SIMILARITY AMONG MULTIPLE GEOGRAPHIC REPRESENTATIONS

Vagner B. N. Coelho [*], **Julia C. M. Strauch** [†] **and Claudio Esperança**

Centro de Tecnologia no Campus da Ilha do Fundão, Laboratório de Computação Gráfica (LCG)
Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ, Brazil
vcoelho@ime.eb.br, julia.strauch@ibge.gov.br, esperanc@lcg.ufrj.br

Commission II/2

KEY WORDS: Multiple Representation, Information Integration, Dataset ambiguity, Cartographic Similarity Index

ABSTRACT:

A key ingredient of systems aiming to cope with multiple representations of geographic features is some method for assessing the correspondence and similarity of such representations. In other words, given two objects from two different data sources, one must be able to tell whether they model the same real world object and, in this case, measure their degree of similarity. This paper proposes an adaptation of the Equivalents Rectangles Method (ERM) to quantify the average distance between ambiguous cartographic representations and uses the Cartographic Similarity Index (CSI) – an index based on areal distances – to evaluate how much a given geometric representation resembles another. To validate the proposal, a prototype system was implemented and experiments were conducted on two geographic databases from two different institutions responsible for mapping the city of Rio de Janeiro. These were first matched using feature names in order to independently establish object correspondence. Then, the *ERM* and *CSI* of 159 districts that make up the city were computed. Results show that 157 districts have an *Adapted ERM* lower than 100.00 m and a CSI of 70% or greater. The method was thus able to detect 2 districts with significant dissimilarity, and these conflicts were later confirmed visually, indicating survey errors. In summary, while the proposed method is being used in a larger framework for *ad hoc* querying geographic data with multiple sources, it is also useful in other circumstances, such as in a preprocessing stage for data source integration or for assessment of data source quality.

1 INTRODUCTION

In many countries, geographic surveys of the same area are frequently developed by different agencies or companies. As a result, the results may differ significantly, even when the employed methodology is similar. It is also common for the *post-processed* geospatial data to be made available via web, meaning that, in principle, any interested party may access it. If two or more surveys of the same feature are available, one must, as a rule, either choose one of them or spend significant effort in integrating these data sources into one unambiguous data set. In other words, geographic databases assume that data about a given feature is unique, correct, and representative of physical reality (see Fig. 1.a).

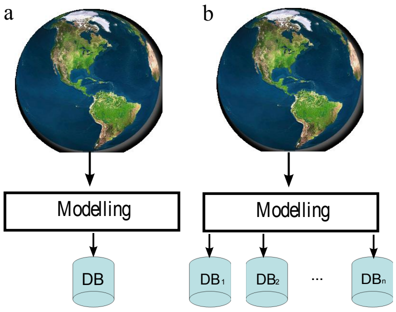

Figure 1: Single and multiple representation

A related problem arises when a given producer employs a given cartographic methodology for surveying a certain theme, but this must later be matched against data pertaining to another theme which was created using some other methodology. This can generate sliver polygons and can easily lead features over other incompatible features, like roads lying inside lakes.

In a nutshell, the current paradigm for modeling and querying geographic databases requires error-free and unique representations. This is a well established concept and was well summarized by Spinoza:

> "There can not exist in nature two or more substances with the same property or attribute"(de Spinoza, 2005).

Of course, this fact is, rationally, readily understood and accepted by human intuition.

To achieve this paradigm, one has to avoid the conflict between data from different producers. Several approaches are common for obtaining a database with no conflicts by data integration. Some of these are the use of Digital Libraries (Pazinato et al., 2002), the Clearinghouse (Goodchild et al., 2007) and the Data Curation approaches (Beargrie, 2006), (Charlesworth, 2006) and (Lord et al., 2008). But, there are some other like a manual schema integration (Kokla, 2006), an extensial determination of schema transformation rules (Volz, 2005), a data matching approaches for different data sets (Mustière, 2006) and a semantic integration (Sester et al., 2007).

