*International Journal of Geographical Information Science*
Accepted 21 Nov 2024, Published online: 03 Dec 2024, 2024, 1–29

# RESEARCH ARTICLE

# Probabilistic Qualitative Spatial Reasoning with Applications to GeoQA

Mohammad Kazemi Beydokhti[a], Matt Duckham[a], Amy L. Griffin[a], Yaguang Tao[a], Ross Purves[b] and Maria Vasardani[a]

a. School of Science, RMIT University, Melbourne, Australia
b. Department of Geography, University of Zürich, Zürich, Switzerland

This paper explores the use of probabilistic and conventional qualitative spatial reasoning (QSR) in the context of geospatial question answering (GeoQA) systems. The paper presents a thorough empirical investigation of the performance of a probabilistic and a conventional qualitative spatial reasoner, across a range increasingly sophisticated scenarios with real data and synthetically generated questions. The results indicate the potential of probabilistic QSR to provide more detailed information about spatial configurations than conventional QSR; but at the cost of less frequent errors in estimating the relative likelihood of different reasoning conclusions. Errors in probabilistic reasoning also tend to be systematically associated with lower probability conclusions. The results have implications for reliable and flexible automated spatial reasoning systems, especially where neither conventional geographic information retrieval (GIR) techniques nor large language models (LLMs) are able to provide a satisfactory solution to GeoQA problems.

**Keywords:** qualitative spatial reasoning; question answering systems; probabilistic reasoning; geoQA; geoAI

## 1. Introduction

GeoQA (geospatial question answering) is the development of systems capable of generating or retrieving valid answers to geospatial questions posed by humans in natural language (Scheider *et al.* 2021). Today's GeoQA systems rely primarily on geographic information retrieval (GIR) techniques to retrieve stored answers from knowledge bases (Scheider *et al.* 2021, Chen *et al.* 2013, Mai *et al.* 2019). However, GIR techniques alone lack the reasoning capabilities to infer new knowledge from stored information, limiting

Corresponding author: mohammad.beydokhti@rmit.edu.au, Mohammad Kazemi Beydokhti

ISSN: 1365-8816 print/ISSN 1362-3087 online
© 2024 Taylor & Francis
DOI: 10.1080/13658816.2024.2434613
http://www.informaworld.com

the range of questions that might potentially be answered using stored data. Large language models (LLMs) have also made significant recent advances in providing plausible responses to natural language questions, including in spatial domains (e.g., Mai *et al.* 2022, Hu *et al.* 2023). However, as statistical models of word frequencies, LLMs are acknowledged to be ill-suited to performing spatial reasoning tasks reliably (Mai *et al.* 2023, Cohn and Hernandez-Orallo 2023, Cohn 2023, Kazemi Beydokhti *et al.* 2024).

This article explores how probabilistic and conventional qualitative spatial reasoning (QSR) approaches can help to bridge this gap, by inferring answers to geospatial questions from stored information. In particular, this paper compares recent advances in probabilistic QSR techniques with conventional "crisp" (in the sense of "not probabilistic") QSR. Logical conclusions inferred using conventional QSR usually form a set of possible relations, without any mechanism to distinguish between more or less probable configurations in that set. By contrast, probabilistic qualitative spatial reasoning is able to estimate the relative likelihoods of different possible spatial configurations, based on the structure of the logic and even the strength of input evidence (using Markov logic networks, described in Duckham *et al.* 2023a).

In addressing the identified gap, this paper compares the performance of conventional and probabilistic qualitative spatial reasoners across a simple and a complex spatial reasoning task. The main contribution of this paper is a systematic comparison of the relative performance of conventional and probabilistic qualitative spatial reasoners on different tasks. The evaluation includes two different reasoning scenarios with different spatial logics (specifically, the cardinal direction calculus, CDC, and region connection calculus, RCC) using real spatial data with synthetically generated questions.

The structure of the remainder of the paper is as follows. Section 2 reviews related literature on the challenges of spatial reasoning in the context of GeoQA, GIR, and LLMs, followed by an overview of qualitative spatial reasoning (QSR). Section 3 and Section 4 step through a sequence of two experiments that compare the performance of conventional and probabilistic spatial reasoning in increasingly sophisticated GeoQA scenarios. Section 5 illustrates how these conventional and probabilistic reasoning can be integrated into a GeoQA system and sketches the key future challenges for QSR-augmented GeoQA in the context of advances in other areas of AI, in particular machine learning and LLMs. Finally, Section 6 concludes the paper, placing these results in the wider context of recent advances in AI.

## 2. Background

Significant recent research effort has focused on the challenge of GeoQA: providing reliable answers to natural language spatial questions (e.g., Zelle and Mooney 1996, Chen 2014, Huang *et al.* 2019, Lobry *et al.* 2020, Kruiger *et al.* 2021, Kazemi Beydokhti *et al.* 2021). Mai *et al.* (2021) has classified existing GeoQA systems based on the types of geospatial questions they answer, including "factoid" (Zelle and Mooney 1996, Chen 2014), "geo-analytical" (Kruiger *et al.* 2021, Kazemi Beydokhti *et al.* 2021), "scenario-based" (Huang *et al.* 2019), and "visual questions" (Lobry *et al.* 2020). Our focus in this research is primarily on the simplest and most common "factoid" category.

Some of the pioneering work on GeoQA is founded on geographic information retrieval (GIR) over knowledge graphs, such as DBpedia (Younis *et al.* 2012). More recently Punjani *et al.* (2018) adopted a similar approach of translating natural language questions into SPARQL queries over linked data knowledge graphs. The approach did not achieve

high performance, with only 22% of questions answered correctly. A key limitation of this and related GIR approaches, such as Hamzei *et al.* (2022), is the focus on retrieving stored spatial relationships, rather than inferring unknown spatial relationships from stored data. In general, only information stored explicitly in the knowledge graph, or that can be inferred through direct (non-spatial) logical inferences supported by SPARQL description logics, can be retrieved using this approach.

In the context of other recent advances in deep learning and LLMs (large language models), it is also natural to ask whether LLMs can help with qualitative spatial reasoning in GeoQA. The use of LLMs for spatial reasoning, and particularly for qualitative spatial reasoning (QSR) has been investigated by several recent studies, including Cohn and Hernandez-Orallo (2023), Cohn (2023), Kazemi Beydokhti *et al.* (2024). Cohn and Hernandez-Orallo (2023) provided an initial evaluation of GPT3.5 for "common-sense" spatial reasoning using a limited number of extended prompts and proposed future improvements for language model capabilities. This initial work has been further elaborated in Cohn (2023), which systematically evaluated the qualitative spatial reasoning capabilities of GPT-4, using two experiments. In the first experiment, GPT-4 was asked to compute the entire region connection calculus (RCC8) composition table, whereas the second experiment investigated how LLM performance compared to human performance. These and many other studies support the conclusion that machine learning models, such as LLMs, are inherently unreliable as deductive reasoners. Indeed, such limitations are only to be expected, since no matter their growing sophistication or training dataset sizes, the essence of LLMs still relies on statistical models of word frequencies (Bender *et al.* 2021, Kazemi Beydokhti *et al.* 2021).

In this context, both GIR and LLMs can only be expected to provide partial answers to GeoQA problems. However, QSR's logical reasoning capabilities may be able to complement and augment these techniques in answering spatial questions.

### 2.1. *Qualitative spatial reasoning*

QSR is concerned with logical reasoning over qualitative spatial relations, rather than over precise spatial quantities (Freksa 1991, Egenhofer and Mark 1995, Cohn and Hazarika 2001). As such, QSR more closely echos how humans reason about space in many everyday situations (cf. the pioneering work on cognitive maps, Tolman 1948, and naïve geography, Egenhofer and Mark 1995, for example).

Qualitative spatial calculi (sometimes referred to as qualitative spatial logics) are concerned with symbolic representation of spatial relationships, enabling spatial reasoning in complex spatial scenarios. Dylla *et al.* (2017) conducted a comprehensive overview of qualitative spatial calculi and their computational properties. The region connection calculus (RCC) is one of the best-known and most widely used qualitative spatial calculi (Randell *et al.* 1992). Like most qualitative spatial calculi, RCC is founded on a set of spatial relations between two regions, namely "inside" (proper part, PP), "contains" (proper part inverse, PPi), "equals" (EQ), "discrete" regions (DR, i.e., disconnected or externally connected regions), and "overlaps" (partial overlap, PO) in the RCC5 calculus (Cohn *et al.* 1997). In combination with these qualitative representations, the definition of mechanisms for the composition and inverse of relations provides the basis for logical qualitative spatial inference (Westphal and Wölfl 2009, Duckham *et al.* 2023b).

Qualitative spatial calculi, such as RCC, accommodate the imprecision of qualitative spatial reasoning in two ways. First, the representation of qualitative spatial relations is inherently imprecise. We often say spatial qualities "supervene" spatial quantities, in the

sense that a qualitative spatial relation can always be derived from a quantitative one, while the converse is not the case (e.g., a bearing of 276.8° may be represented qualitatively as "left", but "left" does not imply the specific bearing of 276.8°). Second, the conclusion of a logical qualitative spatial inference is a (possibly singleton) *set* of qualitative relations between two or more objects. In this way, qualitative spatial reasoning accommodates ambiguity in conclusions about which precise spatial configuration may hold.

Over decades, research has investigated ways of incorporating further uncertainties into QSR. For example, the "egg-yolk" calculus in Cohn *et al.* (1997), Cohn and Gotts (2020) aimed to extend RCC to reasoning about regions with indeterminate boundaries. Similarly, Loerch and Guesgen (1997), Petry *et al.* (2005), Schockaert *et al.* (2009) have explored various approaches to fuzzy RCC relations. More recently, the use of Markov logic networks (MLNs) has proven effective for qualitative spatial reasoning where the underlying evidence or inferred relations are uncertain (Nguyen and Mellor 2020, Nguyen 2019). Duckham *et al.* (2023a) presented the first general purpose technique for probabilistic qualitative spatial reasoning using MLNs. The approach encodes the composition and converse operations as rules in a logical network, much as do conventional qualitative spatial reasoning tools (such as SparQ, Dylla *et al.* 2006, Wolter and Wallgrün 2013, Wallgrün *et al.* 2006, or QSRlib Gatsoulis *et al.* 2016). It then drives inference using a probabilistic MLN, rather than the constraint network satisfaction that drives conventional "crisp" qualitative logics.

More information on the approach can be found in Duckham *et al.* (2023a). Rather than delving into the details of conventional and probabilistic reasoning, this research translates conventional and probabilistic QSR into the context of GeoQA, and explores the relative performance of the two approaches for advancing reliable GeoQA techniques.

## 3. Experiment 1: Questions about simulated direction relations

Our first experiment compared the performance of conventional and probabilistic qualitative spatial reasoners using a simple reasoning task that involved answering simulated questions about actual qualitative direction relations.

### 3.1. *Materials*

A body of spatial questions is crucial to our experimental approach with GeoQA. A few geographic question corpora that cover qualitative spatial questions have appeared in the literature, including the *GeoQuestion201* (Punjani *et al.* 2018) and the *Geoquery* (Zelle and Mooney 1996) datasets[1]. However, both these benchmarks are based on semantically enriched knowledge graphs, with a focus on geographic information retrieval rather than spatial reasoning.

Focusing on the spatial reasoning task specifically, there are only two GeoQA benchmark corpora available today: task 17 from Facebook's *bAbI* project (Weston *et al.* 2015) and *SPARTQA*[2] (Mirzaee *et al.* 2021). The synthetic bAbI dataset concerns tabletop,

[1]Geographic questions involve geographic entities such as place names, place types, or spatial relations within natural language questions. The term "spatial question" is often used interchangeably with the term "geographic question" in relevant studies; geographic question is a broader term, of which spatial questions are a subset.

[2]Unfortunately, the general purpose query language for knowledge graphs, SPARQL; the respected qualitative spatial reasoning engine, SparQ; and the GeoQA benchmark dataset SPARTQA all have confusingly similar

rather than geographic spaces (i.e., so-called "Zubin A-spaces," rather than "Zubin C-spaces," Zubin 1989). The benchmark is also simple, concerning only three objects and four relative directions ("left," "right," "below," "above"). The SPARTQA dataset involves more objects and a wider range of spatial relations (including "touching," "far from," "near to"), but also relies primarily on references to non-geographical objects (e.g., "Block B contains one medium black square") rather than a focus on geographical places.

