

A NEW GEOGRAPHIC CONTEXT MEASURE TO SIMILARITY ASSESSMENT BASED ON THE SHAPE CONTEXT DESCRIPTOR

E. M. A. Xavier ^{1*}, M. A. Ureña-Cámara ², F. J. Ariza-López ²

¹ Brazilian Army Geographic Service, QGEx, 70630-901 Brasília, Brazil – emerson.xavier@eb.mil.br

² Dept. of Cartographic Engineering, Universidad de Jaén, 23071 Jaén, Spain – (maurena, fjariza)@ujaen.es

Commission IV, WG IV/7

KEY WORDS: similarity, matching, conflation, data integration, geographic context

ABSTRACT:

Nowadays there are many geospatial datasets available for the same area. This large availability is derivative from the advances in remote sensing processes, which includes the popularization of drones and the increasing number of satellite platforms. These data are built by distinct producers, with different requirements. It is fair presume that a more complete version of these datasets can be created using data integration techniques. Among them we can find the map conflation methods, in which the first phase often begins with an alignment between the datasets assessed. The procedure of find these correspondences between geospatial datasets is called matching. In this study we present a new geographic context measure that can be used to implement a new matching method at the feature level. This new measure is based on the shape context descriptor proposed by Belongie. The experiments showed that our approach is a feasible solution, which is less sensible to data disturbance than other traditional methods.

1. INTRODUCTION

Recent developments in geospatial data science have increased the availability of data sources. Different providers lead to different needs, and their acquisition rules, designed to fit these purposes. So it is straight suppose that a “best” data can be obtained by the fusion, or conflation, of distinct geospatial datasets. The first phase of map conflation is looking for the correspondences between the datasets (Ruiz et al., 2011), a process known as matching. Recent studies published show that this topic has raised open issues for the geospatial data science community (Wang et al., 2019, Ruiz-Lendínez et al., 2019).

Some authors pointed out that the similarity measures play an important role in matching procedures (Li, Goodchild, 2011). Among the many similarity measures that can be applied for feature matching, there are the context measures, which assess the geographic context of a feature when compared to its neighbourhood. There are few approaches available in the literature focused in feature matching using context measures (Samal et al., 2004, Kim et al., 2010, Zhang et al., 2014). However, the first two studies are based on the selection of landmarks, which can be difficult to determine in a fully-automated system. The third study is based on the Delaunay triangulation, which can be limited when there are many objects nearby.

In this sense, our research question arise from these issues: how can we assess the context similarity of geospatial features? Our hypothesis is that a context measure based on the shape context descriptor proposed by Belongie et al. (2002) configures a solution to effectively assess the geographic context. Therefore, the aim of this study is to develop a new geographic context measure for feature matching based on the shape context designed for object recognition.

In this study we develop our proposal first presented by Xavier (2017). We hope that this new measure can be used to improve

the precision and recall of a matching method, by providing a new similarity aspect to be taken into account when evaluating the correspondences between features. We hope that it could be used jointly with the Euclidean distance, widely-used in many point-based methods (Beerli et al., 2004, Mustière, Devogele, 2008, and McKenzie et al., 2014)

This paper is structured as follows. Section 2 provides a brief background about context measures for matching methods. Section 3 presents the new proposed measure. Section 4 provides our experiments and discussions. Lastly, section 5 brings some conclusions and future work.

2. BACKGROUND

Matching methods for geospatial data requires similarity measures in order to evaluate whether two geospatial datasets are similar or not. Similarity measures define some objective measurements for eliminating, or at least mitigating, the uncertainty inherent to this process. Xavier et al. (2016) organizes the matching measures according to the nature of the measured quantity: geometry, topology, attributes, context, and semantics.

Context measures are used to quantify the geographic context of features, which permits comparing their similarity. According to Samal et al. (2004), “geographic context refers to the spatial relationships between objects in an area”, notably the relationships between an object and a limited set of landmarks. Some authors argued that geographic context has the potential to play an important role in the matching scenario. In this sense, Zhang et al. (2014) affirm that the similarity of geographic features depends on the context, for some ambiguous cases.

To the best of the authors’ knowledge, the first study regarding geographic context measures is that of Samal et al. (2004). The authors applied landmarks to similarity assessment. By

* Corresponding author

combining many metrics, such as positional and attribute, the authors also proposed the use of landmarks to build a proximity graph in order to compare similarity between features. The proximity graph is a weighted directed graph defined to assess geographic context similarity considering the “proximity” relative to some pre-selected landmarks. The similarity is measured using the total vector offset of the corresponding objects in both datasets.