Unfortunately, any approach for integrating data sources may lead to information loss. Whereas a given producer tends to favor one aspect of the real world, another producer will, perhaps, lend more detail to some other aspect. When both sources are integrated into a unique data set, some detail may be lost in the process.

In our research, we propose delaying the solution of these conflicts by integrating query answers rather than data sources. Let

[*] Supported by Instituto Militar de Engenharia (IME)
[†] Supported by Escola Nacional de Ciências Estatísticas (ENCE)

us assume that a certain aspect of the real world has been modeled by different surveyors resulting in several distinct data sources DB_i (see Fig. 1.b). In practice, if a user queries $Q(DB_i)$ each data sources separately, he or she will obtain answers A_i which may or may not be identical (see Fig. 2). In other words, it is possible to have

$$Q(DB_i) \neq Q(DB_j),\ i \neq j\ \vee\ Q(DB_i) = Q(DB_j),\ i \neq j.$$

If all answers A_i agree with each other, then we must concur that no data integration was needed. Otherwise, we may have different kinds and amounts of discrepancy, which, however, may be resolved in a simpler way. For instance, we may find that most data sources produce identical results whereas a single data source may be regarded as an outlier. It seems reasonable that presenting this duly categorized information to the user will lead to safer decisions being made than simply discarding the outlier, even when it really contains erroneous information. One may easily imagine a scenario where the outlier is correct and all other sources are wrong.

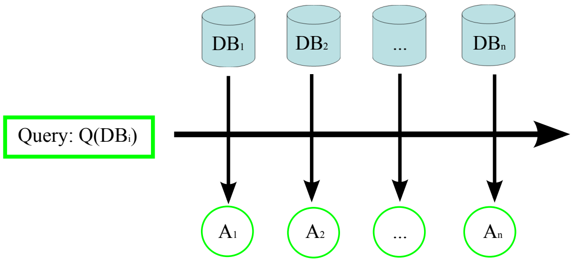

Figure 2: Different answers for different queries

It stands to reason, however, that any process whereby answers must be categorized will require previous knowledge about discrepancies among the sources. In this research we focus on a methodology for analyzing data sources which represent the same set of features in order to establish similarity measures. In particular, we describe an adaptation of the Equivalent Rectangles Method (ERM) (Ferreira da Silva, 1998), a linear discrepancy measure originally proposed for polygonal lines, extended to closed polygons. Furthermore, we use the Cartographic Similarity Index (CSI), an approach for measuring similarity among geographic data sources.

To validate the proposed methods, district boundary databases for the city of Rio de Janeiro, as prepared by two Brazilian institutions, are compared. In this case, databases are represented as a set of closed polygons (not necessarily convex).

The rest of this paper is organized as follows. Section 2 presents the original ERM and shows how it can be adapted to polygons. Section 3 describes the CI, CoI and CSI indexes, and discusses its applicability. Section 4 describes the data sets, methodology used in the experiments, and presents a comparison results. In Section 5, we present our final remarks and suggestions for future work.

2 EQUIVALENT RECTANGLES METHOD (ERM) AND VARIANTS

2.1 Classical *ERM*

The *ERM* methodology was developed to assess the discrepancy between linear representations – polylines, in practice – of the same feature (Ferreira da Silva, 1998). In other words, it tries to measure an average distance between two representations of the same geographic feature. It should be stressed that the methodology can only be used if it is known that both geometric representations are related to same real world feature. So, the *ERM* is very useful in evaluating the quality of data sources.

The approach is based on the well-known formula (Eq. 1)

$$x^2 + S \cdot x + P = 0, \qquad (1)$$

taking into account a "discrepancy polygon" obtained by connecting the initial and final points of the polylines and generating an equivalent rectangle (see Fig. 3). The coefficients assume the values of half the perimeter *P* and area *S* of this discrepancy polygon. Using the formula of Baskara (Eq. 2) two roots for Equation 1 can be determined .

$$\begin{cases} x_1 = \frac{-S+\sqrt{S^2-4\cdot P}}{2} \\ x_2 = \frac{-S-\sqrt{S^2-4\cdot P}}{2} \end{cases} \qquad (2)$$

The absolute value of the first root $|x_1|$ measures an average distance between the representations while the second absolute value $|x_2|$ measures the mean semi-perimeter of the representations (Ferreira da Silva, 1998).