Consequently, for Experiment 1, we generated our own synthetic question corpus with a focus on spatial questions about geographical places. Importantly, the corpus includes questions that require reasoning—not simply retrieval—to answer. The methodology for generating and reasoning with the simulated corpus is explained in the next section, but the designed questions followed the same template introduced in our earlier work in Kazemi Beydokhti *et al.* (2022). Accordingly, there are three basic question types as follows (see Table 1):

- querying the spatial relation between two features, termed "finding relation" (FR);
- querying which feature(s) of a given class that relate to another feature by a particular spatial relation, termed "finding features" (FF); and
- querying whether a feature has a particular spatial relation with another feature, termed "yes/no" (YN)).

| **Code** | **Type** | **Template** | **Example** |
|---|---|---|---|
| FR | Finding relation | What is the spatial relation between ⟨feature⟩ and ⟨feature⟩? | What is the direction of Melbourne from Sydney? |
| FF | Finding features | Which feature(s) is/are ⟨spatial relation⟩ of ⟨feature⟩? | Which states are west of state Queensland? |
| YN | Yes/no | Is ⟨feature⟩ ⟨spatial relation⟩ of ⟨feature⟩? | Is Adelaide east of Melbourne? |

**Table 1.:** Question codes and corresponding templates for each question type, demonstrated with examples (Kazemi Beydokhti *et al.* 2022)

These three question types are a subset of some earlier question corpora, including GeoQuestion201 proposed by (Punjani *et al.* 2018) and SPARTQA by (Mirzaee *et al.* 2021), and most recently, GeoQuestions1089 by (Kefalidis *et al.* 2023).

*Spatial reasoners*

To reason about the answers to spatial questions, we compared a conventional spatial reasoning engine, the well-known and trusted SparQ[1], with the ProbCog[2] MLN engine for probabilistic qualitative spatial reasoning. A full account of using MLNs for probabilistic spatial reasoning is beyond the scope of this paper, and may be found in Duckham *et al.* (2023a). In brief, MLNs are a general framework for probabilistic reasoning, introduced in Richardson and Domingos (2006) and Domingos *et al.* (2008). An MLN is a knowledge base of weighted first-order logical formulae. To enable reasoning, the logical formulae in this knowledge base are mapped to nodes and edges in a Markov random field (MRF). The specifics of this encoding is at the core of both the uniqueness and the complexity of MLNs. With the logical formulae encoded as a MRF, stochastic

names.

[1] `https://github.com/dwolter/SparQ`

[2] `https://github.com/opcode81/Probcog`

sampling methods can be used to drive probabilistic inferences from the knowledge base. To accomplish spatial reasoning, it is necessary simply to translate a spatial logic into the first order logical knowledge base from which an MLN is constructed. For a spatial logic, the knowledge base contains all the qualitative relations and the rules for composition and converse of relations, explored in detail in Duckham *et al.* (2023a).

Using the weights associated with formulae in the knowledge base, MLNs can handle *uncertain* evidence, also explored further in Duckham *et al.* (2023a). In the reasoning tasks in this paper, the weights associated with all formulae in the knowledge base are 1.0 (all formulae are certain). As a result, probabilistic spatial reasoning inferences are always *consistent* with conventional reasoning conclusions. If a conventional reasoning conclusion is logically false, it will have probability zero in the probabilistic reasoner; if the conventional reasoning conclusion is unambiguously true, it will have probability 1.0. However, where conventional spatial logics return an ambiguous result, the probabilistic reasoner provides more detail about the relative likelihoods of the different possibilities.

For example, for three regions $a$, $b$, $c$ such that $a$ is a proper part (PP) of $b$ and $b$ partially overlaps (PO) $c$, a conventional RCC5 reasoner will conclude that the relationship between $a$ and $c$ is one of *either*: proper part (PP), partially overlaps (PO), or discrete region (DR). The probabilistic reasoner likewise concludes that only these three options are *possible* (i.e., non-zero probabilities); but is additionally able to use all the available evidence with the structure of the logic encoded in the MLN to infer a probability associated with each option.

### 3.2. *Methodology*

In generating specific questions to test our approach, we focused on 500 place names in and around the Melbourne central business district (CBD), chosen randomly from the database of Australia's place names, Gazetteer of Australia (This list of places was derived from our previous work in Kazemi Beydokhti *et al.* (2022)).[1]

Next, we randomly selected four key places and computed the root mean square (RMS) of the distances between those places. Sets of places with an RMS below a minimum threshold, chosen manually, were discarded, to ensure a randomly chosen but relatively uniform spatial distribution of key places across the region. The process was repeated until 1000 distinct configurations of key places were generated.

Then, to generate our question corpus, we used the projective cardinal direction calculus (CDC) from Frank (1992), with 9 direction relations (N, S, E, W, NE, NW, SE, SW, and I, the identity relation). The cardinal direction relations between all key places were generated for use as an "evidence database" by implementing the qualification process proposed in Wolter and Wallgrün (2013).

Then, for each of the 1000 configurations, we randomly selected a fifth place $P$ from the gazetteer that was not part of our core set of 500 places. The cardinal direction relation between $P$ and *two* places selected randomly from the four key places in that configuration was also added to the evidence database.

Finally, each of the three types of questions was generated for each configuration, focusing on the relation between place $P$ and one of the two places *not* explicitly stored in the evidence base. In this way, we guaranteed that answering each question required qualitative spatial reasoning, and none of the question answers was stored explicitly in the evidence base. The resulting 3000 questions are freely available in our GitHub repository.

[1] https://placenames.fsdf.org.au/

| aRb↓ / bRc→ | N | NE | E | SE | S | SW | W | NW | I | Converse |
|---|---|---|---|---|---|---|---|---|---|---|
| N | N | NE | NE | $\mathcal{E}$ | $\mathcal{U}$ | $\mathcal{W}$ | NW | NW | N | S |
| NE | NE | NE | NE | $\mathcal{E}$ | $\mathcal{E}$ | $\mathcal{U}$ | $\mathcal{N}$ | $\mathcal{N}$ | NE | SW |
| E | NE | NE | E | SE | SE | $\mathcal{S}$ | $\mathcal{U}$ | $\mathcal{N}$ | E | W |
| SE | $\mathcal{E}$ | $\mathcal{E}$ | SE | SE | SE | $\mathcal{S}$ | $\mathcal{S}$ | $\mathcal{U}$ | SE | NW |
| S | $\mathcal{U}$ | $\mathcal{E}$ | SE | SE | S | SW | SW | $\mathcal{W}$ | S | N |
| SW | $\mathcal{W}$ | $\mathcal{U}$ | $\mathcal{S}$ | $\mathcal{S}$ | SW | SW | SW | $\mathcal{W}$ | SW | NE |
| W | NW | $\mathcal{N}$ | $\mathcal{U}$ | $\mathcal{S}$ | SW | SW | W | NW | W | E |
| NW | NW | $\mathcal{N}$ | $\mathcal{N}$ | $\mathcal{U}$ | $\mathcal{W}$ | $\mathcal{W}$ | NW | NW | NW | SE |
| I | N | NE | E | SE | S | SW | W | NW | I | I |

Note: $\mathcal{N}$ = NE, N, NW; $\mathcal{E}$ = NE, E, SE; $\mathcal{S}$ = SE, S, SW; $\mathcal{W}$ = SW, W, NW; $\mathcal{U}$ = N, NE, E, SE, S, SW, W, NW, I

**Table 2.:** Composition and converse for the projection-based cardinal direction calculus (CDC), after Frank (1992) Ligozat (1998), reproduced from Duckham *et al.* (2023b)

The conventional and probabilistic reasoners introduced above were then used to derive answers to the synthetic question corpus from this evidence base. Specifically, the experiment used the standard projection-based CDC composition and converse operations given in Table 2.

### 3.3. *Results*

Tables 3–5 illustrate the outputs of both conventional and probabilistic reasoning for the three question types for two selected example configurations (C4 and C5). The unknown place $P$ is denoted with place number 501.

| | Query | CR inference | PR inference | Actual relation |
|---|---|---|---|---|
| C4 | (499, 501,?FR) | dir(499, 501, S) ∨<br>dir(499, 501, SE) ∨<br>dir(499, 501, SW) | dir(499,501,SW): 0.242 ∨<br>dir(499,501,NE): 0.000 ∨<br>dir(499,501,I): 0.000 ∨<br>dir(499,501,W): 0.000 ∨<br>dir(499,501,N): 0.000 ∨<br>dir(499,501,SE): 0.751 ∨<br>dir(499,501,E): 0.000 ∨<br>dir(499,501,NW): 0.000 ∨<br>dir(499,501,S): 0.007 | SE |
| C5 | (484, 501,?FR) | dir(484, 501, E) ∨<br>dir(484, 501, NE) ∨<br>dir(484, 501, SE) ∨<br>dir(484, 501, W) ∨<br>dir(484, 501, NW) ∨<br>dir(484, 501, SW) ∨<br>dir(484, 501, N) ∨<br>dir(484, 501, S) ∨<br>dir(484, 501, I) | dir(484,501,SW): 0.252 ∨<br>dir(484,501,NE): 0.000 ∨<br>dir(484,501,I): 0.051 ∨<br>dir(484,501,W): 0.100 ∨<br>dir(484,501,N): 0.047 ∨<br>dir(484,501,SE): 0.000 ∨<br>dir(484,501,E): 0.000 ∨<br>dir(484,501,NW): 0.530 ∨<br>dir(484,501,S): 0.020 | NW |

**Table 3.:** Example outputs for conventional reasoning (CR) and probabilistic reasoning (PR) engines for the "finding relation" (FR) questions in configurations C4 and C5.

For example, for configuration C4 and in response to the question "What is the direction relation between place 499 and 501?"—written (499, 501,?FR)—the conventional reasoner infers that the relation between places 499 and 501 is one of either S, SE, or SW. The probabilistic reasoner, by contrast, computes the likelihood of all nine possible direction relations, with SE being the most probable relation ($p$ =0.751). Note that as highlighted above, the probabilistic reasoning results are always consistent with the conventional reasoner (only relations S, SE, SW have non-zero probabilities), but provide more detail about the relative probabilities (SE is most probably, S least probable).

Figure 1a represents this configuration graphically, along with the inferred probability of each direction relation from the probabilistic reasoner. To aid comparison, likelihoods

have also been associated with all conventional reasoning outputs, calculated as 1 over the number of relations in the returned answer set (i.e., if there are three [equally likely] possible relations returned by the conventional reasoner, each is allocated a probability of 1/3, 0.333; one possible relation is allocated a probability of 1.0; nine possible relations each are allocated a probability of 1/9, 0.111, and so on).