Kim et al. (2010) extended the previous context approach by using Voronoi diagrams and triangulation geometry. In this method, some landmarks are used to determine the “neighbourhood” of a point feature. Comparing with Samal’s method, in the final analysis the two approaches differ. While Samal et al. (2004) use the total offset vector to quantify the context similarity, Kim et al. (2010) use the area/perimeter ratio in the assembled triangles.

Other interesting study was developed by Zhang et al. (2014). The authors, inspired by the k-nearest approach of Zheng and Doermann (2006), proposed the use of Delaunay triangulation to define the neighbourhood of objects, considering a continuous influence from the closest objects.

3. GEOGRAPHIC CONTEXT MEASURE

As shown in the last section, there are few context approaches available in the literature in order to match point features. The studies of Samal et al. (2004) and Kim et al. (2010) are based on the selection of landmarks, and the study of Zhang et al. (2014) is based on the Delaunay triangulation.

We developed a context distance based on the shape context descriptor developed by Belongie et al. (2002). Despite the original study having used the shape context to find correspondences between shapes, we adapted those concepts in order to quantify the geographic context of point features by using their relative positions against the closer objects.

The context distance is computed as follows. For each point in a dataset we compute a common histogram with the polar coordinates of all other close points relative to the assessed point. Figure 1(a) illustrates how the histogram bins are positioned in the space considering the assessed point in the center. It is possible to note that not all points are used, just those inside a neighbourhood limit, or r_{max} . So for each bin we count the number of point features and fill the respective value. After counting all neighbours inside the limit, we have the shape context for the assessed point, which is the context “signature” for that feature.

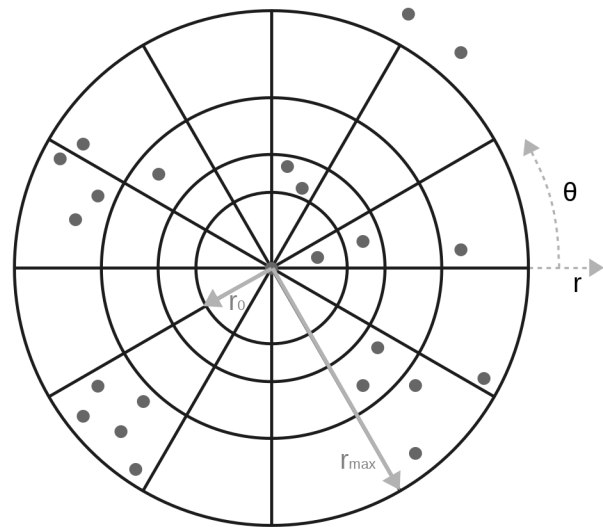
The position of the histogram bins differ from the original Belongie’s study, where the authors considered a log-polar coordinate system (Belongie, Malik 2000, Belongie et al., 2002). In our study we propose that the increment of radius r should occur by the length of the previous arc. So it works as a geometric progression that begins with an r_0 and has a common ratio equal to the angular step in radians.

We adopted the cost function proposed by Belongie et al. (2002) as the context distance. This distance measures the similarity between histograms, i.e. between shape contexts, using a χ^2 test statistic for normalized histograms. This distance assumes values from zero (completely similar) to one (dissimilar).

$$C(p_i, q_j) = \frac{1}{2} \sum_{k=1}^K \frac{[h_i(k) - h_j(k)]^2}{h_i(k) + h_j(k)} \quad (1)$$

where C represents the cost function that measures the context similarity between points p_i and q_j , and $h_i(k)$ and $h_j(k)$ represent the K -th normalized histogram bin for points p_i and q_j , respectively.

(a) Diagram of histogram bins



(b) Resulting histogram

1	0	0	0	1	3	0	5	0	0	2	1
0	0	0	0	1	0	0	0	0	0	2	0
1	0	2	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0

Figure 1. Geographic context measure. (a) Diagram of histogram bins for the point in the center. (b) Resulting histogram representing the number of points in each bin. Source: based on Xavier (2017).

4. EXPERIMENTS AND DISCUSSION

In order to validate our proposal of a context measure for matching geospatial features, we developed a simple matching method which uses the new measure, and tried it using a testbed available to the scientific community, called MatchingLand (Xavier et al., 2017).

This experiment has four essays. The first essay (P1) aims to check the performance of the new context distance. The other essays compare the performance of the new proposed measure with the widely-used Euclidean distance in three configurations: normal conditions (essay P2), in presence of systematic disturbance (P3), and with random disturbance (P4).