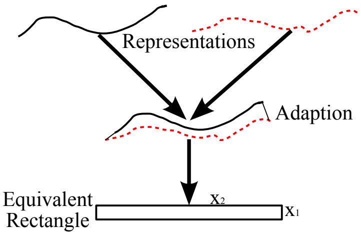

Figure 3: Two line representations and the rectangle used for computing the *ERM*

Incidentally, although the *ERM* has been developed for linear features only, it has been extended to cope with Digital Elevation Models (DEM), having received the name of Equivalent Parallelepiped Method (EPM) (da Rocha Gomes, 2006). In this case, the measure considers the volume, lateral area and perimeter of the generated parallelepiped.

2.2 Polygon *ERM* adaptation

In this work, we propose another extension of the *ERM* so that it can be used for polygonal representations. To obtain this extension, we first observe that a polygon corresponds to a closed polygonal line (see Fig. 4). By analogy with the original *ERM*, a discrepancy polygon can be obtained by computing the difference between the union and the intersection of both polygons. This is then processed in the same way as in the original ERM. Notice that there is no need for joining endpoints.

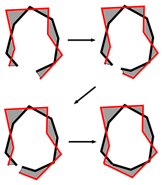

Figure 4: *ERM* adaptation for a pair of polygonal representations

So, is P_i the polygon representing the feature area *A* in the data sources DB_i (Eq. 3). In this case, the coefficient *P*, for the semi-perimeter, and the coefficient *S*, for the area, have value as those obtained by Equations 4 and 5.

$$P_i = DB_i|Polygon(A),\ i = 1, 2 \tag{3}$$

$$P = perimeter((P_1) + perimeter(P_2) \tag{4}$$

$$S = area(P_1 \cup P_2) - area(P_1 \cap P_2) \tag{5}$$

Time complexity of the algorithm for intersection (Žalik, 2000) and union polygonal procedure (Agarwal et al., 2002) is too high. But, it is essential to measure the *CSI* and the *CI* (Sester et al., 2007). Obviously, the quantity of polygon vertices and the number of intersection points are directly related to time complexity of the algorithm. As it was exposed by (Žalik, 2000), an optimal intersection algorithm has a complexity given by $O((k+I) \cdot log_2(k+I))$, where I is the number of intersection points and k is the sum between the number of the input polygon vertices ($k = n + m$). The union operation has a higher complexity. In this case, there are many algorithms, such as, (Agarwal et al., 2002) and (Varadhan and Manocha, 2006). But all of them require non-convex polygons to be decomposed into convex pieces. In this work, we used the algorithm proposed by (Varadhan and Manocha, 2006) to process the union operation and the algorithm proposed by (Žalik, 2000) to produce an intersection polygon.

3 SIMILARITY, COMPLETENESS AND COVERAGE INDICES

When a user considers data from different sources, ambiguities are likely to occur. Measuring the severity of an ambiguity occurrence is not straightforward. Also, it is not clear how to determine the degree of similarity. As a rule, ambiguities may arise in two different scenarios. The first possibility occurs when a single data source has an ambiguous representation. In this case, it is an error of the producer, and a supervised and rigorous inspection on the data source is sufficient to pinpoint this situation and allow it to be corrected. The second case appears when the user has processed data from different producers. This type of ambiguities is a common occurrence because "errors in geographic databases cannot be avoided" (Ali, 2001).

