of 5 *persons/m*^{2} with a standard deviation of 2 [1]. In order to further process the data, we will use *Dempster-Shafer theory* [2] combined with an ontological definition of variable types sharing similar probabilistic behavior. It is important to highlight that the Dempster-Shafer theory is based on reliability or accuracy of the information [3], Therefore, various sources (especially mobile ones) should have different properties concerned with this aspect.

An important contribution of this work is the introduction of a "query language" to formally ask questions.

II. RELATED WORK

Fuzziness in GIS has been studied in several ways, e.g., by defining class memberships functions [3] [4], or through graphical ways to represent fuzzy boundaries [5] [6]. In the last ten years we have seen the emergence of works exploring multi-criteria data analysis in spatial information [7] [8]. These works propose the application of fuzzy measures which are instances of fuzzy membership. However, these works are based on complete and trustable data, which is not the most common case.

Geographic information is usually represented in a discrete manner, with discrete geometries and deterministic values. However, a variable like temperature over a period of time becomes a random variable with frequency distributions and average values. Geographical points can be seen as random variables as well because they are mostly created using sampling procedures (e.g. GPS). Information associated to the geometries of the areas being represented could also have probabilistic and/or epistemic properties, e.g., which is the probability that the boundary runs exactly over a certain line?

In this context, there are some exploratory works that use belief functions to query geo-referenced data [9]. Preliminary results show that combining fuzziness with a belief function it may be possible to obtain "good" results with much less data and resources than using a traditional approach.

The literature regarding geographic information with epistemic properties shows a trend to use belief functions, in particular *Dempster-Shafer Theory* [2]. The *Dempster-Shafer theory* was developed in 1967 by Dempster and extended by Shafer. It proposes to use sets of hypotheses regarding a variable (e.g. the temperature values in X are always between t1 and t2) associated with a probability of being correct. In order to explain this theory, we will use an example: Table 1 shows mean number of persons values associated to a certain location. In addition we have a query Q = [13-23] looking for locations with more than 13 and less than 23 persons. In this case, 3/5 of the locations meet this condition (locations 1 and 3 do not). Now Table 2 contains a "range" of persons registered for each location. In this case, only 2/5 of the locations fully satisfy the condition (positions 4 and 5) and 2/5 more may have a possibility to satisfy it (positions 1 and 2). One location does not fall within any interval of the query range (position 3). The theory defines three types of answers to queries:

- Plausible: is the probability that the random variable takes values within the range of the query.
- Certain: is the probability that the whole range of the distribution of variable (D) is within the range of the query.
- Uncertain: no valuable information can be derived from this data

Location	Mean of #persons
1	12
2	20
3	7
4	19
5	17

Table 1. Location versus mean number of persons

Location	Range of #persons
1	[9-21]
2	[12-23]
3	[5-10]
4	[17-20]
5	[14-22]

Table 2. Location versus range of number of persons

Using the *Dempster-Shafer* evaluation, we can calculate the hypotheses (Plausibility, Certainty and Uncertainty) for each location for the example shown in tables 1 and 2. We see that the Certainty level is 40% and Plausibility level is 80% (Table 3). These values are considered as lower and upper bounds of possibility, i.e. between 40% and 80% of the locations have some possibility to have a similar number of persons to the queried range. Additional to this information, the theory states that a certain weight should be given to each hypothesis. This weight should be assigned by a human expert or a heuristic. Table 4 shows an example where this weight is assigned by an expert. We will show an example where these weights are calculated by a heuristic formula in the next section on this paper (on table 7).

Loc.	Persons	Hypothesis
1	[9-21]	Plausible
2	[12-23]	Plausible
3	[5-10]	Uncertain
4	[17-20]	Certain
5	[14-22]	Certain

Table 3. Location/Persons D-S for Q=[13-23].

Persons	Weight
[9-21]	20%
[12-23]	15%
[5-10]	35%
[17-20]	20%
[14-22]	10%

Table 4. Example of weights assigned by an expert

In this case, since Q = [13-23], the certainty is 30% (20% from location 4 plus 10% from location 5) and the plausibility is 65% (20% from position 1 plus 15% from position 2 which are plausible, plus 30% from the certain positions 4 and 5).

III. SOURCE UNIFICATION MODEL

The *Dempster-Schäfer* theory does not consider the case when two or more hypotheses apply to a certain location, for example, when locations 2 and 3 have some area of intersection. This is a very common problem when working with data obtained form various different sources. In order to

combine multiple sources of information we are going to define that each element on the map (in our example areas) has a **Polygon name** (definition of a certain geographical area by a polygon), **Average value** (e.g. persons/m^2), **standard deviation, Certainty** in %, and **Plausibility** in %, thus the input information we consider will be tables like the one shown in table 5, provided by various sources.