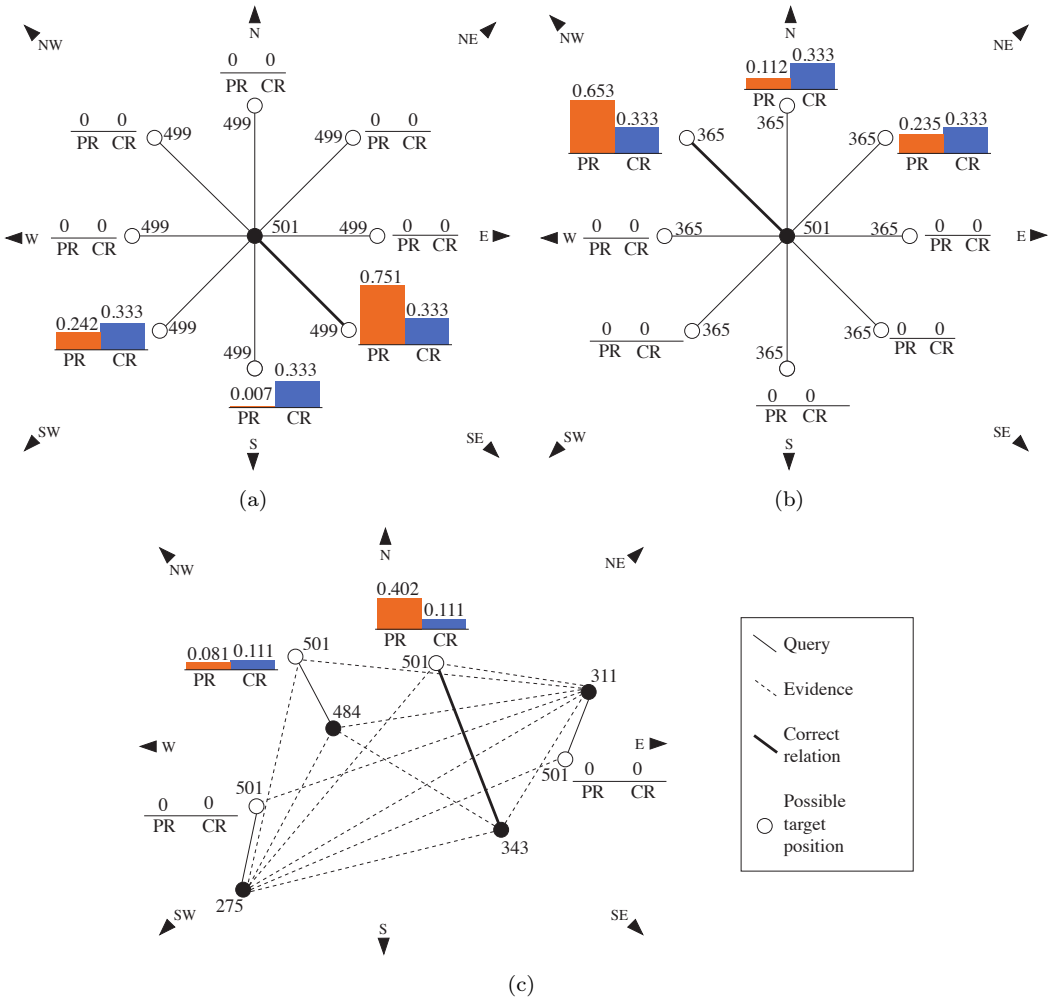



**Figure 1.:** Inference results of conventional reasoning (CR) and probabilistic reasoning (PR) for a) configuration C4 and FR (find relation) with place 499; b) configuration C4 and YN (yes/no) with place 365; and c) configuration C5 and FF (find feature) with place 501.

Similarly, Table 4 shows the reasoning outputs for the example configurations (C4 and C5) and the "finding features" (FF) question (e.g., (?FF, 501, SE), "Which places are SE of place 501?"). In this case, both the conventional and probabilistic reasoners identify the unique answer: place 499. (Note that FF answers do not need to sum to 1, since each probability is in this case independent.) In the case of configuration C5, both places 484 and 343 are equally likely to be SE of unknown place P; while in the probabilistic reasoner, place 343 is calculated to be the most likely ($p$ =0.454) of three possible answers (see also Figure 1c).

Finally, Table 5 presents the reasoning outputs for configurations C4 and C5 and the "yes/no" (YN) question (e.g., (365, 501, NW), "Is place 365 northwest of place 501?"). Fig-

| | Query | CR inference | PR inference | Actual relation |
|---|---|---|---|---|
| C4 | (?FF, 501, SE) | dir(499, 501, SE) | dir(499,501,SE): 0.823 ∧<br>dir(365,501,SE): 0.000 ∧<br>dir(250,501,SE): 0.000 ∧<br>dir(319,501,SE): 0.000 | 499 |
| C5 | (?FF, 501, SE) | dir(484, 501, SE) ∧<br>dir(343, 501, SE) | dir(484,501,SE): 0.081 ∧<br>dir(275,501,SE): 0.000 ∧<br>dir(343,501,SE): 0.402 ∧<br>dir(311,501,SE): 0.000 | 343 |

**Table 4.:** Example outputs for the conventional reasoning (CR) and probabilistic reasoning (PR) engines for the "finding features" (FF) questions in configurations C4 and C5

ure 1b illustrates configuration C4 and its inferred probabilities.

| | Query | CR inference | PR inference | Actual relation |
|---|---|---|---|---|
| C4 | (365, 501, NW) | dir(365, 501, N) ∨<br>dir(365, 501, NE) ∨<br>dir(365, 501, NW) | dir(365,501,NW): 0.653 | YES |
| C5 | (343, 501, SW) | dir(343, 501, E) ∨<br>dir(343, 501, NE) ∨<br>dir(343, 501, SE) ∨<br>dir(343, 501, W) ∨<br>dir(343, 501, NW) ∨<br>dir(343, 501, SW) ∨<br>dir(343, 501, N) ∨<br>dir(343, 501, S) ∨<br>dir(343, 501, I) | dir(343, 501, SE): 0.05 | NO |

**Table 5.:** Example outputs for the conventional reasoning (CR) and probabilistic reasoning (PR) engines for the "yes/no" (YN) questions in configurations C4 and C5

The performance of the probabilistic and conventional reasoners for the three question types is compared systematically below from three distinct perspectives: reasoning accuracy; the distribution of the magnitude of difference in the probability values; and the probability distribution of correct answers.

### *3.3.1. Reasoning accuracy*

In one sense, both reasoners are always accurate because they always generate deductively valid conclusions. However, in cases where multiple different relations are deductively valid, the probabilistic reasoner provides additional information about the relative likelihoods of different possibilities, when compared with the conventional reasoner. Our reasoning accuracy evaluation aims to tease apart these differences, highlighting the ways in which probabilistic and conventional reasoners vary in the level of detail they provide, and the correctness of the additional detail that the probabilistic reasoner generates. To tease apart these differences, we evaluated the accuracy of the generated answers obtained from the conventional and probabilistic reasoners using four classes:

- *Accurate* (A): An answer is accurate if it provides a singleton, most likely relation that matches the ground truth exactly.
- *Incorrect* (I): An answer is incorrect if it comprises a (possibly singleton) set of most likely relations, of which the ground truth is not one.
- *Approximate* (X): An answer is approximate if it contains a set of most likely relations, of which the ground truth is one.
- *Uninformative* (U): An answer is uninformative if it is the universal set of relations, all equally likely.

To help explain these four classes and their interrelationships, Table 6 summarizes the different reasoning accuracy classes and their potential occurrences when using the two

reasoners. In cases where the spatial logic yields a single, unique answer (first column), deductive validity ensures both reasoners are always accurate. For example, in a situation where we know "a is north of b" (aNb), "b is northeast of c" (bNEc), it will always be the case that both reasoners will infer "a is northeast of b" (see Table 2).

| Answer set: | **Unique** E.g., aNb and bNEc implies aNEc | | **Multiple possibilities** E.g., aNb and bSEc implies one of aNEc, aEc, or aSEc | | **All possibilities** E.g., aNb and bSc implies all relations possible | |
|---|---|---|---|---|---|---|
| Reasoner: | CR | PR | CR | PR | CR | PR |
| Accurate (A) | **yes** | **yes** | no | **yes** | no | **yes** |
| Incorrect (I) | no | no | no | **yes** | no | **yes** |
| Approximate (X) | no | no | **yes** | yes* | no | yes* |
| Uninformative (U) | no | no | no | no | **yes** | yes* |

* indicate possible, but unlikely occurrences, infrequent in practice

**Table 6.:** Summary of reasoning accuracy classes, highlighting CR and PR output possibilities

In cases where conventional spatial logic yields all possible answers (the universal set of relations, final column), the CR output will be classed as uninformative. While it is *possible* that the PR will return the universal set of equally probable relations, this is unlikely and very rarely occurred practice in our experiments. Rather, the PR output more normally indicates a single, most likely relation, which may or may not be the actual ground truth relation (i.e., either an accurate answer or incorrect answer, respectively). For example, in a situation where we know "a is north of b" (aNb) and "b is south of c" (bSc), the conventional reasoner can only provide the universal set of possibilities as its answer (Table 2). The probabilistic reasoner, however, usually indicates that one of those possible relations is most likely, and this single top-ranked relation may (accurate) or may not (incorrect) accord with the actual ground truth CDC relation.

Finally, in cases where conventional spatial logic yields multiple but not all possible answers (middle column), the CR output will be classed as approximate. For example, in a situation where we know "a is north of b" (aNb), "b is southeast of c" (bSEc), the possible relations between include northeast (aNEc), east (aEc), and southeast (aSEc) (again, see Table 2). In such cases, the probabilistic reasoner may also generate an approximate answer (i.e., where multiple deductively valid options are equally probable) but this is unlikely and again occurred only rarely in practice. Instead, in most cases the probabilistic reasoner identifies a single, most probable relation from the set, which is then either found to agree with the ground truth (accurate) or not (incorrect).

Figure 2 shows a histogram of the frequency of answers in each of these four categories for each of the three question types, contrasting conventional and probabilistic reasoners. The results immediately highlight a consistent pattern across all question types: the conventional reasoner more frequently generates approximate answers (yellow bars), while the probabilistic reasoner more frequently generates accurate answers (green bars) but at the cost of some incorrect answers (red bars). Further, the conventional reasoner generates a significant number of uninformative answers, while the probabilistic reasoner almost never does.

For example, looking at FR (finding relation) questions, over 70% of generated answers in the probabilistic reasoner are accurate (a single, correct, most probable answer). Only about 20% of answers are similarly accurate using the conventional reasoner, with the majority of CR outputs either approximate (the correct answer is one of several equally likely possibilities) or uninformative (the universal set of relations was returned as the

answer). However, as a result the probabilistic reasoner also returned almost 30% incorrect answers, where the single most probable answer did not match the ground truth. While this may appear a poor result at first sight, as we shall explore further below, the majority of incorrect answers also tend to be systematically low probability answers. For example, more than 75% of incorrect FR answers were associated with a probability of less than 0.5 using PR.

The pattern is similar for FF (finding feature) questions, although in this case the probabilistic reasoner also generated a proportion of approximate answers. In connection with YN questions, the probabilistic reasoner appears to perform significantly better than the conventional reasoner, by returning accurate (A) answers for over 80% of queries (as opposed to the conventional reasoner, which had an accuracy rate of 23%); but it did generate incorrect answers for almost 20% of all queries. Again, these incorrect answers are systematically biased towards low-probability outputs, with only 6% of answers both classed as incorrect and associated with a probability above 0.5. We explore further this connection between reasoning accuracy and associated probability further in the following analysis.

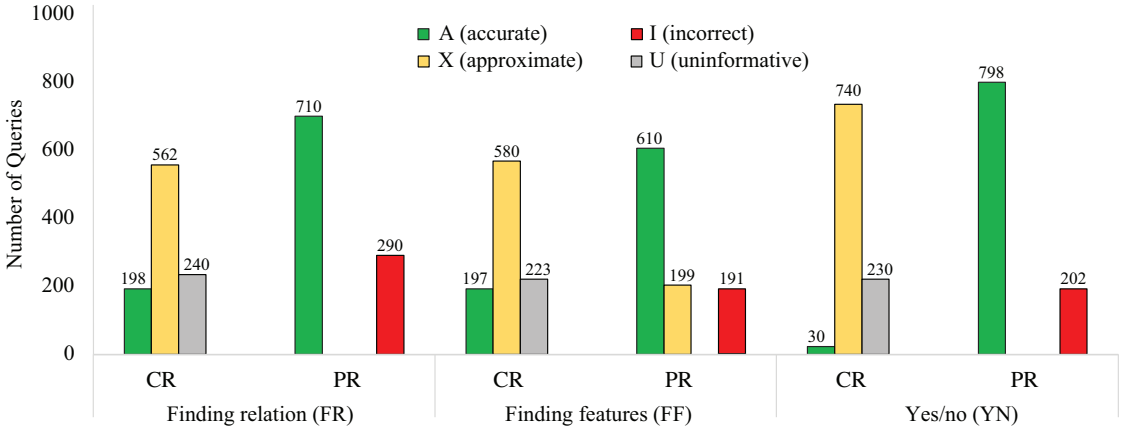



**Figure 2.:** Histogram of answer classes (A, accurate; X, approximate; U, uninformative; and I, incorrect) across the three question types for the conventional reasoner (CR) and the probabilistic reasoner (PR).