This matching method can be defined using three elements: the measure, its associated thresholds, and the selection criteria. This experiment uses two measures: the new context measure

and the Euclidean distance. Each measure has its own thresholds: in context measure is a value for the cost function, and for Euclidean distance it is in meters. The last element that compounds the matching method is the selection criteria. This element defines whether two assessed features should be marked as a “match” or not. In this study we consider that a selection criteria is applied only inside the given thresholds, and we adopt two criteria: both nearest, and closer.

In the *both nearest* criterion two features in different datasets (e.g. datasets A and B) are marked as a match when the maximum similarity (or minimum distance) occurs in the forward matching (from dataset A to dataset B), and in the backward matching (from dataset B to dataset A), so it allows only 1:1 matches. The selection criteria *closer* marks two features as a match when the maximum similarity (or minimum distance) occurs in the forward matching or in the backward matching. So, this approach permits m:n corresponding cases.

In this experiment we assess three variables for matching geospatial features: precision, recall, and F-measure. Precision and recall are concepts that come from the information retrieval field (Van Rijsbergen, 1979). Precision evaluates the presence of wrong matches (false positives) against the real matches (true positives); while recall evaluates the presence of non-matches (false negatives) against those real matches. The F-measure represents the harmonic mean between precision and recall (Do, Rahm, 2002).

4.1 Geographic context measure

In our first essay (P1) we investigate the geographic context distance as similarity measure for point features. We combined a set of thresholds (0.3, 0.5, 0.9) and the two criteria: both nearest and closer. Table 1 summarizes the results for this essay.

Variable	Both nearest			Closer		
	0.3	0.5	0.9	0.3	0.5	0.9
precision	54.6	54.7	62.6	18.7	16.3	14.5
recall	46.4	49.8	53.3	78.3	86.0	95.3
F-measure	49.6	52.0	56.9	25.0	24.0	22.7

Notes: average values for 9 regions in each combination of parameters.

Table 1. Results for the essay P1 (geographic context).

The both nearest criterion with the largest threshold (0.9) presented the best results, with an average F-measure around 56.9%. The closer criteria, which supports m:n cases, presented a worse F-measure than the both nearest criteria in all thresholds. Besides it reaches the best recall (95.3%) the global result was influenced by the low precision.

The configuration with best results, 0.9 threshold and both nearest criterion, will be used in the following essays.

4.2 Comparing the context measure with Euclidean distance

In this essay we compare the performance of the context measure with the Euclidean distance. There are three configurations: Euclidean distance with the closer criterion and threshold of 10 m (10C); Euclidean distance with the both nearest criterion and threshold of 25 m (25B); and the geographic context measure with the both nearest criterion and threshold of 0.9 (CTT). Table 2 summarizes the results for this essay, presenting the average values for 9 regions.

Variable	10C	25B	CTT
precision	97.8	95.7	62.6
recall	98.4	97.8	53.3
F-measure	98.1	96.7	56.9

Notes: 10C means Euclidean distance, closer criterion, 10 m threshold; 25B means Euclidean distance, both nearest, 25 m; CTT means geographic context measure, both nearest, 0.9 threshold.

Table 2. Results for the essay P2 (comparison geographic context vs Euclidean).

Table 2 shows that the Euclidean distance reached the best results, with an average F-measure beyond 96%.

4.3 Systematic disturbance

The next essay refers to systematic perturbations applied over point data (P3). In this essay we used two configurations from the previous essay: 10C and CTT, but we changed the original data for a disturbed version (see Xavier et al., 2017).

These systematic disturbances include 32 different translations, 24 rotations, and 8 scalings. In this essay, we “freezed” the other systematic disturbances and assessed the performance of matching methods using just one kind of disturbance: translation, rotation, or scaling. Table 3 summarizes the results for this essay.

variable	translation		rotation		scaling	
	10C	CTT	10C	CTT	10C	CTT
precision	34.0	62.5	76.4	62.5	71.4	62.8
recall	26.5	52.7	58.5	52.7	53.4	52.9
F-measure	27.8	56.4	62.4	56.5	56.9	56.7

Notes: 10C means Euclidean distance, closer criterion, 10 m threshold; CTT means geographic context measure, both nearest, 0.9 threshold.

Table 3. Results for the essay P3 (systematic disturbance).

The results represent an average value of 288 measured values for translations, 216 for rotation, and 72 for scaling. Table 3 shows that the results with the Euclidean distance were largely influenced by the systematic disturbances, notably in the translations. The geographic context measure presented stability in presence of this kind of disturbance.