An easy way to identify potentially ambiguous representations is by using metadata. Unfortunately, metadata cannot identify ambiguities in many cases, since it may also be incorrect or ambiguous. A saner approach, then, is to analyze the relevant geometric representations in order to extract information about their similarity. In this work, we use the term **Cartographic Similarity Index** (CSI) to refer to the complement of the *areal distance* (Ali, 2001), a measure originally used to evaluate the "distance" *d* between two sets of polygons. In other words, let P_A and P_B be two polygons, then the relation between CSI and distance *d* is expressed by Equation 6.

$$\begin{aligned} CSI(P_A, P_B) &= 100 \cdot (1 - d(P_A, P_B)) \\ &= 100 - 100 \cdot (1 - \tfrac{area(P_A \cap P_B)}{area(P_A \cup P_B)}) \\ &= 100 \cdot \tfrac{area(P_A \cap P_B)}{area(P_A \cup P_B)}. \end{aligned} \tag{6}$$

Notice that the CSI is expressed as a percentage. Thus, two representations are considered identical ($CSI = 100\%$) if they occupy exactly the same locus. Conversely, two disjoint representations have $CSI = 0\%$.

Another useful measure is the so-called **Completeness Index** (*CI*) – (Ali, 2001) and (Kieler et al., 2007) – which tries to establish how much of a given representation P_A agrees with another representation P_B, and is given by Equation 7.

$$CI(P_A, P_B) = 100 \cdot \frac{area(P_A \cap P_B)}{area(P_A)}. \tag{7}$$

We may also define the **Coverage Index** (*CoI*), expressed by

$$CoI(P_A, P_B) = 100 \cdot \frac{area(P_A)}{area(P_A \cup P_B)}, \tag{8}$$

which can be interpreted as a measure of how much a given representation P_A covers points which may actually belong to a feature, given that this feature is estimated by polygons P_A and P_B.

We notice that measures CI and CoI are not symmetric, i.e., in general,

$$CI(P_A, P_B) \neq CI(P_B, P_A) \wedge CoI(P_A, P_B) \neq CoI(P_B, P_A).$$

Notice also, that the CSI, a symmetric measure is related to CI and CoI by

$$CSI(P_A, P_B) = \frac{CI(P_A, P_B) \cdot CoI(P_A, P_B)}{100}.$$

Although the CI, the CoI and the CSI were presented as pairwise operators, they can easily be generalized as n-way operators:

$$\begin{aligned} CI(P_A, \ldots, P_n) &= 100 \cdot \frac{area(P_A \cap \ldots \cap P_n)}{area(P_A)} \\ CoI(P_A, \ldots, P_n) &= 100 \cdot \frac{area(P_A)}{area(P_A \cup \ldots \cup P_n)} \\ CSI(P_A, \ldots, P_n) &= 100 \cdot \frac{area(P_A \cap \ldots \cap P_n)}{area(P_A \cup \ldots \cup P_n)} \end{aligned}$$

4 EXPERIMENTS

In order to investigate the usefulness of the *Adapted ERM* and the *CSI*, a prototype system was used to compare two data sources for the district partitioning of the city of Rio de Janeiro. The prototype exhibits both data sources graphically, thus allowing a visual inspection of ambiguities. It also computes the *Adapted ERM* and *CSI* values for the different polygons. In this case, each polygon represents one of the 159 districts of the city of Rio de Janeiro. The data was obtained from two sources in the same scale ($1 : 10.000$): **Pereira Passos Institute** (IPP in Portuguese), a municipal institution responsible for mapping the city, and the **Brazilian Institute of Geography and Statistics** (IBGE in Portuguese), an entity responsible for the systematic mapping of the country. In fact, the two data sources are, visually, quite similar, but not identical (see Fig. 5).

The data was, initially, acquired in *shapefile* format (ESRI, 1998), but was converted to Geography Markup Language (GML) format (OGC, 2001) using GDAL tools (GDAL, 2008). All subsequent processing was made in *GML* format.