### *3.3.2. Distribution of the magnitude of difference in probability values*

The evaluation in the previous section is concerned primarily with a qualitative classification of inferred answers as one of A, X, U, or I. However, using the probabilities associated with the reasoning outputs, it is also possible to evaluate the *magnitude* of difference between inferred conclusions. The magnitude of difference between conventional and probabilistic inferences is calculated as follows:

$$\text{Magnitude of Difference} = P(\text{PR}) - P(\text{CR}) \tag{1}$$

where $P(\text{PR})$ and $P(\text{CR})$ are the probabilities associated with the probabilistic and conventional reasoner outputs for the same question. A positive difference indicates the extent to which the probabilistic reasoner provided more certainty than did the conventional reasoner, while a negative difference indicates the extent to which the conventional reasoner indicates greater certainty in an answer.

To compare reasoning performance in terms of the magnitude of difference between the two reasoners, Figure 3 shows a horizontal histogram ("population pyramid") of the distribution of the magnitude of difference (y-axis) against the number of answers (x-axis). Negative magnitudes of difference (better performance by the conventional reasoner) for the correct (ground truth) answers appear in blue to the left; positive magnitudes of difference (better performance by the probabilistic reasoner) appear to the right in orange.

Taking Figure 3a as an illustrative example, in 13 answers, the conventional reasoner yielded higher probability answers with a magnitude of difference between 0.05 to 0.1, whereas the probabilistic reasoner yielded a higher probability for 88 answers within the same magnitude of difference range.

The three figures show a clear pattern: that the magnitude of differences in correct answers is skewed strongly towards higher frequency and higher magnitudes of difference for the probabilistic reasoner, when compared with the conventional reasoner.

#### *3.3.3. The probability distribution of generated answers*

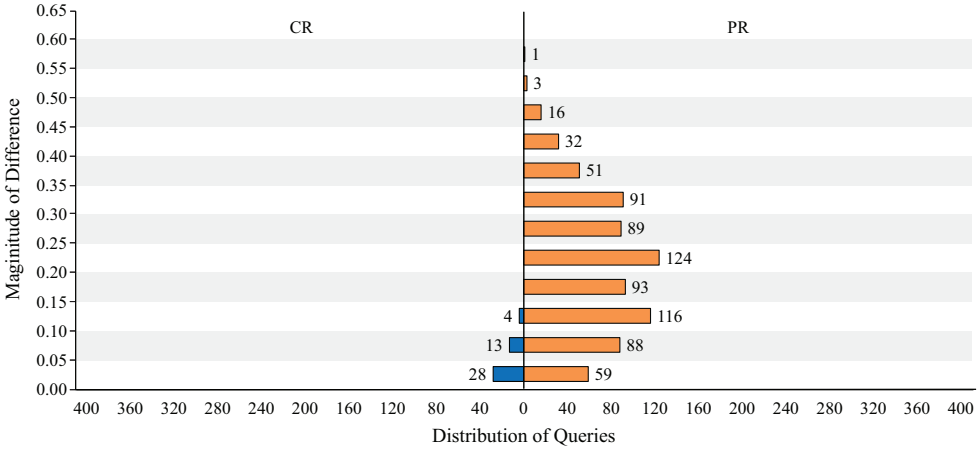
In addition to comparing the conventional and probabilistic reasoners, the probabilities generated natively by the probabilistic reasoner allow a further intrinsic performance evaluation. Figure 4 shows a box-plot of the reasoning probabilities (y-axis) for each accuracy class (x-axis) for the three different question types. The results indicate another clear pattern: correct answers are associated systematically with higher probabilities than are approximate answers, and, more significantly, than incorrect answers.

### 3.4. *Discussion*

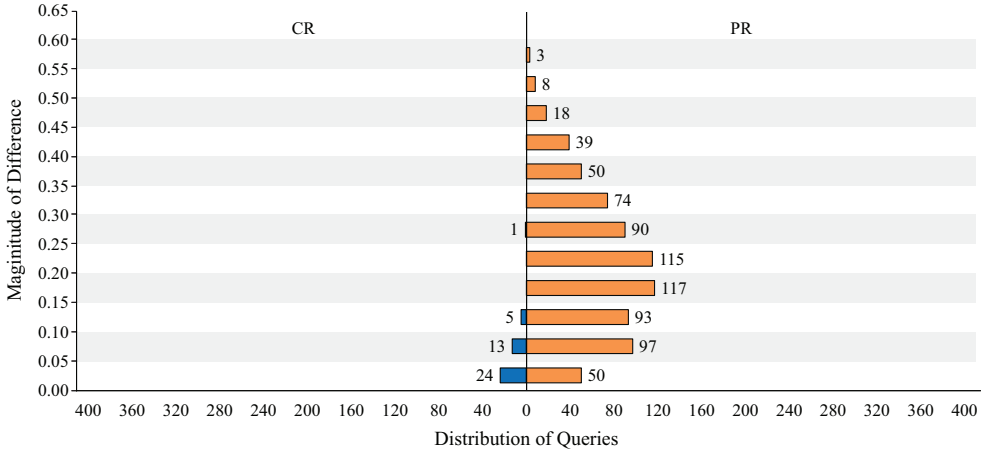
Reflecting on the results we presented above, the comparison between the conventional and probabilistic reasoners can be interpreted to indicate a clear pattern of reasoning performance across the two different reasoning styles as follows.

The probabilistic reasoner produced accurate (A) answers more frequently, at the cost of generating infrequent incorrect (I) answers. The conventional reasoner, on the other hand, never generated incorrect answers, but frequently inferred approximate (A) or uninformative (U) answers. This is the core trade-off in conventional versus probabilistic reasoning: adding detail to approximate or uninformative reasoning outputs also results in occasional errors, where the most probable relation is in actuality not the correct relation. Concomitantly, the guarantee of no incorrectness provided by conventional reasoning is accompanied by systematically less informative conclusions. In terms of GeoQA applications, the most useful reasoning approach is likely to depend on application priorities. In safety-critical and highly automated applications, where inaccuracy carries a high cost, conventional reasoning may be more appropriate; however in most GeoQA applications, with a human in the loop as part of a dialog between the reasoner and operator, the more informative probabilistic conclusions may have greater utility.

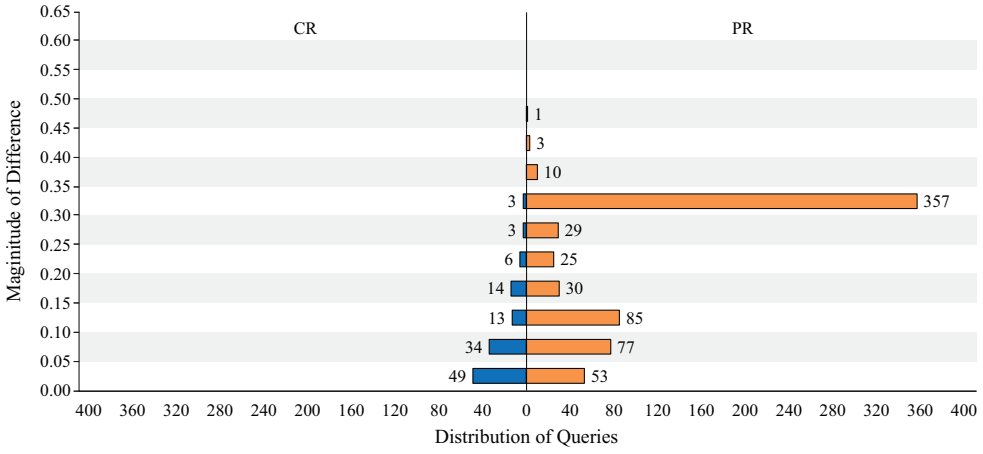
Further, the results in Figure 3 give a picture of how much more confident a probabilistic reasoner can be in its more frequently accurate conclusions. This difference becomes most apparent when considering the tail of the charts, with correct answers with dramatically higher associated probabilities. Almost 30% of the total number of correct probabilistic conclusions to FR (finding relation) and FF (finding feature) questions, for example, have probabilities with a magnitude of difference greater than 0.3 compared to the corresponding conventional reasoning inference. Looking at the distribution of probabilities across all response classes in Figure 4 further underlines this finding, with a clear correlation between higher probabilities and more informative and accurate responses.




(a) FR (finding relation)




(b) FF (finding feature)




(c) YN (yes/no)

**Figure 3.:** Population pyramids showing the frequency of queries by magnitude-of-difference interval for Experiment 1.

## 4. Experiment 2: Questions about region relations

Experiment 1 addressed the extent to which conclusions generated from conventional and probabilistic reasoners are informative and reliable for a simple spatial reasoning sce-

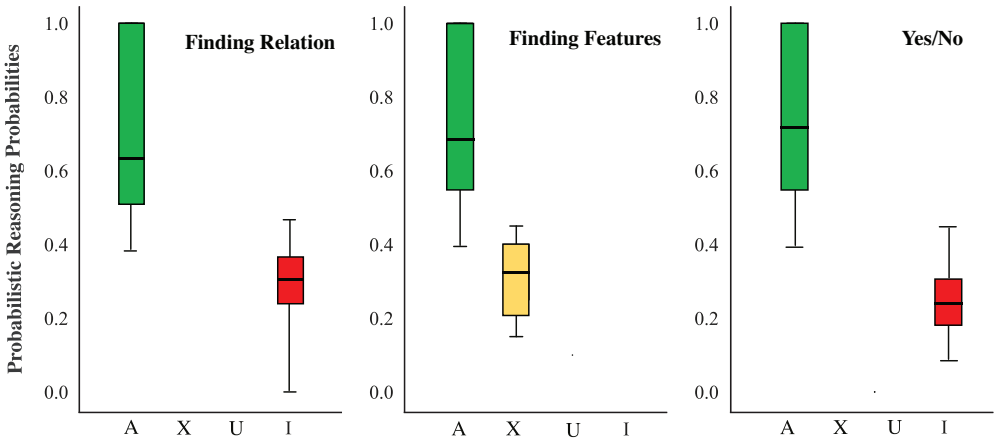



**Figure 4.:** Distribution of reasoning probabilities for each answer category for the three different types of questions for the probabilistic reasoner (PR)

nario based on questions about direction relations using real dataset with synthetically generated questions. Experiment 2 generalizes these results by examining performance on another qualitative spatial calculus, RCC, and validates them using toponym footprints from the United Kingdom. Experiment 2 employs the well-known RCC5, with the composition and converse operations provided in Table 7 (Bennett 1994). While the choice of spatial logic is somewhat arbitrary, RCC5 was selected for these experiments in part due to its relative simplicity when compared to other options. For example, RCC8 introduces three additional qualitative relations that refine relations in RCC5, i.e., distinguishing non-tangential proper part (NTPP) and tangential proper part (TPP) and their inverses from undifferentiated proper part (PP) and its inverse in RCC5; and distinguishing disconnected and externally connected regions (DC and EC) from undifferentiated discrete regions (DR) in RCC5. Indeed, the less detailed RCC5 is arguably more salient in many practical applications without the finer region-boundary distinctions introduced by RCC8.

| aRb↓ / bRc→ | EQ | DR | PO | PP | PPi | Converse |
|---|---|---|---|---|---|---|
| EQ | EQ | DR | PO | PP | PPi | EQ |
| DR | DR | EQ, DR, PO, PP, PPi | DR, PO, PP | DR, PO, PP | DR | DR |
| PO | PO | DR, PO, PPi | EQ, DR, PO, PP, PPi | PO, PP | DR, PO, PPi | PO |
| PP | PP | DR | DR, PO, PP | PP | EQ, DR, PO, PP, PPi | PPi |
| PPi | PPi | DR, PO, PPi | PO, PPi | EQ, DR, PO, PP, PPi | PPi | PP |

**Table 7.:** Composition and converse for RCC5 after Bennett (1994), reproduced from Duckham *et al.* (2023b)

## 4.1. *Materials*

Experiment 2 was based on reasoning about topological relations between UK toponym footprints, derived from crowd-sourced data. The footprint data set contained over 3000 polygonal footprints representing vernacular usage of corresponding toponyms (generated by the authors as part of an unpublished study on "nearness" in 2014, and previously also

used in Li *et al.* 2015). An example of a handful of footprints from this dataset is shown in Figure 5. The complete footprint data set is contained in the accompanying materials to this paper.

