Predicting Distance and Direction from Text Locality Descriptions for Biological Specimen Collections

Ruoxuan Liao ✉
Massey Geoinformatics Collaboratory, Massey University, Auckland, New Zealand

Pragyan P. Das ✉
Massey Geoinformatics Collaboratory, Massey University, Auckland, New Zealand

Christopher B. Jones ✉
School of Computer Science and Informatics, Cardiff University, UK

Niloofar Aflaki ✉
Massey Geoinformatics Collaboratory, Massey University, Auckland, New Zealand

Kristin Stock[1] ✉
Massey Geoinformatics Collaboratory, Massey University, Auckland, New Zealand

Abstract

A considerable proportion of records that describe biological specimens (flora, soil, invertebrates), and especially those that were collected decades ago, are not attached to corresponding geographical coordinates, but rather have their location described only through textual descriptions (e.g. *North Canterbury, Selwyn River near bridge on Springston-Leeston Rd*). Without geographical coordinates, millions of records stored in museum collections around the world cannot be mapped. We present a method for predicting the distance and direction associated with human language location descriptions which focuses on the interpretation of geospatial prepositions and the way in which they modify the location represented by an associated reference place name (e.g. ***near** the Manawatu River*). We study eight distance-oriented prepositions and eight direction-oriented prepositions and use machine learning regression to predict distance or direction, relative to the reference place name, from a collection of training data. The results show that, compared with a simple baseline, our model improved distance predictions by up to 60% and direction predictions by up to 31%.

2012 ACM Subject Classification Computing methodologies → Machine learning

Keywords and phrases geospatial prepositions, biological specimen collections, georeferencing, natural language processing, locative expressions, locality descriptions, geoparsing, geocoding, geographic information retrieval, regression, machine learning

Digital Object Identifier 10.4230/LIPIcs.COSIT.2022.4

Funding This work was partly funded under the New Zealand Ministry of Business, Innovation and Enterprise Endeavour Fund BioWhere project, and used data collections owned by Manaaki Whenua – Landcare Research New Zealand.

1 Introduction

Around the world, vast collections of biological specimens (e.g. plants, fungi, invertebrates, soil samples) held by museums, libraries and government organisations are georeferenced using text locality descriptions such as *North Canterbury, Selwyn River near bridge on Springston-Leeston Rd*. While specimens collected in recent years usually have geographic coordinates (latitude and longitude) captured from GPS, the locations of many millions of specimens, sometimes going back hundreds of years, are recorded only in text format.

[1] Corresponding author

15th International Conference on Spatial Information Theory (COSIT 2022).
Editors: Toru Ishikawa, Sara Irina Fabrikant, and Stephan Winter; Article No. 4; pp. 4:1–4:15
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Many of the text descriptions used are complex, multi-clausal and consist of a mixture of place names/toponyms (*North Canterbury, Selwyn River, Springston-Leeston Rd* in the above example), generic geographic feature types that appear in the landscape but are not sufficiently notable to have toponyms (*bridge* in the above example), and location terms such as prepositions (*near, on* in the above example) and multi-word phrases (e.g. *4km north-east of, 30 miles along the road from*). Furthermore, many of the descriptions include abbreviations, and while attempts are being made to enforce standard approaches to the description of location to support automated georeferencing [2], many historical records do not follow these standards.

Descriptions of location such as the examples above are referred to as locative expressions [16] and are typically expected to include a located object (or locatum), a spatial relational term or phrase, and a reference object or relatum [16]. In biological records the locatum is sometimes implicit (being the sample that was collected), but in more complex phrases, sub-clauses may include a locatum, such as the word *bridge* in the previous example. When georeferencing textual descriptions of locations, Named Entity Recognition methods which recognise specific types of entities in text, including locations [1], have commonly been used to recognise named places, while gazetteers are used to attach coordinates to (i.e. geocode) the names [10, 18]. However, this only provides part of the picture, as location terms (geospatial prepositions and other modifiers) provide offsets relative to place names (e.g. *next to the Manawatu River*), and thus are key to achieving the level of precision needed to make use of this vast repository of biological data for species mapping and monitoring over time. A number of works have identified common forms of location descriptions, such as distance and cardinal direction (e.g. *4km north-east of <place name>*), defining rule-based [13] and probabilistic [12] models for the interpretation of these common forms, and the errors associated with them. However, the range of possible forms of location descriptions is vast, and many do not conform to these common structures. Furthermore, the interpretation of location descriptions is often context dependent, and the way a description is interpreted many depend on the topography and other physical characteristics, or may vary depending on the methods used for data collection (in part a function of date) or the collector.