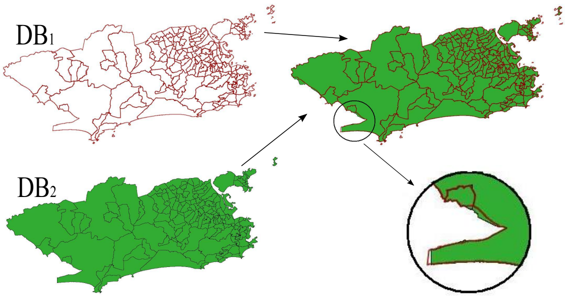

Figure 5: Districts ambiguities

In general, we are interested in performing a procedure to establish matching representations of the same features according to two data sources, say, DB_1 and DB_2. For simplicity, we assume that a data source is comprised solely of two columns, one for the geometric data, and another for the non-geometric information which identifies the table row , which we call "feature name" (see Table 1).

feature name	geometric data
district D_1	list of coordinates
district D_2	list of coordinates
...	...
district D_n	list of coordinates

Table 1: Data source example

The detection of matches consisted of checking all possible pairings between polygons (features) of both data sources. For each pair (P_i, R_j), where P_i is a polygon from the IPP data source and R_j is a polygon from the IBGE data source, both the *Adapted ERM* and the *CSI* were computed. It should be noted that both sets have the same cardinality, but this needs not be the case in general.

The intention was to evaluate the occurence of *matches* between two specific representations by analyzing index values. So, let $ERM_{min}(P_i)$ denote the minimum value for the Adapted ERM among all pairs (P_i, R_j). Then, R_k is considered the candidate match for P_i if $ERM(P_i, R_k) = ERM_{min}(P_i)$. Similarly, let CSI_{max} denote the maximum value for the CSI among all possible pairs (P_i, R_j). Then, R_k is considered the candidate match for P_i if $CSI(P_i, R_k) = CSI_{max}(P_i)$. Notice that the matching functions are not symmetric, i.e., R_j being considered the candidate match for P_i does not imply that P_i is considered a candidate match for R_j.

Within this framework, it is reasonable to suppose that any given feature is represented in both data sources, i.e., there is a multiple representation. One may even call these representations "ambiguous", in the sense that a feature has, thus, two representations. This benign occurrence corresponds to the case where the candidate match (using either index) for P_i is R_j and vice-versa.

Another important consideration is the match between feature names. What happens if a match detected geometrically does not concur with their respective feature names? Conversely, what does it mean to have identical feature names associated with non-matching geometric representations? Clearly, a *true* match must only be considered if geometric representations match each other (according to both index metrics), *and* their feature names also agree. This is expressed in Equation 9, where $FN(x)$ stands for the feature name for polygon x:

$$\begin{aligned} P_i\ matches\ R_j \Leftrightarrow ERM(P_i, R_j) = ERM_{min}(P_i) \wedge \\ CSI(P_i, R_j) = CSI_{max}(P_i) \wedge \\ FN(P_i) = FN(R_j). \end{aligned} \tag{9}$$

4.1 Matching problem

After processing the district data sources, the candidate matches obtained using both indices were exactly the same. In other words, the *Adapted ERM* and the *CSI*, produce the same result. However, the use of feature names reveal that only 158 of the 159 matches were "true" according to Eq. 9. In particular, only one district was not identified correctly. For both data sources, the candidate match for the district named "Parque Columbia" was another district named "Pavuna". In other words, let P_c denote a polygon in the first data source for which $FN(P_c)$ is "Parque Columbia", and P_p denote a polygon for which $FN(P_p)$ is "Pavuna". Let R_c and R_p analogously denote the polygons of "Parque Columbia" and "Pavuna" in the second data source. Then, it was found that $ERM_(P_c, R_p) = ERM_{min}(P_c)$, and, similarly, $CSI_(P_c, R_p) = CSI_{max}(P_c)$. Notice that the district of "Pavuna" was correctly matched, i.e., $ERM(P_p, R_p) = ERM_{min}(P_p)$ and $CSI(P_p, R_p) = CSI_{max}(P_p)$.

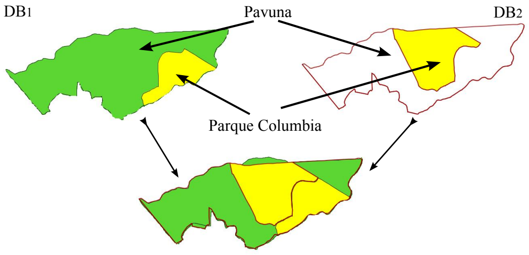

Figure 6: An indefinition example – "Parque Columbia"

In that case, there are indefinitions about the boundaries of the districts. As it is shown in Fig. 6, IPP and IBGE do not agree about the geographic position of "Parque Columbia". These specific districts return the values shown in Table 2.

