Models for Half-Direction based Part-Whole Relationships

Gaurav Singh
Dept. of Geo-Information Processing,
Faculty ITC, University of Twente,
Hengelosestraat 99
7514 AE Enschede, The Netherlands
singh09721@itc.nl

Rolf A. de By
Dept. of Geo-Information Processing,
Faculty ITC, University of Twente,
Hengelosestraat 99
7514 AE Enschede, The Netherlands
deby@itc.nl

ABSTRACT

We present a conceptual framework for interpreting text phrases such as "in central northern Bahia" and "in northern central Bahia" as spatial element of geographic information retrieved from text. Our approach allows spatial computations with such phrases, leading to deeper understanding of places and human spatial cognition associated with them. We develop a number of interpretation models and their placement based on different notions of centre of the reference region. We evaluate these models for the performance characteristics of precision and recall, against an Ornithological gazetteer of Brazil, and draw conclusions on the cognition of, and computation with half-direction based part-whole relations.

Categories and Subject Descriptors

H.2.8 [**DATABASE MANAGEMENT**]: Database Applications—*Spatial databases and GIS*; I.6.5 [**SIMULATION AND MODELING**]: Model Development—*Modeling methodologies*; F.2.2 [**ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY**]: Nonnumerical Algorithms and Problems—*Geometrical problems and computations*

General Terms

Design, Experimentation, Performance

Keywords

GIS, part-whole relationships, spatial models, geostreaming, direction, VGI

1. INTRODUCTION

The advances in communication technology (esp. internet and telecom) have paved the way for people to communicate anytime and anywhere. The opportunity to communicate over text messages or through online social media such as twitter, blogs or facebook allow, for instance, to share travel experiences about places visited. Phrases such as *'last month, I camped in Paraná'*, or *'hiking in central Mato Grosso'* display well the cognitive geographic conceptualization of direction notions used. Understanding and retrieving location information from such textual phrases is an important challenge especially, when users want to know information 'about' or 'within' some portion of a region, described with reference to a compass direction.

In the times of historic expeditions, i.e., way before GPS, when the tools for collecting and sharing information were limited, locations and their names were registered in-situ as textual descriptions by the expeditioners. Such descriptions display strong spatial relationships, especially in terms of geographic location, e.g., with reference to village, town, city and state. Not every place name may be a known locality, in which case it may be described with reference to other, known places. For example, in Figure 1 the location of MATÃO is identified with reference to the well-known place Ribeirão Prêto, indicating distance and direction relations, besides being located in central northern São Paulo state.

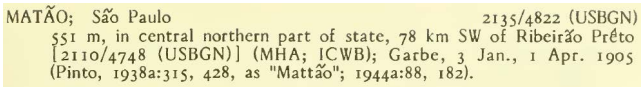

MATÃO; São Paulo 2135/4822 (USBGN)
551 m, in central northern part of state, 78 km SW of Ribeirão Prêto [2110/4748 (USBGN)] (MHA; ICWB); Garbe, 3 Jan., 1 Apr. 1905 (Pinto, 1938a:315, 428, as "Mattão"; 1944a:88, 182).

Figure 1: Example entry from the gazetteer [9].

We call phrases such as 'central northern' and 'northern central' *half-directions* in this paper. The reason is that we do not know whether they are a central part or a border sectoral part of a given region. The interpretation of such half-directions is important to allow distillation of location information from message content. This is the main problem addressed in this paper. It presents an extention of our work on more regular cardinal and ordinal compass direction phrases, tackled in [12].

Our objective is to identify interpretation models for English phrases such as 'central northern São Paulo state' or 'northern central Minas Gerais', with the aim to derive for such part-whole relations a best-as-possible spatial representation. By half-direction-based part-whole relation, we mean a text phrase that combines the word 'central', and a cardinal compass direction (in any order) with the name of a region, for which we know the spatial extent, as in the examples above. The problem of interpreting location with reference to distance and direction has been addressed in [7, 6, 16] while identification of internal cardinal directions parts was studied in [12, 8] and identification of cardinal direction relations between geographic entities has been studied in [4,

ACM SIGSPATIAL IWGS'12, November 6, 2012. Redondo Beach, CA, USA.

5, 13].

Some important work is by [14], who developed algorithms for partitioning polygons into cardinal direction-based sectors. Their theoretical work addressed the four cardinal directions as partitions, but excluded ordinal directions and a notion of central sector; these were addressed by [12] in a more pragmatic way. This work is believed to improve substantially on many of the cases, and revive and improve on the fundamental work by Paynter and Traylor [10, 9]. In this paper we develop a similar approach but for half-direction entries that had not been covered in [12]. We also believe that our previous [12] and current work could be useful to the geostreaming community where the datasets are acquired continually over time and have to be processed on-the-fly, especially when handling text data from news, SMSs and twitter.

2. RELATED WORK

Large amounts of geographic information have been collected historically in the form of textual content, with smaller amounts as maps, and hardly anything as spatial data. An important case is that of specimen labels in biological museum collections, and that of travel logbooks of expeditioners of the past. Descriptions in these collections exhibit a variety of spatial relations between features [16]. In natural language, identification and understanding of spatial relations is important, and is fundamental to building geospatial semantics [2].