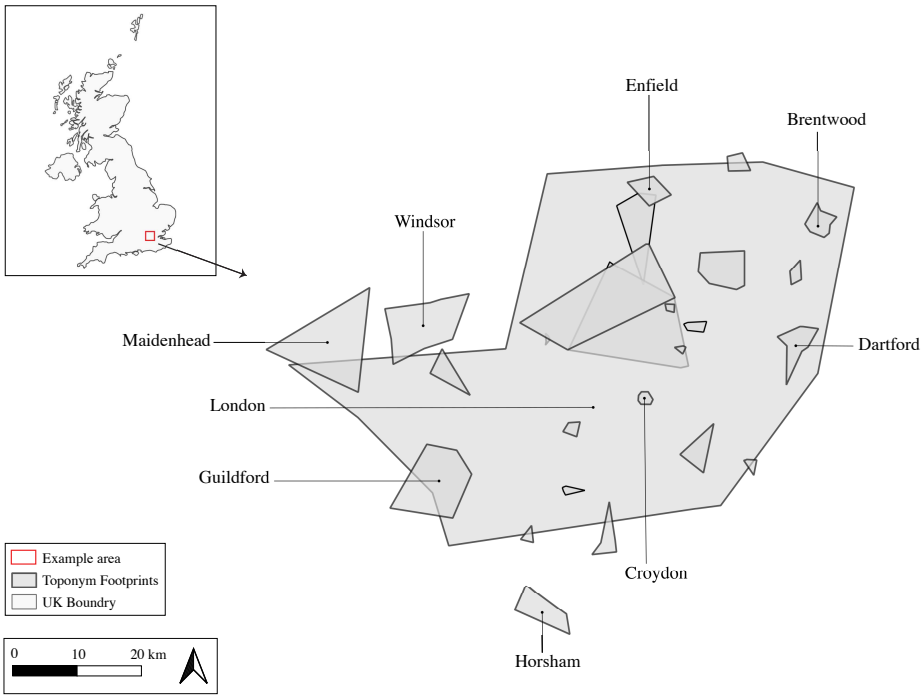



**Figure 5.:** Examples of crowd-sourced geographic toponym footprints in the UK

The footprint data was generated from a combination of the official UK Ordnance Survey gazetteer and the GeoGraph[1] crowd-sourced data set, using the following procedure:

(1) *Spatial data fusion*: Searching through the user-generated descriptions of geolocated photos in the UK GeoGraph crowd-sourcing initiative (see `https://www.geograph.org.uk/`), we extracted all locations labeled as "near $P$," where $P$ could be unambiguously matched with an authoritative Ordnance Survey Gazetteer toponym. The result was almost 90,000 points, relatively evenly distributed across the UK, each labeled by a human GeoGraph user as being "near" a place $P$.
(2) *Footprint generation*: Kernel density estimation was used to filter the fused point data to remove outlier points before generating non-convex footprints for each place $P$ using the non-convex hull "chi-shapes" algorithm in Duckham *et al.* (2008). The result was over 3500 labeled zones of influence (ZoIs or "footprints") from the conflated point sets.

The qualitative, topological relationships between footprints could then be straightforwardly generated, and stored as evidence data sets for qualitative spatial reasoning. For example, from Figure 5, (Windsor, Dartford, DR), (London, Guildford, PO), (London, Brentwood, PPi), (Enfield, London, PP), and so on.

[1] `https://www.geograph.org.uk/`

### 4.2. *Methodology*

Similar to Experiment 1, we implemented an automated approach to generating the same three types of qualitative spatial questions from this dataset. Our algorithm selected four toponyms from the footprint dataset randomly, subject to the constraint that at least half of the topological relations between the corresponding footprints were not disjoint. In our large, national dataset of toponym footprints, four footprints chosen entirely at random are highly likely to all be mutually disjoint, limiting the range of different topological relations reasoning scenarios that we could explore.

The algorithm then randomly generated an unknown point within the combined extents (minimum bounding rectangle) of those selected footprints. The topological relations between the point—called the "incident," echoing one potential application in locating an emergency from a qualitative, natural language description—and two of the selected footprints were also added to the evidence database. Next, the algorithm generated three questions using the question templates already defined in Table 1. Similarly to Experiment 1, all the generated questions required spatial reasoning to be answered. In total, the algorithm generated 1000 different configurations, along with each of the three types of questions. This question corpus is provided in the accompanying materials for this paper. Finally, we provided the generated questions to the conventional and probabilistic reasoners for inference. Figure 6 illustrates the spatial configuration of footprints and the incident point for the first configuration, along with the generated questions for the three question types.

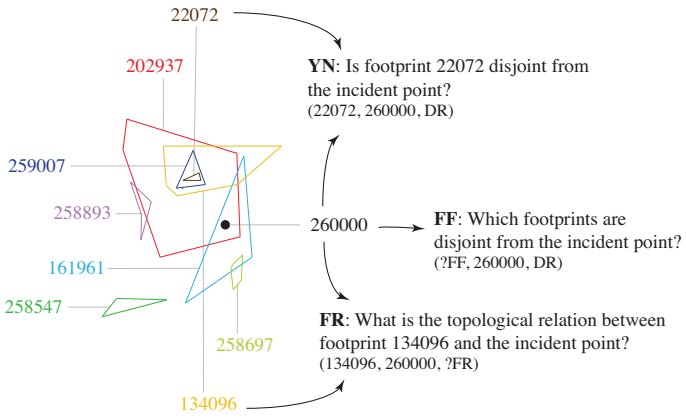
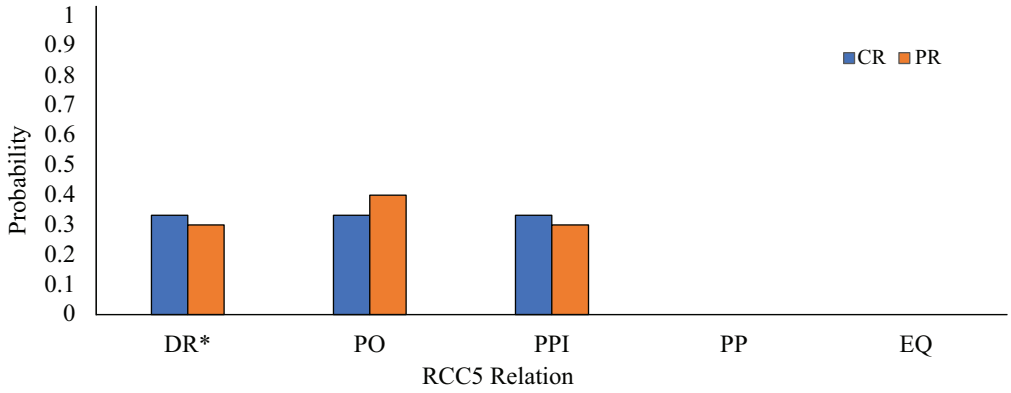



**Figure 6.:** Example of generated questions for FR (find relation), FF (find features), and YN (yes/no) question types for the first configuration (C1). Each footprint is represented by a distinct colored region, with the incident point represented by a black filled circle.

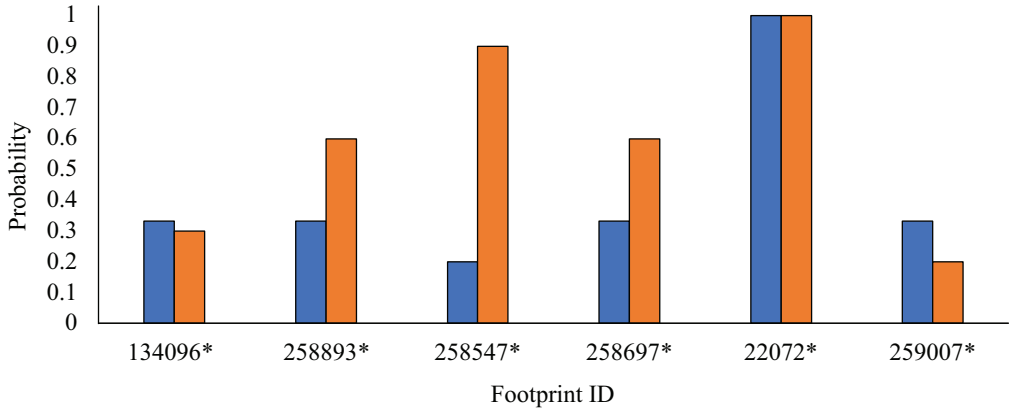
We note that "regions" in region connection calculus are defined axiomatically. Thus, RCC "regions" may include a wide variety of objects in geometry, including points, lines, polygons, volumes, hypervolumes (and indeed non-geometric "regions," such as conceptual regions, see Gärdenfors 2000) as long as they satisfy the basic axiomatic properties of connection in Clarke's calculus of individuals (Clarke 1985), upon which RCC is founded. Thus, RCC can be used directly to represent and reason flexibly about the relations between a mix of geometric objects, including between points and polygons.

### 4.3. *Results*

To aid comparability, we report the results for Experiment 2 in the same manner as Experiment 1. Again, to enable direct comparisons between different reasoning engines, we estimate the likelihood associated with the conventional reasoner's outputs as the inverse of the size of the set of (equally likely) possible relations. Figure 7 illustrates the conclusions inferred from two questions: finding the relation (FR) between footprints 134096 and 260000 (Figure 7a) and finding the regions (features, FF) that are disjoint with place 260000 (Figure 7b). For example, Figure 7a indicates the probabilistic reasoner incorrectly infers PO (partial overlap) as the most likely RCC5 relation between footprints 134069 and 260000, while the conventional reasoner identifies PO, PPi, and DR (the correct answer) as the (equally likely) possibilities. Figure 7b indicates both reasoners agree that footprint 22072 is certainly a discrete region (DR) from footprint 260000, but they differ on the relative likelihoods of five other DR footprints.




(a) Find relation (FR) query (134096, 260000, ?FR)




(b) Find feature (FF) query (?FF, 260000, DR)

**Figure 7.:** Example inferences from the conventional and probabilistic reasoners for two question types (Correct answer(s) marked with *).

#### *4.3.1. Reasoning accuracy*

As for Experiment 1, we evaluated reasoning accuracy with respect to four distinct classes (A, accurate; X, approximate; U, uninformative; and I, incorrect) and provided the frequency of the answers in each of these four classes in Figure 8. The results are

in agreement with what was observed in the first experiment: the probabilistic reasoner more frequently pinpoints the correct answers (i.e., accurate, class A green bars) as a result of generating more detailed probabilities for otherwise approximate (yellow bars) or uninformative (gray bars) conventional reasoning outputs, but at the cost of occasional incorrect answers (class I, red bars).

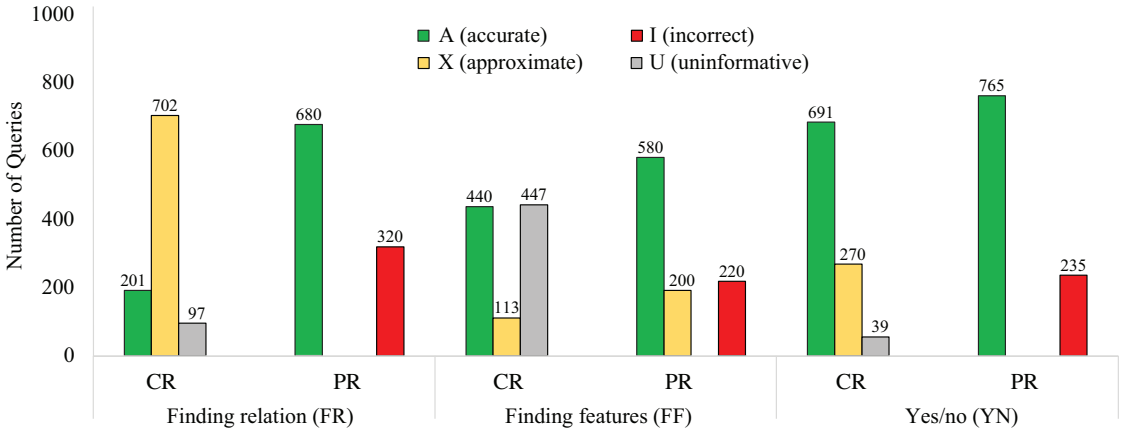



**Figure 8.:** Histogram of answer classes (A, accurate; X, approximate; U, uninformative; and I, incorrect) across the three question types for the conventional reasoner (CR) and the probabilistic reasoner (PR).