Polygon	Average	Std. dev.	Certainty	Plausibility
Pol_1	4	2	20%	50%
Pol_2	6	2	15%	30%
Pol_3	3	3	5%	20%
Pol_4	5	2	30%	60%
Pol_5	4	3	17%	34%

Table 5. Location/Persons weights

As example we will use the scenario of analyzing the possibility of finding a certain number of persons in a metropolitan area like Shibuya in Tokyo, as shown in Fig. 1. In this figure the various available data sets provided by different data sources are represented by polygons (**Pol_i**); the data represents human density (*persons/m^2*).

Fig. 1. It shows multiples sources, each one represented by a geometric element, shown with same color, with a degree of transparency, so the color has more density when these elements overlap. For the rest of the map there is no data.

The combination process of merging data from various sources must be executed for each polygon. In order to perform this computation the entire area which should be analyzed is divided into small squares (10x10 meters), and the computation process is executed for each of those squares. In order to answer the query we need to represent the real world in a proper way.

The first step is to compute the probability of having a certain range of human density between two values, v1 and v2 for each square. This is done by using a probabilistic representation of the values.

When the values are discrete and the cumulative probability distribution (CDF) associated with the random variable is known, we can estimate it from the samples. For example, in our scenario we can obtain aerial photos from Shibuya, select some representative areas and count the number of persons per square meter inside them. Using the distribution functions derived from the samples we can calculate the probability of having a certain range of number of persons per square meter in that area (**prob**).

The second step is to include a distance probability (**pdist)**; this value will decrease (using a normal distribution) according to the distance from the edges of the area to the evaluated square. This is because we assume that the geometric representation has some error margin with respect to reality. If the square is in the area the probability is 1. Using a fuzzy representation of the polygons we estimate the probability that the data from the polygon influences the result data in each square. Finally, we calculate a set of new certainty values (**cert**) pondering the range probability with the presence probability with the certainty of each polygon present on the square.

$$cert(square) = cert(Pol_j) * pdist(square, Pol_j) * prob(v1, v2) \quad (1)$$

The result of the previous process is a set of Certainty values for each square, and each value corresponds to a pondered version of the source initial value. In order to combine the Certainty value of each source in the square we use *Dempster-Shafer* combination rules (see mathematical formulas in [10]). In Fig. 2 the certainty is shown in a red-yellow-green scale (red 0%-green 100%).

The result of the combination is a fuzzy certainty function in 2 dimensions, space and probability of occurrence. In Fig. 2 we see the results of applying (1) for the query "which is the certainty of finding between 4 and 7 persons per square meter?"

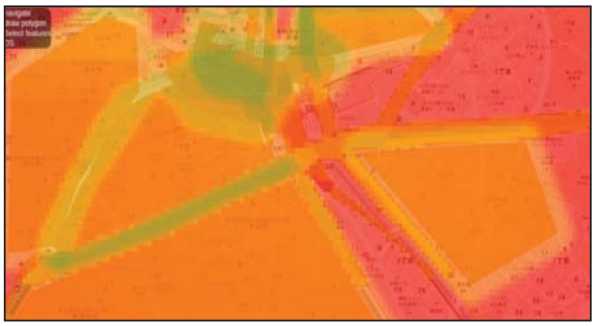

Fig. 2. Fuzzy Certainty representation.

In Fig. 2 there are no squares without data; this occurs because the according to the propagation rules, the probability never goes to 0 even in zones where no data was available. By zooming out the figure, it is possible to see that according to it, there is a non-zero certainty value of finding persons on the rails of the train station; of course this is a wrong conclusion. Since the certainty propagates over the map, it becomes important to define a way to include rules, constraints, or other characteristics in the processes which take into account these kinds of situations. Our approach to avoid this problem was developing an Ontology structure for the data types with class interactions methods. The interaction methods are executed for two or more objects on each square, and they can modify

the Certainty value of the square. The base class definition of the ontology is shown in table 6.

```
GeoObject{
        geometry, //area description (polygon, line, point)
        val,  //mean value of the area
        valattr, // density probability function attributes
        cert, //certainly
        plau, //plausibility
  pinterval: function(from,to), // density probability function
  pdist: function(geo), //distance probability function
  addInteraction: function(classname,func), //define interactions between objects
  interact: function(object), }  //executes an interaction with an object
```