In this paper, we present a method for determining the distances and directions associated with a set of location terms (prepositions/prepositional phrases and cardinal directions) using machine learning regression. We predict the offset distance and direction from a reference object associated with the location term using semantic and contextual features of the locality description and the reference object, compare the results to a baseline and evaluate the importance of the features in the model. We demonstrate that regression is a useful tool for predicting the distance or direction associated with location descriptions (depending on the type of description), and that there is scope to further expand this method with additional features and machine learning models. Our method is evaluated using 15311 locality descriptions from the biological collections held by Manaaki Whenua - Landcare Research (MWLR), New Zealand. We identified the most frequently occurring prepositions or nouns/adverb + preposition pairs (e.g. *base of* and *north of* respectively), resulting in eight distance-oriented (*near, above, below, at, head of, end of, mouth of, tributary of*) and eight direction prepositions (*north of, south of, east of, west of, north-east of, north-west of, south-east of, south-west of*). We confined our attention to the last place name in the location description that is preceded by one of those terms. In future work, multiple clauses and place names should be parsed to make use of all the information in the descriptions and further improve results.

The paper is structured as follows: Section Two describes previous work; Section Three presents the method used for extracting relevant terms from the descriptions and calculating the dependent variables (the values we wish to predict, in this case distance and direction) and the features used in our model. Section Four presents the results and evaluates feature importance, and Section Five contains conclusions.

2 Previous work

One of the most common methods for georeferencing text location descriptions is the use of Named Entity Recognition to identify place names, referred to as geoparsing, before extracting their coordinates from a gazetteer, known as toponym disambiguation or resolution [10, 17]. However, this simply georeferences the place names mentioned, but ignores the impact of spatial relation terms that describe a location offset from the place name (e.g. *near the Manawatu River; north-west of Lincoln; outside Auckland*). Such descriptions are known as relative location descriptions, because they describe a location relative to a reference object, and rely on a spatial relation term to do this. Many spatial relation terms are prepositions (e.g. *at, on, near, beside*), though they may be other parts of speech including verbs and adverbs [6].

A number of works have developed models of spatial relation terms, including mapping of terms to spatial relations formally modelled with qualitative spatial reasoning methods (QSR) addressing topological, proximity, orientation and projective relations, e.g. [7, 9, 24]. Such models have only had quite limited application to quantification of distances or angles associated with specific natural language spatial relational terms due to the challenges of interpeting the vagueness inherent in human language.

Several studies have proposed fuzzy logic models of proximity relations and conducted human subjects experiments, including for geographical contexts [26, 31, 8, 11], but such models do not appear to have been applied to the interpretation of natural language texts. A regression model of various forms of nearness and farness that considers several contextual factors, again in a geographical context, was presented in [32] but its application there was to predict the linguistic description for given metric measures. A quantitative analysis of the use of *near* within n-grams was conducted in [5] based on text sources mined from the web. Triples of the located object, spatial relation and reference object were used to examine distances between points of interest within three cities, and between populated places and each of the cities. They found that distances were smaller for *near* in New York compared to San Francisco and Los Angeles, but did not study context specific differences in distances relating to the different feature types. Another study in a similar context analysed the proximity of *close, near* and *next to* spatial relation terms [30] to derive reference object locations. The study discusses the importance of contextual variables like geometry, size and travel distance in deriving coordinates. However, factors like the cardinal direction and angle of the reference object to the located object were not taken into consideration in this study.