4.4 Random disturbance

The last essay refers to assess the performance of matching methods over the presence of random disturbance in the datasets (essay P4). We used the same two configurations from the previous essays: 10C and CTT.

These random disturbances include 100 iterations over three distinct standard displacements: 5 m, 25 m, and 50 m. The testbed includes a fourth value: 12.5 (Xavier et al., 2017), but it was not used here in order to simplify the analysis. Henceforth the two configurations (10C and CTT, Euclidean and context measure) were tested against 900 disturbed datasets (9 regions x 100 iterations) for each standard displacement (5, 25, and 50 meters). Table 4 summarizes the results for this essay.

variable	5 m		25 m		50 m	
	10C	CTT	10C	CTT	10C	CTT
precision	97.2	62.6	65.2	62.4	32.3	61.4
recall	97.6	52.8	33.1	52.1	9.0	50.6
F-measure	97.4	56.5	42.1	56.1	13.2	54.8

Notes: 10C means Euclidean distance, closer criterion, 10 m threshold; CTT means geographic context measure, both nearest, 0.9 threshold.

Table 4. Results for the essay P4 (random disturbance).

Table 4 shows a similar result from that obtained with the systematic disturbance in the previous essay (P3). The overall performance of the Euclidean distance was getting worse with an increasing displacement. Figure 2 illustrates these results by presenting the variations in the average F-measure relative to the random disturbance for each standard displacement applied.

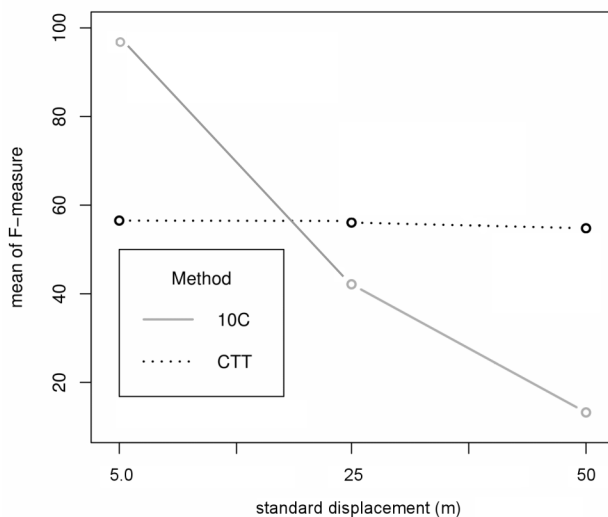


Figure 2. Average F-measure for each matching method in function of standard displacement for random disturbs.

4.5 Discussions

The experiment revealed that the geographic context measure is a feasible option when choosing a similarity measure for matching geospatial point data.

Besides those methods based on Euclidean distance achieved the best performance (precision, recall, F-measure) when compared with the geographic context measure, these methods were shown to be more sensitive to systematic or random disturbances. On the other hand, the geographic context measure showed off less influenced by these kinds of disturbance.

If there is the possibility of large displacements between point datasets, we recommend using the geographic context measure, both nearest criterion, threshold equal to 0.5, prior to any Euclidean matching. This way, it is possible to define some “control points” in order to determine whether the Euclidean distance could be applied.

Other possibility is to using these control points to perform a register between assessed datasets in order to get down the

displacement between datasets. Once the displacement is reduced, some Euclidean-based method could be applied.

5. CONCLUSIONS

The integration of distinct geospatial data has the potential to improve datasets by aggregating new information to those datasets. The geographic key for this integration requires the perfect alignment between the considered datasets, in a process called geospatial data matching. In this paper we have presented a new geographic context measure that can be used to match point datasets.

The proposed solution is based on the shape context descriptor proposed by Belongie et al. (2002). The results obtained in the experiment showed the feasibility of this approach to matching point features.

This approach presents interesting features. The geographic context measure proved to be less sensible to data disturbance (systematic or random) than other measures. We believe that this new geographic context measure can be performed as a first phase of a matching method, in which it could be applied to an initial align between the assessed datasets. Since its main characteristic is taking into account the feature context, it can be used if there are large displacements, rotations, or translations between these datasets. This measure could also be used in other research issues, beyond the geospatial data matching. As a basic measure, which assess the context signature of features, it could even be used in data quality studies.

The proposed solution has some limitations: it is sensible to parameter setting. If a small neighbour limit (a search box) is used, it may reduce the effectiveness of this procedure for feature matching, since many objects possibly cannot be reached.