Correlation	*Adapted ERM (m)*	*CSI (%)*
$P_p \times R_p$	232.50	54.16
$P_p \times P_c$	708.22	0.00
$P_p \times R_c$	415.08	21.32
$R_p \times P_c$	457.20	31.60
$R_p \times R_c$	859.13	0.00
$P_c \times R_c$	680.19	0.00

Table 2: "Pavuna" and "Parque Columbia" comparison

4.2 ERM Analysis

We notice that the values obtained with the ERM_{min} index were fairly high both for the district of "Pavuna" and for the district of "Parque Columbia", but generally low for the other districts, rarely surpassing $40m$, as shown in Table 3.

Incidentally, Brazilian law tries to establish standards to assess the quality of the systematic mapping of the country, called the Cartographic Accuracy Standard (Brasil, 1984). In this case, the standard prescribes that a class A map should have 95% of field samples lying within $5m$ of the corresponding map feature in mapping scale. Thus, using the ERM index, it is possible to affirm that at least one data source used in the experiments would not pass said standard.

ERM_{min} range (m)	number of districts
$0 \leq 10$	2
$10 \leq 20$	67
$20 \leq 30$	60
$30 \leq 40$	18
$40 \leq 50$	4
$50 \leq 60$	4
$60 \leq 80$	0
$80 \leq 100$	2
$100 \leq 250$	1
$250 \leq 500$	1

Table 3: ERM range analysis

Another curious aspect of the *ERM* index is that it sometimes yields non-intuitive similarities. For instance, the match for the district of “Oswaldo Cruz” yields

$$ERM_{min}(P_o) = ERM_{min}(R_o) = ERM(P_o, R_o) = 9.10m,$$

the lowest among all ERM values. Looking, however, at the next best candidates for matching that district, i.e., the next 5 lowest values of $ERM(P_o, \ldots)$, we do not find neighboring districts as can be seen on Figure 7 and Table 4, instead of the highest CSI values, as can be seen in Table 5.

$FN(R_j)$	$ERM(P_o, R_j)$	$CSI(P_o, R_j)$
Oswaldo Cruz	9.10	97.00
Cosme Velho	421.70	0.0
Santa Teresa	431.75	0.0
Paquetá	437.49	0.0
Urca	447.29	0.0

Table 4: Lowest ERM values for matching P_o, the district of “Oswaldo Cruz”

$FN(R_j)$	$ERM(P_o, R_j)$	$CSI(P_o, R_j)$
Oswaldo Cruz	9.10	97.00
Bento Ribeiro	728.15	0.19
Madureira	784.79	0.02
Turiaçú	597.36	0.01
Campinho	575.03	0.01

Table 5: Highest CSI values for matching P_o, the district of “Oswaldo Cruz”

4.3 CSI Analysis

It is also useful to look at the highest value for the CSI among all pairs, which corresponds to the district of “Bangu”. A table with the next highest CSI values for that district are shown in Table 6. As can be seen in Figure 8, these correspond to districts neighboring “Bangu”, as expected.

We also ranked the matches obtained with the CSI in increasing order, as shown in Table 8. As expected, the two lowest values are associated with the districts of “Parque Columbia” and “Pavuna”, whereas all other districts yielded CSI values bigger than 70%. Thus, a cut-off value of 70% would be enough to pinpoint matching problems, even in the absence of feature name information.

In Figure 9, it is possible to observe the *CSI* distribution with respect to geographic locations. Districts with *CSI* lower than 70% are painted in red, districts between 70% and 90% in yellow, and above 90% in green. We notice that smaller values usually correspond to districts with smaller areas. This is understandable since errors are more likely to occur on district boundaries.