Table 6. Structure of the ontology defined via GeoObjects

Using this base class we can extend a "person", a "rail", a "street" and a "building" class. The "rail" and "building" classes have an interaction with "person", this interaction returns 0 value of certainty (because we want to exclude that possibility). The street class has an interaction with "person", this interaction returns 1.2 (amplify this possibility). The new certainty definition which includes the interactions is shown in Table 7.

cert : $\text{Geoobject}_j -> cert$

pdist : $\text{Geoobject}_j -> pdist(square)$

prob : $\text{Geoobject}_j -> pinterval(v1, v2)$

interaction: foreach(Geoobject $-> i$) {

$interaction *= Geoobject_j -> interaction(Geoobject_i)$ }

$$Cert(square) = cert(Pol_j) * pdist(square, Pol_j) * prob(v1, v2) * interaction \quad (2)$$

Table 7. Calculation of the Certainty value for a square including the interactions with GeoObjects

Figures 3 and 4 show the results of the interactions processing applying (2).

Fig. 3. 2D Fuzzy certainly with GeoObject interactions.

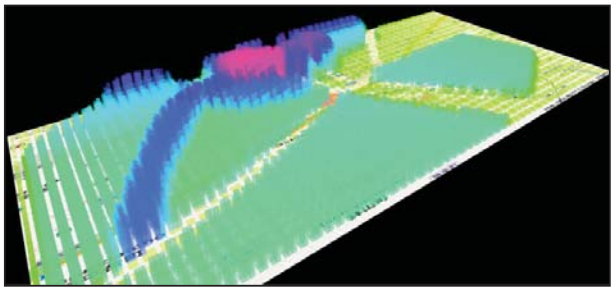

Fig. 4. 3D Fuzzy certainly with class interactions.

IV. DECISION SUPPORT AND GIS WITH UNCERTAINTY

According to [16] Decision support systems (DSS) are computer technology solutions that can be used to support complex decision making and problem solving tasks. The complexity of decision making processes is often due to uncertainties in the available data which is important to make the right decision. Geographic Information Systems (GIS) are often used to support decision making processes for which intensive use of geo-referenced information is needed in order to generate and evaluate the outcome of the various alternative scenarios. As argued in the first section, this type of data is often uncertain and based on hypotheses. As a consequence, decision making processes related to geographical issues often involves among other activities, generating a large set of different maps, each one showing the results of applying different hypotheses regarding the alternatives with multiple evaluation criteria. In geography, these maps have been called "suitability maps" [17]. In the previous sections we described a process for generating suitability maps with uncertain data generated by various sources. In order to take advantage of that process a decision maker must have the possibility of generating them in a flexible and versatile way. Making an analogy with a user not necessarily expert in computer science consulting a database with simple SQL sentences to find the desired information, in order to simplify the alternatives generations, modification and impact evaluation, we designed a simple Scenario Generation Language (SGL). SGL is inspired by SQL (Simple Query Language), however it is not designed to query data; it is designed to generate a scenario based on expert knowledge, empirical data, and existing environmental models. SGL is divided in three main statements: Analysis definitions, hypothesis definitions and model restrictions.

In the analysis definition part, the decision maker can define the kind of visualization to be generated, and apply some basic filters. For example, to generate the certainty map for persons density in a range between [2,4] persons/m^2 filtering the results for shops with capacity bigger than 20 and a certainty larger than 30% the user should input the following sentence:

```
"certainty-map @persons[2,4] where
@shops.capacity > 20 and certainty > 30%"
```

In the Hypothesis definitions part, the expert can express his knowledge using the *Dempster-Shafer* framework, for example, if the expert is looking for persons, then one hypothesis may be "*the given range of persons/m^2 are in cinemas with a certainty of 20%*" or "*the given range of persons/m^2 are in schools or workplaces with a certainty of 30%*" or "*the given range of persons/m^2 are in shops with a certainty equal to the one in coordinates x,y*". In the hypothesis statement the expert can define multiple hypotheses, which are combined using Dempter-Shafer combination rules. Furthermore, this complex scenarios are designed by the expert without requiring any GIS expertise.

```
"hypothesis {@cinema}20%
{@school,@workplace}30%  {@shops}?  at
point(X,Y)"
```

Finally, the model restriction statement is designed to represent real world interactions between the elements in the data source. For example, a high density of people is not expected in a lake or sea. Contrarywise, we expect a high density of people inside a sports stadium. This kind of interaction complements the behavior statement by adding environmental rules. This rule can be expressed using values in an interval. For example, if we are generating a scenario for persons, we add in the model statement a -100% value for lake areas and 50% value for stadiums. This value can decrease or increase the certainty level in the indicated area. The following command defines this rule.

```
"Model @stadium{50} @lake{-100}"
```

A full Scenario definition will look like the following:

```
"certainty-map @persons[2,4] where
@shops.capacity > 20 and certainty > 30%
hypothesis {@cinema}20%
{@school,@workplace}30%  {@shops}?  at
point(X,Y) Model @stadium{50} @lake{-100}"
```

The SGL language explained above has been designed to be extended in order to allow users to specify other types of maps which could be generated (not only suitability maps based on certainty but belief or plausibility maps) and to incorporate other filters like e.g., defining a polygon where the map should be shown, and depending on the scenario and data available, specify temporal attributes for the data which should be used to generate the map.