Early work by [4] suggest two methods for sector recognition: angle-based and grid-based sectors, which were used to interpret cardinal directions between two spatial objects. These models were applied mostly for static applications of point object referencing, in which the grid-based approach is cognitively more acceptable, whereas angle-based directions proved to be a better model for movement applications. Subsequently, [5] proposed an envelope approach for representing cardinal direction relations between two non-point objects. In this approach, the partition lines of the envelope of the reference object are extended until they intersect the envelope lines of the other object. This approach clearly gives unequal cardinal direction zones that depend on the shape of the reference objects. Some other challenges of that model were discussed in [13]. The latter approach does not approximate a region to a point.

Various methods and models have been proposed to compute direction from a reference object to a target object. In [16], a point-and-radius method is proposed to georeference locality descriptions. Many factors — like distance, direction, map scale — are taken into consideration while computing the georeference of a target locality. The method determines not only the georeference, but also the uncertainty associated with the respective inputs. We plan to publish separately on these associated uncertainties.

The above approaches identify cardinal direction relations between two objects externally, not between a region and its subregions. Recent work has addressed the latter to some extent. In [8], external techniques are used to determine region/subregion compass directions. Three approaches are offered, one of which is similar to the model of [5]. Amongst others, in a cognitive experiment, subjects were asked to assign a direction and level of accuracy to a number of presented points on a map [8], in an attempt to determine the ideal value of ρ, being the scaling factor from region to central sector. Varying over values for ρ from $\frac{1}{12}$ to $\frac{2}{3}$, an optimal value of $\rho = \frac{1}{3}$ was found. Using this value, the other eight sectors were determined.

The problem of determining direction-based sectors was also addressed in [14]. That body of work addressed only the four cardinal directions, and presented both criteria for splitting the original, as well as efficient algorithms for determining the extents of sectors, meeting those criteria. Also, their work did not address criteria for deriving the ordinal sectors from the polygon and the notion of central region in a polygon. Work by [12] provides the computational models of determining not only cardinal and ordinal directions but also the computational models for central notions of direction. It attempts to address more pragmatically the problems for a wider range of directions, including the central sector, while evaluating also different models and not address the complexity of the general case of algorithms as that of [14]. Our work for this paper addresses another set of non-standard directions (complementing [12]), we call half-directions, and was briefly addressed in [11].

3. DATA USED

Our gazetteer data derives from the two volumes of the Ornithological Gazetteer of Brazil [10, 9], which provides over 8,000 entry descriptions in natural language text. These descriptions cover over 3,200 sites visited by expeditioners almost two centuries of natural history expeditions in Brazil. The books are a publication in a longer series by the same authors, covering most other Latin American countries, which were published between 1975 and 1991.

In our gazetteer, place name descriptions from 25 states of Brazil apply a variety of spatial relations – distance, direction and topological. The most important spatial relation in our gazetteer descriptions is regional containment. Here, we focus on directional containment cases as described above. In the gazetteer, 25 compass directions are found in containment relation patterns,. These compass directions can be categorized as cardinal direction (e, w, n, s, and as extension: c); ordinal direction (se, sw, ne, nw) and half-direction (nc, cn, amongst others). The first two categories are standard direction categories of the compass rose, whereas half-directions are those directions that are neither cardinal nor ordinal directions but are somewhere between them, thereby forming non-standard direction notions such as cn, nc, wn, cs and others. Of all directions, ordinal directions (38%) are more frequent than cardinal directions (32%) and various half-directions (30%). For this paper, we focus on the third group, i.e., the half-directions, giving a total number of 556 corpus entries with known location and known regional containment. Of these, our main focus is on the eight half-direction (i.e., nc, cn, ec, ce, sc, cs, wc, cw) amounting to 532 entries.

The state boundary polygons were obtained from Esri's ArcWorld Supplement Map data, dated 1998. We merged the states of Goiás, Tocantins and the Distrito Federal as a single state 'Pre-1988 Goiás' to reflect the gazetteer's notion of Goiás state. Our methods are equally valid on higher accuracy data than these.

4. DATA ANALYSIS

The half-direction entries are not evenly spatially distributed, as is illustrated in Figure 2. Unsurprisingly, there is a rather

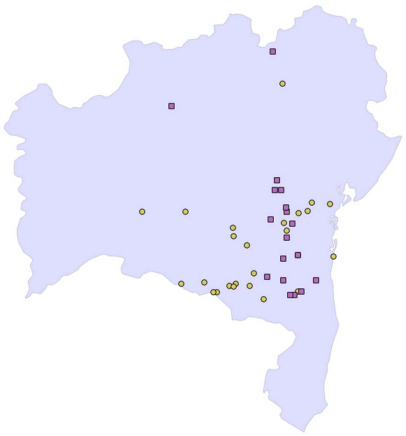


Figure 2: Distribution of half-direction entries in the state of Bahia, Brazil.

strong coastal and southeastern bias at not just national but also at province level. This can be explained from historic accessibility and expeditioners' region preferences, with the Brazilian Atlantic Forest being a main target for early explorations. Notwithstanding this history, it is also apparent from the data (Figure 3) that entries tagged cs, sc and ce, ec outnumber the entries tagged with other half-directions. Below, we analyse the distance and angle characteristics of half-direction entries.

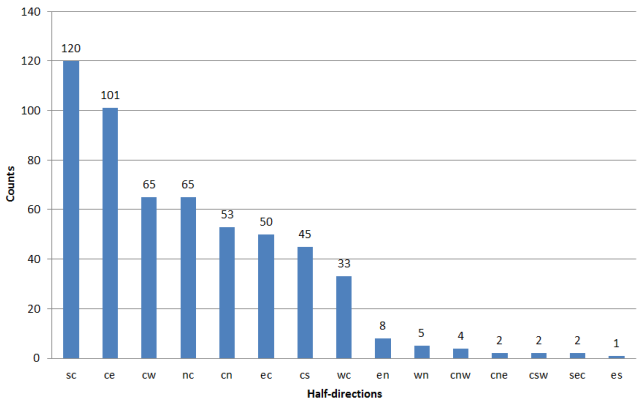


Figure 3: Histogram of distribution of half-direction entries.