In addition, we conducted a comparative analysis of the performance of two reasoners, where the following cases could occur between them:

(1) $P(\text{PR}) > P(\text{CR})$: Where PR assigns a higher probability to the correct answer than the corresponding CR conclusion.
(2) $P(\text{PR}) < P(\text{CR})$: Where PR assigns a lower probability to the correct answer than the corresponding CR conclusion.
(3) $P(\text{PR}) = P(\text{CR}) = 1$: Where both reasoners agree on the correct answer.
(4) $P(\text{PR}) = P(\text{CR}) = 0$: Where both reasoners agree on an incorrect answer.

Figure 9 presents the frequency of reasoning answers across the above four cases. The bar chart includes only DR and PPi relations as they are the only two possible RCC5 relations between a point (the incident) and an areal region (the place footprint). These results indicate that probabilistic reasoning (PR) tends to outperform conventional reasoning (CR) in cases where the correct relation is DR (i.e., reasoning accuracy case 1 above); while the converse is true (CR outperforms DR, reasoning accuracy case 2 above) in scenarios where the correct relation is PPi.

Table 8 summarizes the performance using overall relative accuracy, as:

$$\text{Relative Accuracy}_{\text{PR}} = \frac{N(P(\text{PR}) > P(\text{CR}))}{N(P(\text{PR}) > P(\text{CR})) + N(P(\text{PR}) < P(\text{CR}))} \tag{2}$$

where $\text{Relative Accuracy}_{\text{CR}} = 1 - \text{Relative Accuracy}_{\text{PR}}$.

### *4.3.2. Distribution of the magnitude of difference in probability values*

Similar to Experiment 1 in Section 3.3.2, we evaluated the magnitude of difference between inferred outputs. Figure 10 compares the performance of the two reasoners across the three question types. Contrasting these results with Figure 3, the relative differences

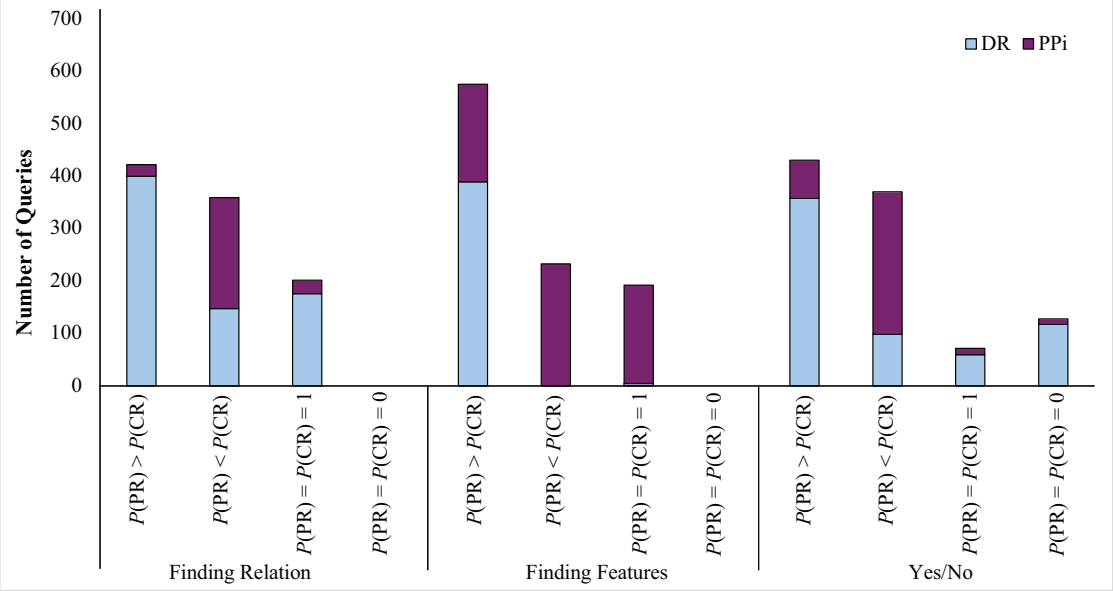



**Figure 9.:** Experiment 2 reasoning accuracy across three question types, highlighting cases where the correct relation is DR versus PPi.

| | **Finding Relation** | | **Finding Features** | | **Yes/No** | |
|---|---|---|---|---|---|---|
| | **PR** | **CR** | **PR** | **CR** | **PR** | **CR** |
| **Relative Accuracy** | 55% | 45% | 71% | 29% | 55% | 45% |

**Table 8.:** Experiment 2 relative accuracy for probabilistic and conventional reasoning across three question types.

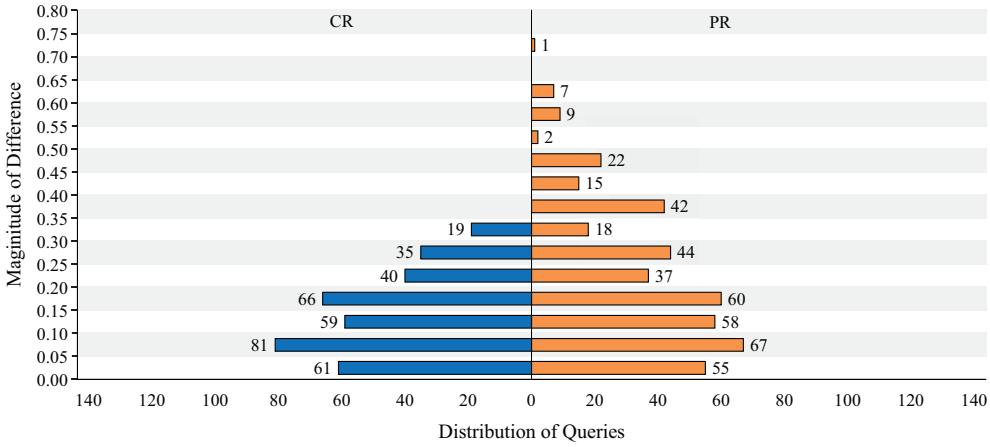
in PR and CR performance are less pronounced compared to Experiment 1. However, a similar pattern—of a higher proportion of correct conclusions with higher magnitudes of difference generated by probabilistic reasoning—is still indicated in the Experiment 2 results.

### *4.3.3. The probability distribution of generated answers*

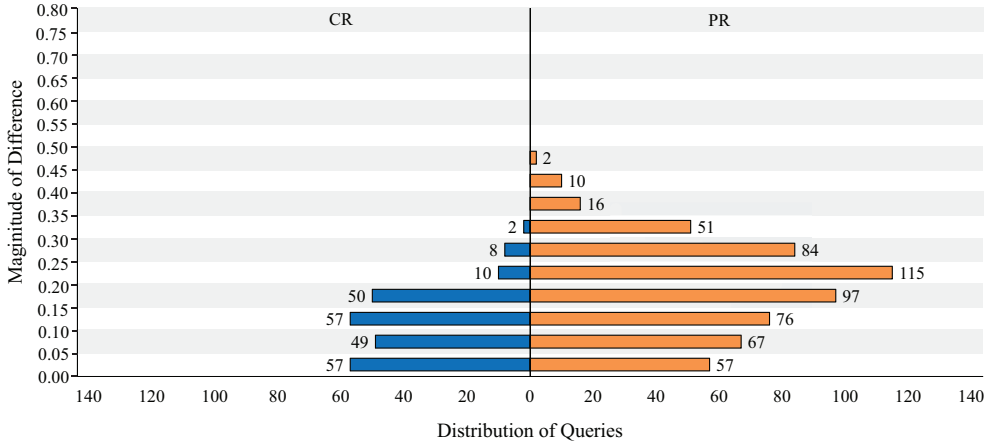
Finally, we evaluated performance of the probabilistic reasoner in Figure 11, similarly to Experiment 1 in Section 3.3.3. Figures 11(A–C) show box plots of the probability distributions of correct answers for the three question types. Figures 11(D–F) also decompose those distributions by the correct relation, either DR and PPi again highlighting the difficulties the probabilistic reasoner has with PPi in our experimental scenarios.

## 4.4. *Discussion*

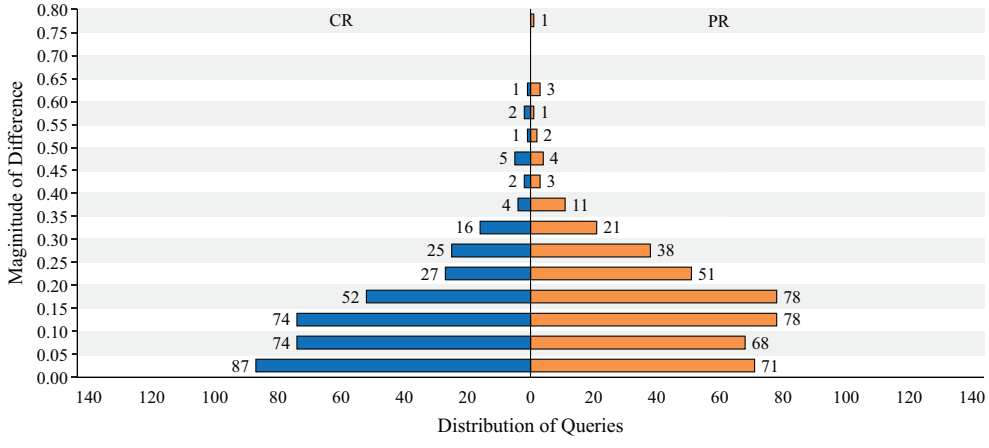
While the first experiment showed a clear advantage for probabilistic reasoner performance compared with the conventional reasoner, the differences between the two reasoners in Experiment 2 were less pronounced. For example, the probabilistic reasoner performed with much higher relative accuracy than the conventional reasoner for FF (finding features) questions. However, the probabilistic reasoner exhibited only marginally higher accuracy than conventional reasoner for FR (finding relation) and YN (yes/no) questions (see Table 8 and Figure 9). Similarly, Figure 10 indicates similar magnitudes of difference for both reasoners in response to YN (yes/no) questions.




(a) FR (finding relation)




(b) FF (finding features)




(c) YN (yes/no)

**Figure 10.:** Population pyramids showing the frequency of queries by magnitude-of-difference interval for Experiment 2.

There are two likely explanations for this relative softening of performance of the proba-

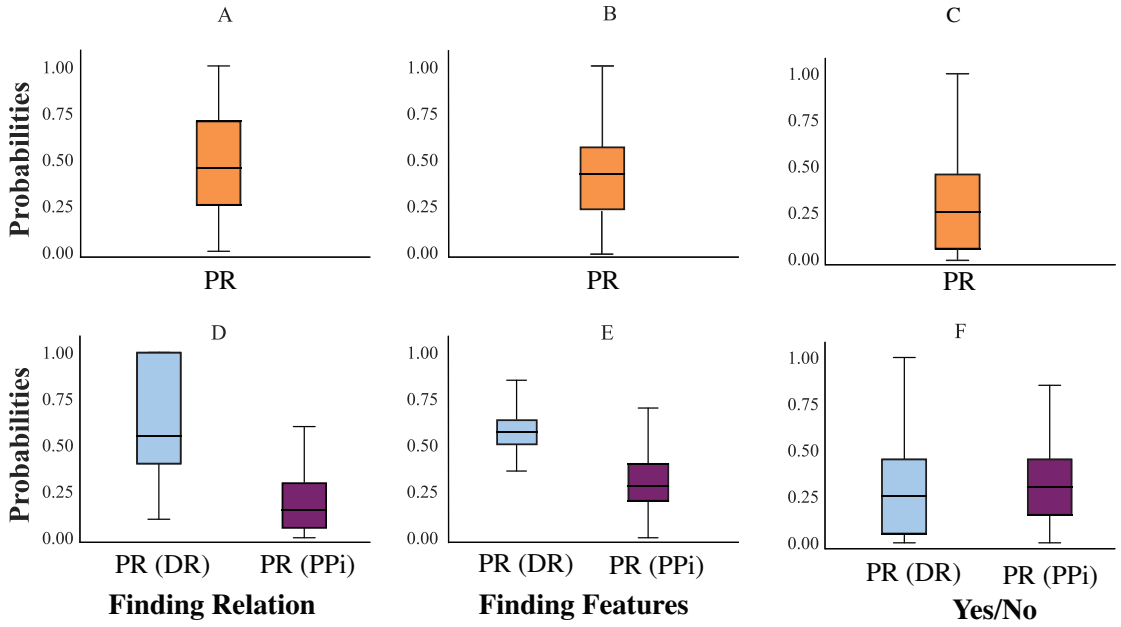



**Figure 11.:** Box plots of probabilities of correct answers for three question types using probabilistic reasoning (A–C) (represented in orange boxes), decomposed by RCC relation (D–F) (DR represented in blue and PPi represented in purple boxes).

bilistic reasoner using real world data. First, being based on real user-generated content, the footprints data is quite "messy," with, for example, many oddly shaped footprints at wildly different scales engaging in unusual and complex spatial relations. It is possible that systematically less-likely relations may occur more frequently in this data set, tending more often to confound the probabilistic reasoner, which operates on the regular structure of the logic, rather than the wildness of unusual geometries.