V. Prototype Implementation

As proof of concept, a prototype system was implemented in order to explore the feasibility of constructing a decision support tool based on the concept presented in this paper: on the one hand, combining fuzzy data or data from various sources with different degrees of certainty, and on the other, using a command language for generating suitability maps in order to assist the decision maker in analyzing the results of applying various hypotheses on a certain scenario. The application first asks the user to define a project, which mainly consists of a set of objects, each one with the following attributes:

- **Identification Attributes:** Name, Description, Source identification
- **Value Attributes:** Mean value, Probability Distribution Function (PDF) of the mean value, PDF attribute values. Standard Deviation in the normal function case.
- **Spatial Attributes:** Discrete Geometry representation, Spatial fuzziness function and its attributes.
- **Dempster Schaffer Attributes:** Weight, Certainty, Plausibility

After all elements of the project are entered (and eventually stored persistently) the user can start "querying" the system and generating maps. Figure 5 shows the interface of the system while performing this task. As we see, the left half of the interface is used to enter the queries and the right half to show the results. SGL commands can be entered in 2 different ways, in order to help beginners using the system in an easy way and allow experts take all advantages of using SGL commands at the same time. For beginners there is a QBE (Query by example) – like section where a simple SGL command can be assembled by selecting and/or typing in the parameters of the sentence. After doing this, when the user presses the button labeled "Evaluate" the application builds the corresponding SGL sentence and computes the corresponding map which is shown at the right half of the screen. The Generated map can be stored and retrieved later. The application can show a set of previously generated maps at the same time, side by side, in order to support the user comparing the results of different hypotheses.

The parameters for the pre-formed SGL sentences intended for first users are:

- Type of map: Certainty map, Belief map, Plausibility map, Selectable from a pull-down menu
- Evaluation element: a pull-down menu shows all types of elements that where inputted to the project which can be measurable. In our example we used people density and were introduced with the class name "persons". All input measurable elements will be displayed in a pull-down menu.
- Range: this is the range of values for the evaluation element which will be considered for the calculation of Certainty, Belief or Plausibility. The numbers for bottom and upper limits should be typed in.
- Hypotheses: currently there is a list of sites with special characteristics available for all projects which can be selected from a pull-down menu. For each selected one, a % value for the hypotheses should be input. The list of amenities can be tailored for each project. Multiple hypotheses can be defined.
- Rule: The same list of special sites presented by the pull-down menu for defining hypotheses is presented

here. The definition of the rule follows the same logic as the hypotheses.

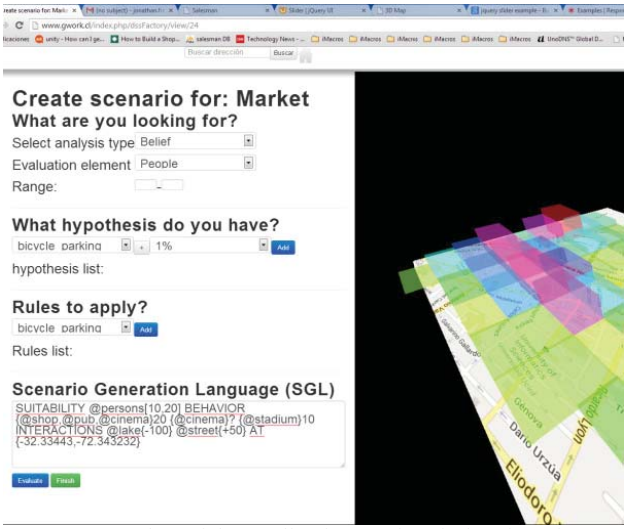

Fig. 5. Screenshot of the application main window interface

VI.CONCLUSIONS

In this paper we have presented a process for using geo-referenced data from various sources which have various degrees of certainty to be combined in order to answer questions about the possibilities of finding certain elements in a geographical area. Here we presented the example regarding the probable number of persons in a certain area, which is strongly related to context awareness. However, we see that this method can also be applied in many scenarios e.g. for calculating the probability of finding a certain vegetal or animal species in a geographical region given its climate and geomorphology or calculating the probability of the outbreak of a certain disease given the data of occurrences in nearby regions and general conditions of the weather.