4.1 Distance analysis

Our previous work suggests strongly that centroid (mass centre of state) is the much preferred notion of centre [12]. States differ substantially in size, making comparisons between states difficult, and derivation of generic rules of size and delineation of a central sector per state also hard. So, to overcome these differences, we devised a metric named avg_dist2centroid, per region P, defined as

$$\text{avg\_dist2centroid}(P) = \frac{1}{\text{area}(P)} \int_{p \in P} \text{dist}(\text{centroid}(P), p)\, \delta p.$$

The metric determines the average distance to centroid over all points within a region P. We use this metric to normalise distances to centroid within different regions, so as to make them comparable [3]. We analysed the distribution of all normalised distances-to-centroid, for both central and non-central half-direction entries.

We initially hypothesised that our 'c*' (e.g., central northern) and '*c' (northern central) entries are separate cases and that '*c' entries occur within the central region, and 'c*' entries are somehow central to the '*' border sector. We therefore coined '*c' entries 'central half-direction entries' and 'c*' entries 'non-central half direction entries'. Further, we analyzed the distribution of normalized distances-to-centroid for central and non-central half-direction entries, by their histograms and probability density functions. The latter (Figure 4) were derived by fitting standard normal distribution curves, for the non-central half-direction entries giving $N[\mu = 0.8707; \sigma = 0.3133; n = 286]$, and a normal distribution for the central half-direction entries, giving $N[\mu = 0.7441; \sigma = 0.2841; n = 272]$, where μ and σ are expressed with the respective state's average normalised distance as unit. This statistical analysis revealed that the average distances for '*c' and 'c*' entries were rather close, namely, 0.74 and 0.87 normalised unit distance. Should '*c' and 'c*' indeed be different types of entry, their differences would need to be larger, and the amount of overlap cases should be smaller.

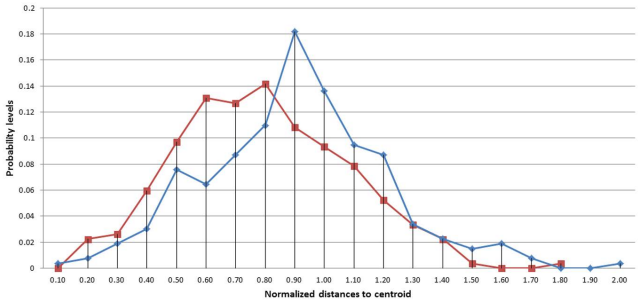


Figure 4: Frequency of normalised distance-to-centroid for central half-direction (red) and non-central half-direction (blue) entries in the study.

We further tested our hypothesis by applying two independent sample t-tests on the normalized distances on the groups of '*c' and 'c*' entries. The null hypothesis for this test is:

H_O: The mean of normalized distances to centroid is equal for both samples of the group, i.e., $\mu_1 = \mu_2$.

We use four groups, reflecting cardinal compass directions, with two independent samples each to test our null hypothesis. The test is done with confidence of 95% and significance value of 0.05. We report the results of each group in Table 1 which suggests that our null hypothesis is true only for two groups. There is still room for more statistical analyses to test our hypothesis and we will come to this again in Section 4.2.

We also analysed the distance to the outer boundary, normalised again per state, for central and non-central half-directions. Normal probability density functions (Figure 5) were derived by fitting a normal distribution giving $N[\mu = 0.355; \sigma = 0.211; n = 264]$ for non-central half-direction en-

Table 1: Results of independent sample t-test with unequal variance on half-direction entries based on normalized distance to centroid

Notation	nccn	ecce	sccs	wccw
Sample 1	cn	ce	cs	cw
$\mathbf{n}_1$	53	101	45	65
μ_1	0.848	0.903	0.871	0.771
Sample 2	nc	ec	sc	wc
$\mathbf{n}_2$	65	50	120	33
μ_2	0.761	0.739	0.740	0.732
t	1.374	-3.704	2.862	-0.583
df	108.7	118.2	81.7	62.2
p **sig.(1-tailed)**	0.086	0.001	0.002	0.281
Results ($\mathbf{H}_o$)	valid	rejected	rejected	valid

tries, and$N[\mu = 0.381; \sigma = 0.250; n = 268]$ for central half-direction entries.

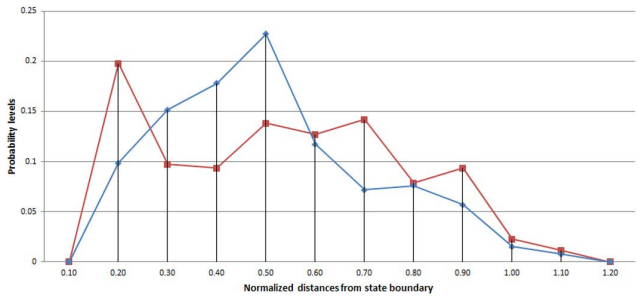


Figure 5: Frequency of normalised distance-from-boundary for central half-direction (red) and non-central half-direction (blue) entries in the study.

Again, we carried out a t-test for four groups of half-directions to test our null hypotheses. The null hypothesis being:

H_0: The mean of normalized distances to the respective state boundary is equal for both samples of the group, i.e., $\mu_1 = \mu_2$.