Our tests indicate that the *Adapted ERM* and *CSI* tend to detect the same matches. However, the *Adapted ERM* is not as sensitive as the *CSI*. The former produces a large dispersion in the results when compared to the latter, as shown by Table 3 and 8. Thus, the identification of ambiguities is probably easier when the *CSI* is used. The advantage of the *Adapted ERM* lies on its yielding measures in distance units. On the other hand, the *CSI* is more adequate for quantifying similarity.

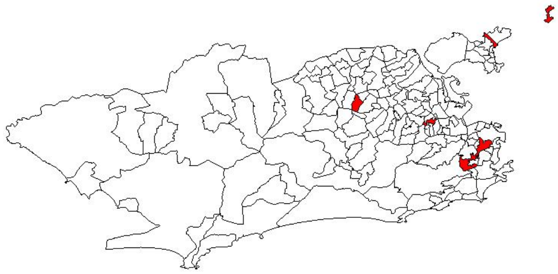

Figure 7: Districts yielding the 5 lowest values of ERM with respect to the “Oswaldo Cruz” district

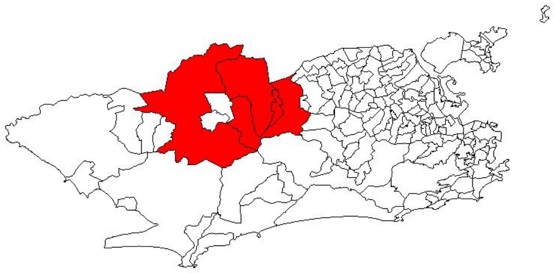

Figure 8: Districts yielding the 5 highest values of CSI with respect to the “Bangu” district

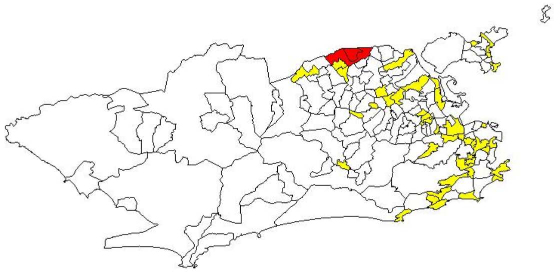

Figure 9: Similarity distribution on Rio de Janeiro city

5 CONCLUSIONS AND FUTURE WORK

This work is part of a doctoral thesis that proposes a methodology to enable an user to obtain any information from a query to multiple data sources. The next step of the research will be to use the *CSI* as a qualifier of ambiguities and facilitate the integration of responses. The main idea is to shy away from *a priori* integration of data sources in favor of an *a posteriori* treatment of answers obtained by querying these data sources separately. Thus, given a *query* applied to any multiple representation, it is necessary to process the multiple responses in order to provide support for decisions. This, also, helps quantifying the certainty, coverage and completeness of the query answers.

To reach this goal, this work proposed, initially, an extension of *ERM*, and then proposed a new use for a known index, the *CSI*. The idea was to seek a way to identify possible ambiguities, in order to facilitate a further integration of responses. It can also

$FN(R_j)$	$ERM(P_b,R_j)$	$CSI(P_b,R_j)$
Bangu	20.98	98.32
Padre Miguel	2081.72	0.24
Campo Grande	2789.99	0.11
Senador Camará	2273.43	0.09
Realengo	2292.59	0.03

Table 6: Highest CSI values for matching P_b, the district of “Bangu”

$FN(R_j)$	$ERM(P_o,R_j)$	$CSI(P_o,R_j)$
Bangu	20.98	98.32
Santa Teresa	1690.78	0.0
Barra de Guaratiba	1907.89	0.0
Cidade Universitária	1913.72	0.0
Centro	1934.76	0.0

Table 7: Lowest ERM values for matching P_o, the district of “Bangu”

be observed that the proposed index serves as a certifier of geographic data to be used in digital curation. Identifying ambiguous representations and offering them a value of similarity is essential to obtain the largest possible amount of information. It is our belief that this approach will help making ready use of *web* data sources without incurring the costly effort of integrating them in a single database.