The contributions of this paper are threefold. First, it shows a way to deal with maps containing uncertain or incomplete data. Furthermore, it is possible to work with data with varying degrees of certainty. Second, it introduces SGL, a language designed to formally describe and query uncertain data associated to maps. Third, it presents a software prototype to process SGL statements. This software has been developed in HTM5 which means that can be run on virtually every modern internet browser, including those from mobile devices.

All in all, the prototype is a step towards advancing traditional GIS capabilities from basic map displaying functionalities to GSS features: individual users by themselves or collaborating ones may make decisions based on gathered probabilistic geo-referenced data from various different sources.

As presented in the Introduction, the widespread availability of geo-referencing mobile devices may make systems evolved from the prototype presented in this paper very useful.

VII. REFERENCES

[1] D. Pereira, L. Loyola, «Inferring User Context from Spatio-Temporal Pattern Mining for Mobile Application Services,» de Procs. of the The IEEE/WIC/ACM International Joint Conferences on Web Intelligence

[2] G. Shafer, A mathematical theory of evidence, vol. 1, Princeton University press Princeton, 1976.

[3] L. Hegarat-Mascle, I. Bloch, D. Vidal-Madjar, «Application of Dempster-Shafer evidence theory to unsupervised classification in multisource remote sensing,» Geoscience and Remote Sensing, IEEE Transactions on, vol. 35, nº 4, pp. 1018-1031, 1997.

[4] R. Cheng, D. V. Kalashnikov, S. Prabhakar, «Querying imprecise data in moving object environments,» Knowledge and Data Engineering, IEEE Transactions on, vol. 16, nº 9, pp. 1112-1127, 2004.

[5] U. C. Benz, P. Hofmann, G. Willhauck, I. Lingenfelder, M. Heynen, «Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information,» ISPRS Journal of Photogrammetry and Remote Sensing, vol. 58, nº 3, pp. 239-258, 2004.

[6] W. Shi, K. Liu, «A fuzzy topology for computing the interior, boundary, and exterior of spatial objects quantitatively in GIS,» Computers & geosciences, vol. 33, nº 7, pp. 898-915, 2007.

[7] F. Wang, G. B. Hall, «Fuzzy representation of geographical boundaries in GIS,» International Journal of Geographical Information Systems, vol. 10, nº 5, pp. 573-590, 1996.

[8] Y. Chen, S. Khan, Z. Paydar, «To retire or expand? A fuzzy GIS-based spatial multi-criteria evaluation framework for irrigated agriculture,» Irrigation and drainage, vol. 59, nº 2, pp. 174-188, 2010.

[9] H. Jiang, J. R. Eastman, «Application of fuzzy measures in multi-criteria evaluation in GIS,» International Journal of Geographical Information Science, vol. 14, nº 2, pp. 173-184, 2000.

[10] M. H. Tangestani, «A comparative study of Dempster--Shafer and fuzzy models for landslide susceptibility mapping using a GIS: An experience from Zagros Mountains, SW Iran,» Journal of Asian Earth Sciences, vol. 35, nº 1, pp. 66-73, 2009.

[11] S. Schaer, Mapping and predicting the Earth's ionosphere using the Global Positioning System, vol. 59, Institut für Geodäsie und Photogrammetrie, Eidg. Technische Hochschule Zürich, 1999.

[12] J. B. McDonald, Y. J. Xu, «A generalization of the beta distribution with applications,» Journal of Econometrics, vol. 66, nº 1, pp. 133-152, 1995.

[13] D. Kundu, M. Z. Raqab, «Generalized Rayleigh distribution: different methods of estimations,» Computational statistics \& data analysis, vol. 49, nº 1, pp. 187-200, 2005.

[14] F. A. Haight, F. A. Haight, «Handbook of the Poisson distribution,» 1967.

[15] M. Annett, «The binomial distribution of right, mixed and left handedness,» The Quarterly journal of experimental psychology, vol. 19, nº 4, pp. 327-333, 1967.

[16] J. P. Shim , M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, C. Carlsson, Past, present, and future of decision support technology Decision Support Systems vol. 33 pp. 111 –126, 2002

[17] L. D. Hopkins Methods for Generating Land Suitability Maps: A Comparative Evaluation Journal of the American Institute of Planners vol. 43, n°4, pp. 386-400, 1977