Table 2: Results of independent sample t-test with unequal variance on half-direction entries based on normalized distance from state boundary

Notation	nccn	ecce	sccs	wccw
Sample 1	cn	ce	cs	cw
$\mathbf{n}_1$	53	101	45	65
μ_1	0.384	0.309	0.398	0.371
Sample 2	nc	ec	sc	wc
$\mathbf{n}_2$	65	50	120	33
μ_2	0.375	0.369	0.405	0.317
t	0.187	1.785	0.195	-0.988
df	109.9	82.04	106.12	55.37
p **sig.(1-tailed)**	0.426	0.04	0.423	0.164
Results ($\mathbf{H}_o$)	valid	valid	valid	valid

Table 2 suggests that our null hypothesis is accepted for all four groups and that '*c' and 'c*' entries belong to similar population in each group.

We tested the presence of any correlation between normalised distances from centroid of all half-direction entries within a state and the state sizes. The correlation coefficient R^2 was 0.225 for smaller state sizes and 0.164 for large state sizes. Since these values are small, we concluded that there is no substantial correlation between normalised distances to centroid and state sizes.

On testing the correlation between normalised distances from state boundary and state sizes the results suggest that there is no correlation between them as R^2=0.08 for small state sizes and 0.004 for large state sizes.

4.2 Azimuth analysis

To study the angle spread of 'c*' and '*c' entries in the '*' direction we calculated the azimuth of entries from mass centre of state and azimuth difference for each entry from their respective cardinal directions. For example, to study the angle spread of cn and nc entries their azimuth difference was calculated from the N direction. After a comprehensive analysis and percentile calculations, circular box plots [1] were developed for all eight half-directions showing the spread of all entries for10/25/50/75/90-percentiles as deviation from the main direction (see Figure 6). From the figure it can be inferred that barring a few exceptions in 'nc, sc, cs' half-directions, all other entries fall within the boundaries of ordinal directions. This finding is used at the time of model creation.

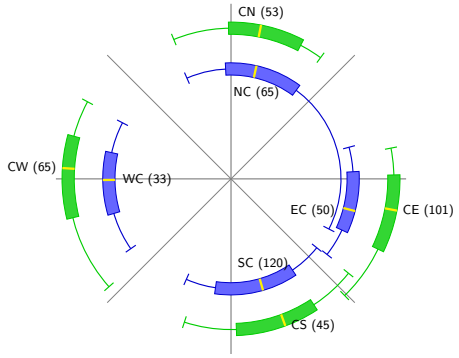


Figure 6: Azimuth spread indicated as 10/25/50/75/90% boxplots from region mass centre over the compass rose for: for individual '*c' and 'c*' entries.

In the last section, we tested our hypothesis if '*c' and 'c*' entries belong to different populations through distance analyses. In this section we test our hypotheses using the azimuth analysis to see of the results obtained are different from distance analyses.

A plot of polar coordinates of all half-direction entries using their azimuths from centroid illustrates the case. While mapping polar coordinates on the $x - y$ plane we expected that '*c' entries fall closer to origin and 'c*' entries are further away from origin to agree with our working hypothesis. However, the plot showed that '*c' and 'c*' entries shows substantial overlap and are close on average (Figure 7).

We tested our hypothesis further, by applying t-test on the azimuth difference of all half-direction entries. The hypothesis are:

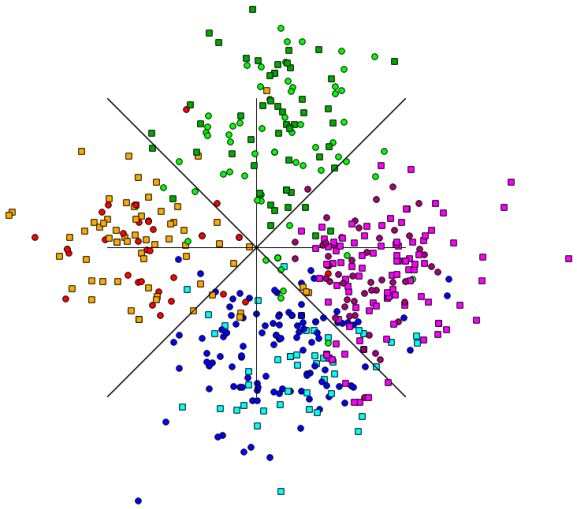


Figure 7: All half-direction entries shown in polar coordinates. Entries shown as boxes represent 'c*' entries and with circles represent '*c' entries.

Table 3: Results of independent sample t-test with unequal variance on half-direction entries based on azimuth differences

Notation	**nccn**	**ecce**	**sccs**	**wccw**
Sample 1	cn	ce	cs	cw
$\mathbf{n}_1$	53	101	45	65
μ_1	-11.343	-13.292	17.740	5.329
Sample 2	nc	ec	sc	wc
$\mathbf{n}_2$	65	50	120	33
μ_2	-25.026	-11.735	12.054	6.367
t	-1.865	0.431	-1.179	-0.124
df	99.187	98.565	93.617	66.236
p **sig.(1-tailed)**	0.033	0.334	0.121	0.451
Results ($\mathbf{H}_o$)	valid	valid	valid	valid

H_O: The means of azimuth difference to centroid is equal for both samples of the group, i.e., $\mu_1 = \mu_2$.

The results (Table 3) of this test validate our null hypothesis for all the four groups i.e., nccn, ecce, sccs, wccw. In other words, there exists no reason to consider the two samples in each of the four groups as dissimilar. Hence, we combine the samples in each group and create another circular box plot [1] for those four groups. Figure 8 shows the spread of all entries within 10/25/50/75/90-percentiles for deviation from the main direction.