CSI_{max} range	number of districts
$0 \leq 70$	2
$70 \leq 80$	6
$80 \leq 90$	36
$90 \leq 95$	72
$95 \leq 100$	43

Table 8: CSI range analysis

The admittedly small experimental evidence shown in this paper indicates that the *CSI* is more sensitive to the identification of possible ambiguities than the *Adapted ERM*. Nevertheless, the latter, being able to return distances rather than correlations, may be of use in queries involving metric reasoning.

6 ACKNOWLEDGMENTS

The authors wish to thank the anonymous reviewers for many helpful comments and suggestions. Thanks are also due to CNPq, ENCE, IBGE, IME and IPP by financial and academic supports to this research.

REFERENCES

Agarwal, P. K., Flato, E. and Halperin, D., 2002. Polygon decomposition for efficient construction of minkowski sums. Computational Geometry 21, pp. 39 – 61.

Ali, A. B. H., 2001. Positional and shape quality of areal entities in geographic databases: quality information aggregation versus measures classification. In: ECSQARU´2001 Workshop on Spatio-Temporal Reasoning and Geographic Information Systems, Toulouse, France.

Beargrie, N., 2006. Digital curation for science, digital libraries, and individuals. International Journal of Digital Curation.

Brasil, 1984. Decreto n^o 89,817, de 20 de junho de 1984. estabelece as instruções reguladoras de normas técnicas da cartografia nacional. Diário Oficial da República Federativa do Brasil.

Charlesworth, A., 2006. Digital curation: Copyright and academic research. International Journal of Digital Curation.

da Rocha Gomes, F. R., 2006. Avaliação de discrepâncias entre superfícies no espaço tridimensional. Master’s thesis, Instituto Militar de Engenharia, Rio de Janeiro, Brazil.

de Spinoza, B., 2005. Ética demonstrada à maneira dos geômetras. first edn, Martin Claret, São Paulo, Brazil.

ESRI, 1998. Esri shapefile technical description: An esri white paper. http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf.

Ferreira da Silva, L. F. C., 1998. Avaliação e integração de bases cartográficas para cartas eletrônicas de navegação terrestre. PhD thesis, Escola Politécnica da Universidade de São Paulo, São Paulo, Brazil.

GDAL, 2008. Geographic data abstraction library. http://www.gdal.org/.

Goodchild, M. F., Fu, P. and Rich, P., 2007. Sharing geographic information: An assessment of the geospatial one-stop. Annals of the Association of American Geographers 97(2), pp. 250 – 266.

Kieler, B., Sester, M., Wang, H. and Jiang, J., 2007. Semantic data integration: data of similar and different scales. Geoinformation.

Kokla, M., 2006. Guidelines on geographic ontology integration. In: Proceedings of the ISPRS Technical Commission II Symposium, Vol. 36.

Lord, P., Macdonald, A., Lyon, L. and Giaretta, D., 2008. From data deluge to data curation. International Journal of Digital Curation.

Mustière, S., 2006. Results of experiments on automated matching of networks at different scales. In: ISPRS - Workshop on Multiple Representation and Interoperability of Spatial Data, pp. 92 – 100.

OGC, 2001. Geography markup language. http://www.opengis.net/gml/01-029/GML2.html.

Pazinato, E., Baptista, C. and Miranda, R., 2002. Geolocalizador: Sistema de referência espaço-temporal indireta utilizando um sgbd objeto-relacional. SBC Geoinfo ´02: Anais do IV Simpósio Brasileiro de GeoInformática pp. 49 – 56.

Sester, M., von Gösseln, G. and Kieler, B., 2007. Identification and adjustment of corresponding objects in data sets of different origin. In: 10th AGILE International Conference on Geographic Information Science, Aalborg University.

Varadhan, G. and Manocha, D., 2006. Accurate minkowski sum approximation of polyhedral models. Graphical Models 68(4), pp. 343 – 355. PG2004.

Volz, S., 2005. Data-driven matching of geospatial schemas. In: 6th COSIT International Conference, pp. 115 – 132.

Žalik, B., 2000. Two efficient algorithms for determining intersection points between simple polygons. Computer and Geosciences 26(2), pp. 137 – 151.