Future work includes: (1) adapt this measure for internal matching; (2) investigate how decrease the dependence of parameter setting; and (3) publish an implementation of this method as a plug-in for the QGIS software, called Matching_Box¹.

ACKNOWLEDGEMENTS

This work has been funded by the Ministry of Science and Technology (Spain) and the European Regional Development Fund under grant no. BIA2011-23217. The authors also wish to acknowledge the support of the Brazilian Army's Department of Science and Technology, and the support of the Regional Government of Andalucía (Spain) for the Research Group “Ingeniería Cartográfica” (TEP-164).

REFERENCES

- Beeri, C., Kanza, Y., Safra, E., Sagiv, Y., 2004. Object fusion in geographic information systems. *Proceedings of the 30th VLDB Conference*, 16–827. Toronto, Canada.
- Belongie, S., Malik, J., 2000. Matching with shape contexts. *IEEE Workshop on Content-Based Access of Image and Video Libraries*, CBAIVL 2000, 20–26. Hilton Head Island: IEEE. doi.org/10.1109/IVL.2000.853834.

¹ https://github.com/emxav/matching_box

- Belongie, S., Malik, J., Puzicha, J., 2002. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4), 509–522. doi.org/10.1109/34.993558.
- Do, H.-H., Rahm, E., 2002. COMA - A system for flexible combination of schema matching approaches. *28th International Conference on Very Large Databases*. VLBD. Hong Kong, China.
- Kim, J. O., Yu, K., Heo, J., Lee, W. H., 2010. A new method for matching objects in two different geospatial datasets based on the geographic context. *Computers & Geosciences*, 36(9), 1115–1122. doi.org/10.1016/j.cageo.2010.04.003.
- Li, L., Goodchild, M. F., 2011. An optimisation model for linear feature matching in geographical data conflation. *International Journal of Image and Data Fusion*, 2(4), 309–328. doi.org/10.1080/19479832.2011.577458.
- McKenzie, G., Janowicz, K., Adams, B., 2014. A weighted multi-attribute method for matching user-generated points of interest. *Cartography and Geographic Information Science*, 41(2), 125–137. doi.org/10.1080/15230406.2014.880327.
- Mustière, S., Devogele, T., 2008. Matching networks with different levels of detail. *GeoInformatica*, 12(4), 435–453. doi.org/10.1007/s10707-007-0040-1.
- Ruiz, J. J., Ariza, F. J., Ureña, M. A., Blázquez, E. B., 2011. Digital map conflation: a review of the process and a proposal for classification. *International Journal of Geographical Information Science*, 25(9), 1439–1466. doi.org/10.1080/13658816.2010.519707.
- Ruiz-Lendínez, J. J., Maćkiewicz, B., Motek, P., Strykiewicz, T., 2019. Method for an automatic alignment of imagery and vector data applied to cadastral information in Poland. *Survey Review*, 51(365), 123-134, doi.org/10.1080/00396265.2017.1388959.
- Samal, A., Seth, S., Cueto, K., 2004. A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science*, 18(5), 459–489. doi.org/10.1080/13658810410001658076.
- van Rijsbergen, C. J., 1979. *Information retrieval* (2nd ed). Newton: Butterworth-Heinemann.
- Wang, S., Guo, Q., Xu, X., Xie, Y., 2019. A Study on a Matching Algorithm for Urban Underground Pipelines. *ISPRS International Journal of Geo-Information*, 8(8), 352. doi.org/10.3390/ijgi8080352.
- Xavier, E. M. A., Ariza-López, F. J., Ureña-Cámara, M. A., 2016. A survey of measures and methods for matching geospatial vector datasets. *ACM Computing Surveys*, 49(2), Article 39. doi.org/10.1145/2963147.
- Xavier, E. M. A., 2017. *Automatic evaluation of geospatial data quality using web services*. Universidad de Jaén, Jaén, Spain.
- Xavier, E. M. A., Ariza-lópez, F. J., Ureña-cámara, M. A., 2017. MatchingLand, geospatial data testbed for the assessment of matching methods. *Scientific Data*, 4, 170180. doi.org/10.1038/sdata.2017.180.
- Zhang, X., Ai, T., Stoter, J., Zhao, X., 2014. Data matching of building polygons at multiple map scales improved by contextual information and relaxation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 92, 147–163. doi.org/10.1016/j.isprsjprs.2014.03.010.
- Zheng, Y., Doermann, D., 2006. Robust point matching for nonrigid shapes by preserving local neighborhood structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4), 643–649. doi.org/10.1109/TPAMI.2006.81.