It is interesting to see that there are a high number of entries in the eastern and southern directions, something we already observed in [12]. Hence there is a general skew in the east and south direction for not just half-direction entries but also cardinal and ordinal direction entries. It is interesting to know that out of 25 states in Brazil eleven states are located on the e/se/s coast. These coastal states also account for 61% of the number of half-direction entries. We believe that this skew in the coastal directions is best explained from explorer and collector arrival in Brazil by

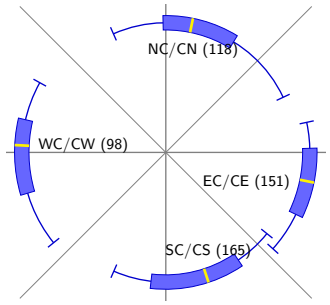


Figure 8: Azimuth spread indicated as 10/25/50/75/90% boxplots over the compass rose for nccn, ecce, sccs, wccw.

the coast side, and a relatively unexplored and undeveloped coastal hinterland at that time. In the following paragraph, we examine the presence of directional skew for coastal and non-coastal states separately.

We divided the states into two groups: coastal and non-coastal states, based on their location on the e/se/s coast, and identified eleven states as coastal and the remaining 14 as non-coastal states. A comprehensive percentile calculation was again carried out for four groups i.e., nccn, ecce, sccs, wccw separately for coastal and non-coastal states (Figure 9).

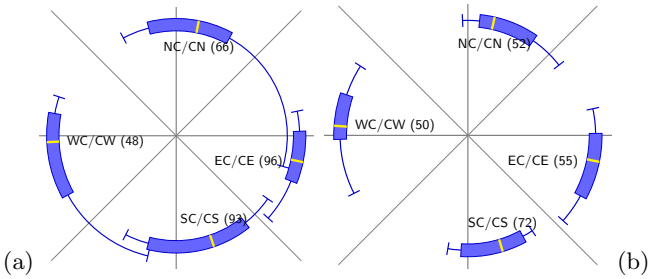
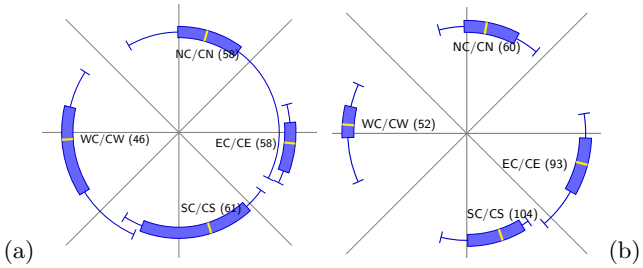


Figure 9: Azimuth spread indicated as 10/25/50/75/90% boxplots over the compass rose for nccn, ecce, sccs, wccw: (a): for coastal states; (b): for non-coastal states.

It can be seen from the figures that for non-coastal states the number of entries and their azimuth differences, are quite nicely spread in all the four directions. However, coastal states have an evident bias towards the east/south-east direction. One can see that even some of the entries classified as wc or cw, and nc or cn appear to fall in east/south-east direction. This emphasises the eastern skew in our data set once again and supports our earlier hypothesis.

Similar to distance analysis in Section 4.1, we tested whether less compact and more compact states show different angle spreads. We used the same groups of states — less compact and more compact states, using the same Roeck scores and carried out percentile analysis to work out the angle spread of nccn, ecce, sccs, wccw half-direction entries. Figure 10, shows the angle spread in 10/25/50/75/90-percentiles for less compact states and more compact states.

It is interesting to see from Figure 10(a) that for less compact states even though some points were identified as nc/cn they fall in the eastern direction. Similarly, some points



(a) (b)

Figure 10: Azimuth spread indicated as 10/25/50/75/90% boxplots over the compass rose based on Roeck scores for: (a): for less compact states (low Roeck scores) for combined '*c' and 'c*' entries; (b): for more compact states (high Roeck scores) for combined '*c' and 'c*' entries.

identified as wc/cw by the collectors in fact fall in the southern direction. Since less compact states are expected to have a geometry deviant from approximately circular shapes (almost evenly spread in all direction) it is possible that the collectors could not accurately understand in which part of the state they were positioned. This explains why some of the points that fall in southern and eastern direction were not identified as such. It can also be seen that more compact states (almost evenly shaped) show higher frequency of about 63% of the entries falling in south/east direction, yet again an eastern/southern skew.

5. MODELS FOR INTERPRETATION OF HALF-DIRECTIONS

5.1 Organisational principles for models of interpretation

Before we use our analysis to create interpretation models, we discuss the important organisational principles. The distinctions that one can draw directly affect the spatial assignations to the sectors, as well as the reasoning that one can perform over those assignations. Our work is organised using the following principles:

1. every compass sector falls with its region,
2. every compass sector in a region contains two half-direction sectors,
3. half-direction sectors of the compass sector can overlap,
4. different compass sectors do not overlap,
5. angle computations take place in a geographic reference system, while length and area computations take place in a (metric) projected reference system,
6. interpretation models are always relative to a centre point in the region,
7. the choice of the centre defines the placement of our models, and
8. all compass sectors together cover the region.

This list partially coincides with that of [12], [14]. In Sections 5.2, we discuss various choices of centre respectively. This determines placement of our model for interpretation of half-direction based part-whole relations.