Second, with only five possible relations in RCC5 (compared with 9 in the cardinal direction calculus in Experiment 1), the relative probabilities associated with the approximate or uninformative answers more frequently generated by the conventional reasoner present a tougher challenge to a probabilistic approach. For example, uninformative answers received a nominal probability of ⅑ (one of 9 possibilities) with the cardinal direction calculus in Experiment 1, but a nominal probability of ⅕ (one of 5 possibilities) with RCC5 in Experiment 2. Hence, our comparison metrics may make it relatively harder for PR to outperform CR on spatial logics that involve smaller sets of relations.

Looking at the probability distribution of correct answers in the probabilistic reasoner in Figure 11 underlines that the probabilistic reasoner also struggles in cases where the correct relation is PPi rather than DR. In addition to the input dataset's inherent messiness, the dataset predominantly comprises DR relations, accounting for approximately 95% of the total relations. This significant imbalance increases the likelihood that the probabilistic reasoner will favor DR over PPi, as PPi is less represented and, therefore, less likely to be correctly inferred with high probability in such an imbalanced dataset.

Nevertheless, the results in Figure 10 do continue to suggest that the probabilistic reasoner retained an edge in performance over the conventional reasoner. For example, the probabilistic reasoner is more accurate in inferring relations when faced with FR (finding relation) and FF (finding features) questions, especially when considering the tail of the charts. Almost 31% of correct inferences to FR and FF questions by the probabilistic reasoner have probabilities with a magnitude of difference to the corresponding conventional

reasoner of greater than 0.3. This edge in less ambiguity and greater certainty in correct answers does still come with the cost, though, of occasional incorrect inferences with the probabilistic reasoner. There were cases in which the probabilistic reasoner assigns the highest probability to an incorrect answer (Figure 7a as an example). The conventional reasoner maintains deductive validity and never generates incorrect answers, but does generate more frequent approximate or uninformative answers. Consequently, in the 70 out of 1000 FR queries where the most probable relation was not correctly inferred by the probabilistic reasoner, the conventional reasoner generated only approximate answers in these queries.

## 5. Applications to GeoQA

Before concluding this paper, we end with a brief discussion of the integration of conventional and probabilistic reasoners into a GeoQA system we are developing, called "QSR-QA." This system has three main components as follows, and illustrated in Figure 12.

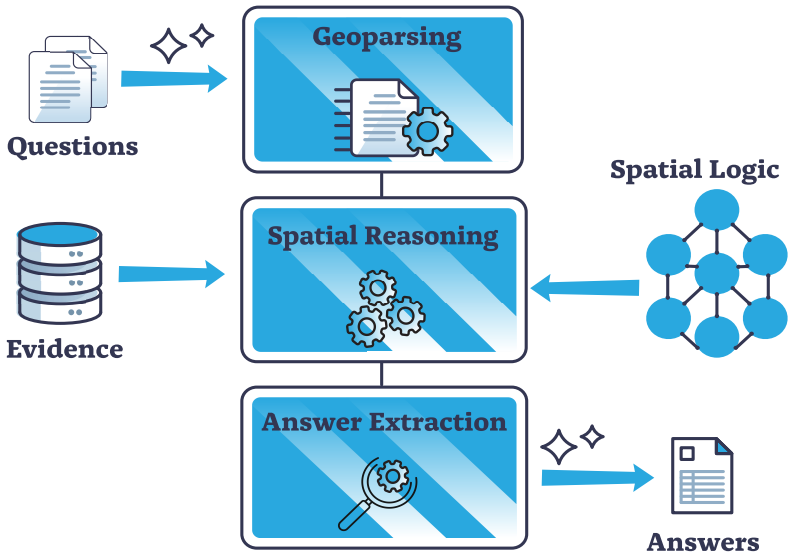



**Figure 12.:** Architecture of our proposed QSR-QA system

The system is composed of three main components:

- *Question processing*: A user's natural language input query may need to first be parsed to extract the embedded qualitative spatial knowledge in a subject-predicate-object representation suitable for reasoning. This is called as the "linguistic conceptual representation" in Kordjamshidi (2013). The next step involves mapping the extracted (subject-object-predicate) triples at the linguistic level into a formal spatial representation (e.g., CDC, RCC). This is called as the "formal semantic representation" in Kordjamshidi (2013). The output of this step translates the unstructured input query into a structured and machine-readable format that are compatible with existing spatial relation logics and reasoning engines.
- *Qualitative spatial reasoning:* Spatial reasoning engines (including the conventional and probabilistic reasoning engines that are the subject of this research) are integrated in this central component. Known spatial relations in various spatial configurations are

stored separately in an evidence database. This stored information must be formatted as triples, consistent with the linguistic level in the question-processing stage. A selected reasoning engine can then be used to infer new spatial knowledge based on the structure of the spatial relation logic and the input evidence.

- *Answer extraction:* The inference results, which are produced in triple format during the reasoning stage, will usually need to be processed to generate natural language answers to the user's input question.

Recent advances in generative AI are assisting greatly with developing this general architecture. While LLMs present inherently low reliability for logical reasoning tasks, they are extremely well-adapted for converting text from one form to another, and especially between human natural language and machine-readable representations and languages. Hence, both the question-processing stage (converting knowledge embedded in natural language spatial questions into machine-readable triples) and the answer-extraction stage involve tasks that are ideally suited to generative AI and LLMs.

A motivating scenario for such use of GeoQA might be identifying an unknown location based purely on natural language questions. Such a GeoQA capability could be useful in responding to an emergency event or criminal activity, with the natural language text perhaps originating from emergency calls or intercepted messages. Linguistic descriptions of an unknown location in such texts (e.g., "north of X" or "near Y") could drive a reasoner, such as those illustrated in this research, to infer more information about the event's location. The balance between accuracy and informativeness highlighted in our experiments may have direct bearing on which reasoning engine should be used. Safety-critical applications, such as emergency response, may have a lower tolerance for incorrect conclusions, and so may rely more on conventional spatial reasoners. Intelligence applications, such as identifying illegal activity, may benefit from the more informative responses produce by probabilistic reasoners, even with occasional inaccuracies.

Like generative AI, however, spatial reasoning today can be computationally challenging, especially in connection with probabilistic reasoning with MLNs (Duckham *et al.* 2023a, Kazemi Beydokhti *et al.* 2022). In our experiments, certain queries, such as finding features (FF), are consistently more computationally demanding, requiring several hours to process, even when using cloud computing resources. Increasing the size of the evidence base also increases computation time. Increasing the number of relations in the logic (e.g., from RCC5's 5 relations, to RCC8, to CDC's 9 relations) also tends to increase computational demands. In fact, the development of MLN reasoning engines today remains at a relatively early stage, with today's MLN engines not yet scalable to even medium sized real-world real world applications. However, work on MLN scalability is progressing, for example, by focusing on a restricted subset of first order logic in the knowledge base (Fang *et al.* 2023); employing a database to keep track of critical weights during the computation (Shin *et al.* 2015, De Sa *et al.* 2016); or using targeted approximation of truth values (Chen *et al.* 2024). Further, MLNs are but one of the leading techniques in the field of statistical relational learning (SRL), dedicated to the broader challenge of probabilistic reasoning Getoor and Taskar (2007). With a large and active community of SRL researchers around the world, and myriad applications including but extending far beyond spatial reasoning, it seems reasonable to look hopefully towards further major progress in this field in the medium term.

## 6. Conclusions

This study explored conventional and probabilistic qualitative spatial reasoning approaches within the domain of GeoQA. The experiments demonstrated how these two types of reasoners perform with simple spatial reasoning tasks using real datasets with synthetic questions, including cardinal direction and topological relations. The results have shown that the probabilistic qualitative spatial reasoner generated correct answers more frequently, but at the cost of occasionally incorrect inferences. Whereas the conventional qualitative spatial reasoner never generated incorrect answers, it provided less informative answers more often.

In both experiments, we used an almost complete evidence set, to ensure sufficient experimental controls enabled us to compare performance againts an unambiguous baseline ground truth. However, this approach should not be taken to imply that such near-complete datasets are required for real-world applications. On the contrary, both reasoning approaches are well-adapted to operating in environments with sparse spatial knowledge. Such capabilities compare favorably with popular machine-learning approaches to the problem, which require voluminous training data required to construct as well as suffering from low levels of explainability and generalizability.

In the context of the rising popularity of deep learning techniques, including LLMs, and their remarkable capacity to handle and emulate human natural language, knowledge-based GeoAI techniques such as conventional and probabilistic spatial reasoning may have fallen in favor in recent years. Nevertheless, we argue in this paper that advances in knowledge-based GeoAI techniques remain fundamental to the geospatial sciences. LLMs remain poorly suited to handling even simplified spatial reasoning, analysis, and logical tasks. Combining the rich variety of knowledge-based GeoAI techniques with the remarkable advances in LLMs, GeoAI "orchestration" holds the potential to leverage the semantic understanding and human-like response generation of LLMs alongside knowledge-based GeoAI reasoners and authoritative data repositories to produce trustworthy answers to geospatial questions.

AI orchestration is already integrated into advanced tools, such as Microsoft's Copilot and Amazon Bedrock, and some recent geospatial technologies with generative AI, such as ESRI's AI assistants for ArcGIS. However, these tools do not yet seamlessly integrate reliable spatial reasoning and analytical capabilities. There remains huge potential for using GeoAI orchestration to generate reliable, human-like answers to user questions in textual format. This could enable decision makers to talk directly with their authoritative spatial data and receive instantly trustworthy answers for sophisticated queries, instead of relying on human data analysts for basic spatial queries.

As such, our research represents a step further toward developing automated spatial reasoning systems, especially in scenarios where GIR techniques in GeoQA and LLMs fall short. On further exploring this vision, several issues emerge as promising areas for future investigation. First, the proposed QSR-QA system could be implemented and compared with existing prototypes. This could be a promising direction for some existing work in GeoQA using GIR techniques, such as Punjani *et al.* (2018) and Li *et al.* (2021), where the proposed systems only retrieved correct answers for 22% and 38% of all questions, respectively. Second, our research thus far has not considered the performance of probabilistic qualitative spatial reasoning with uncertain evidence. As highlighted by Duckham *et al.* (2023a) regarding the capacity of probabilistic qualitative spatial reasoning to use MLN for handling uncertain evidence, future research may benefit from exploring how probabilistic reasoning performs under different levels of uncertainty in the evidence set

within the GeoQA domain.

## Data and codes availability statement

The data and codes that support the findings of this study are available at `https://doi.org/10.6084/m9.figshare.26789650`.