5.2 Characteristics for model creation

There are three important characteristics in creating models for half-direction-based part-whole relations. They are: *shape*, *placement* and *size*. Our models attempt to be minimal, yet sufficient spatial extents for each of the possible half-directions for each state.

We allow circle and hull *shape* for creating '*c' and 'c*' models. The extent of the circle and hull models are derived from the distance analysis in Section 4.1. The hull models are convex hulls of state scaled down to the size of the circle models, to allow fair comparison. We further create angle sectors that form cones originating from some choice of centre in the region. In our models, all four directions in both 'c*' and '*c' sectors have a natural and identical fan-out angle of 45° on left and right from the main direction, and this gives us the angular sectors. Also, the compass spread in Figure 6 and polar coordinate plot in Figure 7 suggest that '*c' and 'c*' entries fall within the natural sectors formed at 45° from the main directions, respectively.

It is imperative for any half-direction model to have a *placement* in its state. In [12], we studied four different notions of centre for model placement and recognized the region mass centre as the best performing placement option over other centres. Other notions of centre considered in our previous study were *envelope mass centre*, *circle mass centre* and *mass box centre*. Since we concluded in [12] that region mass centre is the best option for model placement in a state of any shape and size, we use that notion of centre for placement of models created for half-directions.

The *size* of a half-direction model differs between '*c' and 'c*' sectors. We use results from our analysis on distances from centroid to determine appropriate sizes of our models. Different sizes for '*c' and 'c*' models are discussed below.

5.3 Model creation for central half-direction entries

A model for '*c' half-directions determines a sector in the region in which those entries are expected to be located. As before, two categories of central model are proposed, based on *shapes* i.e., the circle and hull. We first discuss the circle models and then the hull models. The models have names with convention 'MC/x/xx', in which MC stands for model of central category, x refers to the shape (c for circle, and h for hull) and xx refers to the cut-off percentage.

In our first approach, we create central models based on circular shape and 90-degrees based compass sectors. The size of the model, i.e., its minimum and maximum extent is derived from a choice of percentage levels ranging from 65 to 95 % giving a two-tailed, symmetric split over the Gaussian curve Figure 4. These percentage levels represent the probability of the '*c' entries falling within the model sector. For instance, for the 95-percentage level the minimum and maximum extents are defined at a distance of 0.20 to 1.30 times the average distance of the state, respectively, from the state centroid. These factors are derived from a Gaussian curve fitted through all '*c' entries in our corpus with the symmetric split chosen at 95-%. We call this model 'MC/c/95'. Similarly, we create models at percentage levels

of 65, 70, 75, 80, 85 and 90%, thereby giving us a total of seven central models with the circular shape. See Figure 11 for 'MC/c/75'.

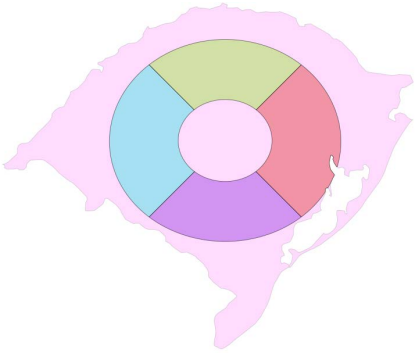


Figure 11: A circle shape central model 'MC/c/75' for Rio Grande do Sul state, Brazil.

In the second approach, we create central models based on hull shape and 90-degrees based compass sectors. Hulls are convex hulls of state scaled down to the size of the earlier created circle extents, to allow fair comparison. Hence, we use the seven models created above by changing the shape from circle to hull. The convex hulls are created in such a way that the area bound by the inner and the outer convex hull of a model, is equal to the area bound by the inner and outer circles, respectively, of the original model created in our first approach. Thus, we have seven more central models but now based on hull shape and at the same percentage levels as in the first approach. For example, the model 'MC/h/95' is derived from the 'MC/c/95' created in the first approach. See Figure 12 for 'MC/h/75'.

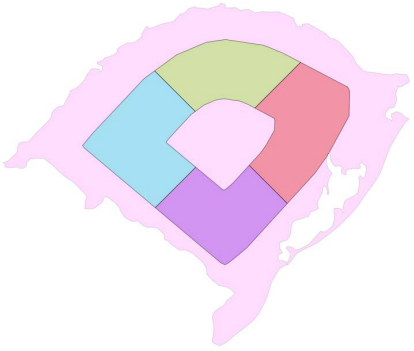


Figure 12: A hull shape central model 'MC/h/75' for Rio Grande do Sul state, Brazil.

5.4 Model creation for non-central half-direction entries

Models for 'c*' entries determine a region in the state where such entries are expected to fall. Here, we discuss variation in size for non-central half-direction models. The models are named with convention 'MN/x/xx', in which MN stands for model of non-central category, x refers to the shape (c for circle, and h for hull) and xx refers to the cut-off percentage.

In our first approach, we created non-central models based on circular shape and 90-degrees based compass sectors. The minimum and maximum extents are derived from the Gaussian curve at various percentage levels ranging from 65 to 95 %. For the 95 % level, the minimum and maximum extents are defined at a distance of 0.28 to 1.52 times the mean distance of the state. We call this model 'MN/c/95' and likewise we created models at percentage levels of 65,70,75,80,85 and 90% thereby giving us a total of seven non-central models on circular shape. See Figure 13 for 'MN/c/75'.

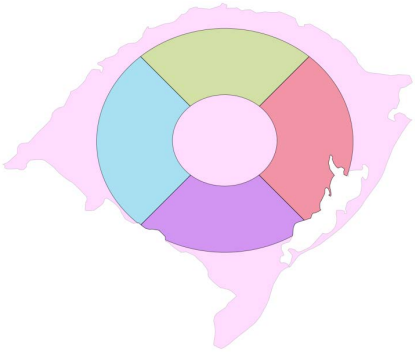


Figure 13: A circle shape non-central half-direction model 'MN/c/75' for Rio Grande do Sul state, Brazil.

In the second approach, we create non-central models based on hull shape and 90-degrees based compass sectors. As in Section 5.3, we use the seven models created above by changing the shape from circle to hull. Again, the convex hulls are created in such a way that the area bound by the inner and the outer convex hull of a model, is equal to the area bound by the inner and outer circles respectively of the original model created in our first approach in this section. Thus, we have seven more non-central models but now based on hull shape and at the same percentage levels as in the first approach in this section. For example, the model 'MN/h/95' is derived from the 'MN/c/95' created in the first approach. See Figure 14 for 'MN/h/75'.

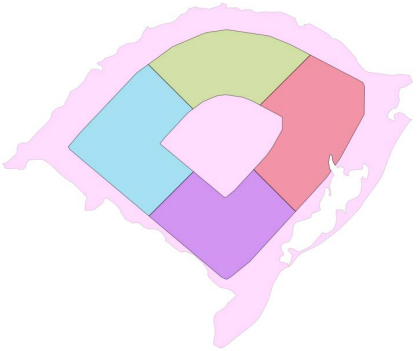


Figure 14: A circle shape non-central half-direction model 'MN/h/75' for Rio Grande do Sul state, Brazil.

6. RESULTS OF MODEL COMPARISON

The comparison for all '*c' and 'c*' models are made with precision/recall scores [15] for all 532 half-direction entries in Brazil. The results of recall score R range between 0.53 and 0.82, and precision scores P range between 0.84 and 0.91. Below, we report the P, R scores for central and non-central models for both circle and hull shapes.

6.1 Comparison of central models

We compare both circle and hull shape models pairwise when applied with same percentage levels. We find that models with circle shapes show higher recall scores than hull models for all percentage levels. At the same time however, the models with hull shapes show higher precision scores than circle models for all percentage levels. Overall, the precision P ranges between 0.84 and 0.91 and recall R range between 0.53 and 0.79 for all circle and hull models.

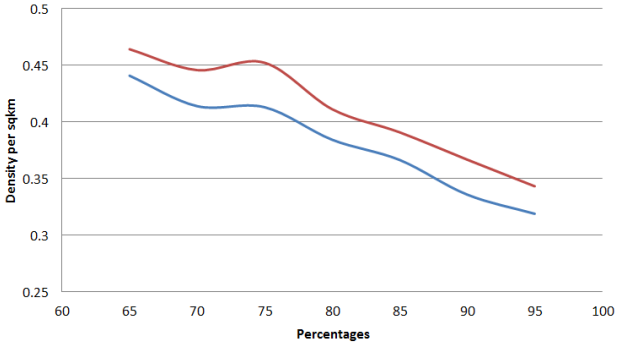


Figure 15: Comparison of densities achieved by circle (red) and hull (blue) central models.

We further evaluated the models on the basis of density of the relevant and retrieved entries per unit area of the model. In Figure 15 we show the densities achieved in various circle and hull central models. It is quite clear from this evaluation that the circle models outperform the hull models for all percentage levels. It is safe to conclude that for central models, circle shapes are better than hull shapes. Within the circle shape models, 'MC/c/95' is found to be the best model since it carries the highest recall score. We prefer to rely on the recall score and not precision because we wish to maximise the relevant entries that are retrieved by a model and this score is reflected in the recall. In order to confirm that we are not making a compromise on precision scores, we tested the trend in precision and recall scores w.r.t increase in percentage values as shown in Figure 16. It was found that by increasing the percentage values (from 60% to 95%) precision scores drop slightly but recall increases notably. By going for maximum precision, we have to settle for considerably lower recall values but not vice versa. Hence, we confirmed that by choosing the recall score as our decisive score, we have not compromised on the precision scores.

6.2 Comparison of non-central models

For non-central models too, we compare the circle and hull shape models for all percentage levels. We find that circle models show higher precision scores than hull models from 65 to 80-percentage levels while the hull models show higher recall scores for the same percentage levels. Beyond

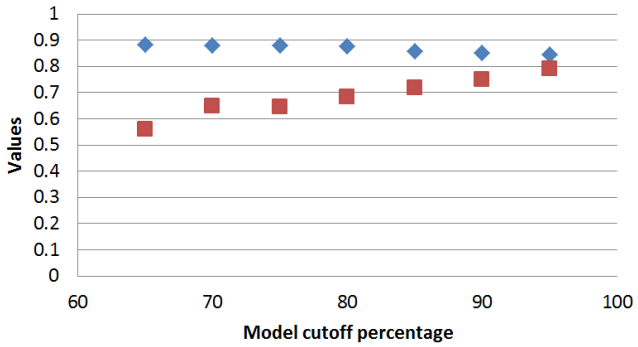


Figure 16: Trend of precision (blue) and recall (red) w.r.t percentage values for circle based *c models.

the 80-percentage level, the scenario reverses such that the circle models show higher recall scores where hull models show higher precision. The precision scores P for all non-central models range from 0.86 to 0.91 and the recall scores R range from 0.58 to 0.82.