## Acknowledgments

We would like to express our profound appreciation to the RMIT AWS Cloud Supercomputing (RACE) team for providing us with a reliable platform for running computations on AWS, which has contributed significantly to the success of our project. We also thank the two anonymous reviewers whose insightful comments greatly enhanced the quality of this paper.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Notes on contributors

***Mohammad Kazemi Beydokhti*** is a PhD candidate in the School of Science at RMIT University, Melbourne, Australia. He concentrates his research on knowledge representation and reasoning, spatial reasoning in large language models, geographic question answering systems, and spatial knowledge graphs.

***Matt Duckham*** is a professor and Director, Enabling Impact Platform (EIP) at RMIT University, Melbourne, Australia. As the author of one of the leading textbooks in GIS and geospatial computing, Matt brings over 25 years experience in spatial databases, spatial reasoning, and geo AI.

***Amy L. Griffin*** is a senior lecturer at RMIT University, Melbourne, Australia. She is interested in how people use maps and other information visualisations both in everyday life and to support their thinking about real-world spatial problems.

***Yaguang Tao*** is a postdoc researcher at RMIT University, Melbourne, Australia. Yaguang brings experience in movement analytics, cloud computing, machine learning, knowledge graphs, and LLMs.

***Ross Purves*** is a professor of Geographic Information Science at the Department of Geography, University of Zurich. His research interests are Geographic Information Retrieval (GIR), modeling, representing and analyzing geographic information associated with vagueness, ambiguity and uncertainty, and environmental modeling.

***Maria Vasardani*** is a senior consultant, data and analytics at Aurecon, Melbourne, Australia. She used to be as senior lecturer & vice chancellor research fellow in

geospatial at RMIT University, specializing in spatial cognition and natural language processing.

## ORCID

Mohammad Kazemi Beydokhti: https://orcid.org/0000-0003-1887-1560
Matt Duckham: https://orcid.org/0000-0002-7249-6709
Amy L. Griffin: https://orcid.org/0000-0001-6548-7970
Yaguang Tao: https://orcid.org/0000-0002-6335-1452
Ross Purves: https://orcid.org/0000-0003-4301-7024
Maria Vasardani: https://orcid.org/0000-0003-0208-0561

## References

Bender, E.M., *et al.*, 2021. On the dangers of stochastic parrots: Can language models be too big? *In*: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

Bennett, B., 1994. Spatial reasoning with propositional logics. *In*: *Principles of Knowledge Representation and Reasoning*. Elsevier, 51–62.

Chen, S., *et al.*, 2024. Differentiable neuro-symbolic reasoning on large-scale knowledge graphs. *In*: *Proc. 37th International Conference on Neural Information Processing Systems (NeurIPS)*.

Chen, W., 2014. *Developing a framework for geographic question answering systems using gis, natural language processing, machine learning, and ontologies*. Thesis.

Chen, W., *et al.*, 2013. A synergistic framework for geographic question answering. *In*: *2013 IEEE seventh international conference on semantic computing*. IEEE, 94–99.

Clarke, B., 1985. Individuals and points. *Notre Dame Journal of Formal Logic*, 26 (1), 61–75.

Cohn, A.G., 2023. An evaluation of ChatGPT-4's qualitative spatial reasoning capabilities in RCC-8. *arXiv preprint arXiv:2309.15577*.

Cohn, A.G., *et al.*, 1997. Qualitative spatial representation and reasoning with the region connection calculus. *geoinformatica*, 1, 275–316.

Cohn, A.G. and Gotts, N.M., 2020. The 'egg-yolk'representation of regions with indeterminate boundaries. *In*: *Geographic objects with indeterminate boundaries*. CRC Press, 171–187.

Cohn, A.G. and Hazarika, S.M., 2001. Qualitative spatial representation and reasoning: An overview. *Fundamenta informaticae*, 46 (1-2), 1–29.

Cohn, A.G. and Hernandez-Orallo, J., 2023. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of LLMs. *arXiv preprint arXiv:2304.11164*.

De Sa, C., *et al.*, 2016. DeepDive: Declarative knowledge base construction. *ACM SIGMOD Record*, 45 (1), 60–67.

Domingos, P., *et al.*, 2008. Markov logic. *In*: *Probabilistic inductive logic programming*. Springer, 92–117.

Duckham, M., *et al.*, 2008. Efficient generation of simple polygons for characterizing the shape of a set of points in the plane. *Pattern Recognition*, 41, 3224–3236.

Duckham, M., *et al.*, 2023a. Qualitative spatial reasoning with uncertain evidence using

markov logic networks. *International Journal of Geographical Information Science*, 37 (9), 2067–2100.

Duckham, M., Sun, Q.C., and Worboys, M.F., 2023b. *Gis: A computing perspective*. CRC press.

Dylla, F., *et al.*, 2006. SparQ: A toolbox for qualitative spatial representation and reasoning. *In*: *Qualitative Constraint Calculi: Application and Integration, Workshop at KI*. 79–90.

Dylla, F., *et al.*, 2017. A survey of qualitative spatial and temporal calculi: algebraic and computational properties. *ACM Computing Surveys (CSUR)*, 50 (1), 1–39.

Egenhofer, M.J. and Mark, D.M., 1995. Naive geography. *In*: A.U. Frank and W. Kuhn, eds. *Spatial Information Theory: A Theoretical Basis for GIS*, Berlin, Heidelberg. Springer, 1–15.

Fang, H., *et al.*, 2023. MLN4KB: an efficient Markov logic network engine for large-scale knowledge bases and structured logic rules. *In*: *Proc. ACM Web Conference (WWW '23))*. Association for Computing Machinery, 2423—-2432. Available from: `https://doi.org/10.1145/3543507.3583248`.

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

Freksa, C., 1991. Qualitative spatial reasoning. *Cognitive and linguistic aspects of geographic space*, 63, 361–372.

Gärdenfors, P., 2000. *Conceptual spaces: The geometry of thought*. MIT Press.

Gatsoulis, Y., *et al.*, 2016. Qsrlib: A software library for online acquisition of qualitative spatial relations from video. *In*: *Proc. Workshop on Qualitative Reasoning (QR16) at IJCAI-16*.

Getoor, L. and Taskar, B., 2007. Introduction to statistical relational learning. *MIT Press*, 10.

Hamzei, E., Tomko, M., and Winter, S., 2022. Translating place-related questions to geosparql queries. *In*: *Proceedings of the ACM Web Conference 2022*. 902–911.

Hu, Y., *et al.*, 2023. Geo-knowledge-guided gpt models improve the extraction of location descriptions from disaster-related social media messages. *International Journal of Geographical Information Science*, 37 (11), 2289–2318.

Huang, Z., *et al.*, 2019. Geosqa: A benchmark for scenario-based question answering in the geography domain at high school level. *arXiv preprint arXiv:1908.07855*.

Kazemi Beydokhti, M., *et al.*, 2021. Geo-event question answering systems: A preliminary research study. *In*: *Proc. 11th International Conference on Geographic Information Science (GIScience 2021)*, Santa Barbara, United States. UC Santa Barbara: Center for Spatial Studies, 6.

Kazemi Beydokhti, M., *et al.*, 2022. Qualitative spatial reasoning over questions (short paper). *In*: *15th International Conference on Spatial Information Theory (COSIT 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Kazemi Beydokhti, M., *et al.*, 2024. Integrating large language models and qualitative spatial reasoning. *Big Data: Techniques and Technologies in Geoinformatics*, 316.

Kefalidis, S.A., *et al.*, 2023. Benchmarking geospatial question answering engines using the dataset geoquestions1089. *In*: *International Semantic Web Conference*. Springer, 266–284.

Kordjamshidi, P., 2013. Structured machine learning for mapping natural language to spatial ontologies. *Radboud University Nijmegen*.

Kruiger, J.F., *et al.*, 2021. Loose programming of gis workflows with geo-analytical concepts. *Transactions in GIS*, 25 (1), 424–449.

Li, H., *et al.*, 2021. Neural factoid geospatial question answering. *Journal of Spatial Information Science*, (23), 65–90.

Li, S., *et al.*, 2015. On redundant topological constraints. *Artificial Intelligence*, 225, 51–76.

Ligozat, G., 1998. Reasoning about cardinal directions. *Journal of Visual Languages & Computing*, 9 (1), 23–44. Available from: `https://www.sciencedirect.com/science/article/pii/S1045926X97999997`.

Lobry, S., *et al.*, 2020. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58 (12), 8555–8566.

Loerch, U. and Guesgen, H.W., 1997. Qualitative spatial reasoning under uncertainty in geographical information systems. *In*: *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*. IEEE, vol. 3, 2169–2174.

Mai, G., *et al.*, 2022. Towards a foundation model for geospatial artificial intelligence (vision paper). *In*: *Proc. 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, New York, NY, USA. Association for Computing Machinery. Available from: `https://doi.org/10.1145/3557915.3561043`.

Mai, G., *et al.*, 2023. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*.

Mai, G., *et al.*, 2021. Geographic question answering: Challenges, uniqueness, classification, and future directions. *AGILE: GIScience Series*, 2, 1–21.

Mai, G., *et al.*, 2019. Relaxing unanswerable geographic questions using a spatially explicit knowledge graph embedding model. *In*: *International Conference on Geographic Information Science*. Springer, 21–39.

Mirzaee, R., *et al.*, 2021. SpartQA: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2104.05832*.

Nguyen, V., 2019. On combining probabilistic and semantic similarity-based methods toward off-domain reasoning for situational awareness. *In*: *2019 22th International Conference on Information Fusion (FUSION)*. IEEE, 1–8.

Nguyen, V. and Mellor, L., 2020. Fuzzy MLNs and QSTAGs for activity recognition and modelling with RUSH. *In*: *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. IEEE, 1–8.

Petry, F.E., Robinson, V.B., and Cobb, M.A., 2005. *Fuzzy modeling with spatial information for geographic problems*. Springer.

Punjani, D., *et al.*, 2018. Template-based question answering over linked geospatial data. *In*: *Proceedings of the 12th Workshop on Geographic Information Retrieval*. 1–10.

Randell, D.A., Cui, Z., and Cohn, A.G., 1992. A spatial logic based on regions and connection. *KR*, 92, 165–176.

Richardson, M. and Domingos, P., 2006. Markov logic networks. *Machine learning*, 62 (1), 107–136.

Scheider, S., *et al.*, 2021. Geo-analytical question-answering with GIS. *International Journal of Digital Earth*, 14 (1), 1–14.

Schockaert, S., De Cock, M., and Kerre, E.E., 2009. Spatial reasoning in a fuzzy region connection calculus. *Artificial Intelligence*, 173 (2), 258–298.

Shin, J., *et al.*, 2015. Incremental knowledge base construction using DeepDive. *Proceedings of the VLDB Endowment*, 8 (11), 1310–1321.

Tolman, E.C., 1948. Cognitive maps in rats and men. *Psychological review*, 55 (4), 189.

Wallgrün, J.O., *et al.*, 2006. Qualitative spatial representation and reasoning in the SparQ-toolbox. *In*: *International Conference on Spatial Cognition*. Springer, 39–58.

Weston, J., *et al.*, 2015. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Westphal, M. and Wölfl, S., 2009. Confirming the QSR promise. *In*: *AAAI Spring Symposium: Benchmarking of qualitative spatial and temporal reasoning systems*. 41–46.

Wolter, D. and Wallgrün, J.O., 2013. Qualitative spatial reasoning for applications: New challenges and the SparQ toolbox. *In*: *Geographic information systems: Concepts, methodologies, tools, and applications*. IGI Global, 1639–1664.

Younis, E.M., *et al.*, 2012. Hybrid geo-spatial query methods on the Semantic Web with a spatially-enhanced index of DBpedia. *In*: *International Conference on Geographic Information Science*. Springer, 340–353.

Zelle, J.M. and Mooney, R.J., 1996. Learning to parse database queries using inductive logic programming. *In*: *Proceedings of the National Conference on Artificial Intelligence*. 1050–1055.

Zubin, D., 1989. Natural language understanding and reference frames. *In*: D. Mark, A. Frank, M. Egenhofer, S. Freundschuh, M. McGranaghan and R.M. White, eds. *Languages of Spatial Relations: Initiative 2 Specialist Meeting Report*, Santa Barbara, CA. National Center for Geographic Information and Analysis, 13–16.