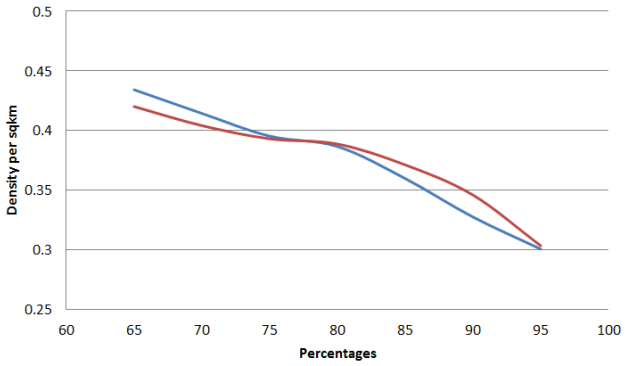


Figure 17: Comparison of densities achieved by circle (red) and hull (blue) non-central models.

The densities achieved by the models are shown in Figure 17. A similar pattern is observed in the figure as well since the densities achieved by hull models are higher than those of circle models, like their recall scores, until approximately the 78% level. This changes however, after this level and circle models show higher densities than hull models, like their recall scores. It is worth mentioning that with the increase in percentage levels, the densities achieved decrease. To pick the best performing non-central model we rely again on the recall scores and find that 'MN/c/95' performs the best. This model is closely followed by 'MN/h/95'.

7. PAPER CONTRIBUTION AND FUTURE WORK

In this paper, we provide interpretations of non-standard direction notions which we call half-directions and this work emanates out of the need highlighted in [12]. We present a computational approach to derive geometric extent of half-direction in a state. We developed a number of models, and validated these against a body of gazetteer entries. The model construction is based on different shapes and sizes.

On validation, we found that models with circle shape performed better than those with hull shape for both central and non-central half-directions, when considering their recall scores.

In our future work, we hope to extend our framework to other than directional part-whole relations, allowing more textual interpretation to geometry, even including measures of spatial uncertainty in this context. This aims to allow for synthetic interpretation models that assign linguistically steered spatial uncertainty to features in this context.

8. ACKNOWLEDGMENTS

Gaurav Singh's research is supported by the EU Erasmus Mundus program External Cooperation Window 2009-LOT 15, and by the Faculty ITC research fund. Rolf de By's work on this publication was in part supported by the Dutch national program COMMIT.

9. REFERENCES

[1] A. H. Abuzaid, I. B. Mohamed, and A. G. Hussin. Boxplot for circular variables. *Computational Statistics*, pages 1–12, 2011.

[2] I. B. Arpinar, A. Sheth, C. Ramakrishnan, E. L. Usery, M. Azami, and M.-P. Kwan. Geospatial ontology development and semantic analytics. *Transactions in GIS*, 10:551–575, 2006.

[3] R. A. de By. A funny distance computation and ... the power of spatial SQL. Technical report, Faculty of Geo-information Science & Earth Observation (ITC), University of Twente, 2012.

[4] A. U. Frank. Qualitative spatial reasoning about distances and directions in geographic space. *Journal of Visual Languages & Computing*, 3(4):343–371, 1992.

[5] R. K. Goyal and M. J. Egenhofer. Consistent queries over cardinal directions across different levels of detail. In *Proceedings 11th International Workshop on Database and Expert Systems Applications, Greenwich, U.K.*, pages 876–880, 2000.

[6] Q. Guo, Y. Liu, and J. Wieczorek. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, 22:1067–1090, January 2008.

[7] Y. Liu, Q. Guo, and M. Kelly. A framework of region-based spatial relations for non-overlapping features and its application in object-based image analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 63(4):461–475, 2008.

[8] Y. Liu, X. Wang, X. Jin, and L. Wu. On internal cardinal direction relations. In A. Cohn and D. Mark, editors, *Spatial Information Theory*, volume 3693 of *Lecture Notes in Computer Science*, pages 283–299. Springer Berlin/Heidelberg, 2005.

[9] R. A. Paynter, Jr. and M. A. Traylor, Jr. *Ornithological Gazetteer of Brazil*, volume N–Z, 2. Harvard University, Museum of Comparative Zoology, Bird Department, Cambridge, Ma., U.S.A., 1991.

[10] R. A. Paynter, Jr. and M. A. Traylor, Jr. *Ornithological Gazetteer of Brazil*, volume A–M, 1. Harvard University, Museum of Comparative Zoology, Bird Department, Cambridge, Ma., U.S.A., 1991.

[11] G. Singh and R. A. de By. Interpretation models for non-standard compass directions. In *Extended Abstracts: Proceedings 7th International Conference on Geographic Information Science, Columbus, OH, U.S.A.*

[12] G. Singh, R. A. de By, and I. Ivánová. Concepts, compass and computation: Models for directional part-whole relationships. In B. Murgante, O. Gervasi, S. Misra, N. Nedjah, A. M. A. Rocha, D. Taniar, and B. O. Apduhan, editors, *Computational Science and Its Applications – ICCSA 2012*, volume 7334 of *Lecture Notes in Computer Science*, pages 286–301. Springer, 2012.

[13] S. Skiadopoulos and M. Koubarakis. Composing cardinal direction relations. *Artificial Intelligence*, 152(2):143–171, 2004.

[14] M. J. van Kreveld and I. Reinbacher. Good NEWS: Partitioning a simple polygon by compass directions. *International Journal of Computational Geometry & Applications*, 14:233–259, 2004.

[15] C. van Rijsbergen. *Information Retrieval*. Butterworth, London, Boston.

[16] J. Wieczorek, Q. Guo, and R. J. Hijmans. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, 18(8):745–767, 2004.