Another approach defines spatial templates [21], also known as applicability models or probabilistic density fields, for particular spatial relation terms, that describe the areas in which a term may apply, and depict the variation in applicability for example in the form of a density surface. Thus proximity relation terms such as *near* may be highly applicable at distances close to an object, but gradually become less so as the distance increases. Individual templates can be constructed using multiple examples of observations of the location of a located object relative to a reference object. While most applications of spatial templates have been in table-top space [27, 19, 28], they have also been applied in the geo-spatial domain [14], including for purposes of interpreting location descriptions [13], with models of several proximal and projective relations being instantiated with data from the Geograph photo-sharing web site and human subjects experiments. Various forms of applicability models were used in the study of [3] to infer distances implied by individual spatial relations between places in text describing city locations. They used the gazetteer coordinates of known locations to derive the coordinates of the non-gazetteered places, based on applying their models of the respective spatial relations. The approach was rule based and depended upon the prior existence of a graph ('place graph') representation of the respective locations.

In the context of image analysis and retrieval, spatial templates have been used with machine learning models to predict applicable spatial relations and the locations at which spatial relation terms apply, allowing the context of particular situations to be taken into account. In [22] deep learning methods are used with spatial templates (constructed from multiple examples of, mostly projective, spatial relations) to infer the spatial relations between objects in images. In a related study [4], deep learning is used with spatial templates to infer the coordinates of objects in images that have been described with verbal action relations to a given subject. Their input includes the word embeddings of the subject, the relation and the object, along with the location and size of the subject. Notably the latter study [4] provides an example of using regression methods to infer coordinates relative to a reference object. These studies were both conducted in the context of image analysis and retrieval without reference to geographic space.

This paper differs from the previous work in developing a predictive, machine learning model of a selection of distance and direction oriented terms, and incorporating novel contextual factors in the model. We do not go as far as georeferencing since we only predict either distance or direction, but this does enable the area to which a description refers to be narrowed down to a more precise location relative to the reference place name.

3 Method

3.1 Data

The biological specimen dataset that we use in this paper comes from Manaaki Whenua - Landcare Research (MWLR), New Zealand, and consists of four separate collections, the details of which are listed in Table 1. The MWLR database is constantly updated and maintained, and the MWLR version we used was extracted on July 30, 2021. Table 2 shows some examples of the kinds of locality descriptions that appear in the database. While these collections in combination contain many millions of records, only a small proportion are digitised and have geographic coordinates, and we use a subset of these to train and test our model.

A notable characteristic of the data set is that the coordinates are highly variable in spatial accuracy. The biological specimens in the collections range in age from those collected during Cook's voyages of New Zealand (1769-1779) to the present day. Older specimens may rely on place names that did not continue to be used and whose location has been lost, and textual location descriptions were sometimes very imprecise. Specimens collected in the last few years have been coordinated with GPS, but before that a range of practices were used to derive the coordinates that we used in this work. Some were heavily manual processes involving examination of maps, aerial photos and records to allocate coordinates, but also several automated processes were applied, one example being the use of map sheets recorded as part of the collection record to derive coordinates, using either the centre or a specified corner of the map sheet as an approximation of the location. The result of this is

Table 1 The composition of the MWLR database.

Dataset	**Original number of records**	**Number of records after dropping null values**
Allan Herbarium (plants)	321891	13692
International Collection of Micro-organisms from Plants	22345	909
NZ Fungarium	106945	426
NZ Arthropod Collection	202676	14
Total	**633857**	**15311**

Table 2 Example locality descriptions.

Locality Description	**Latitude**	**Longitude**
Buller, Paparoa Mountains, north flank of Mt Euclid, c. 1-1.5km east of Morgan Tarn.	-41.9562	171.6032
Auckland Island, lower slopes about Musgrave Inlet	-50.6469	166.1533
Nelson, about l km SE of Lake Peel, in the track to Balloon Hut	-41.1316	172.6001
Marlborough, hills about Queen Charlotte Sound	-41.3859	173.7136
Lake Ellesmere Spit = Kaitorete Spit - About Midway along length.	-43.874	172.2679

that the accuracy of individual records in the data set is unknown. To illustrate this point, Appendix A provides scatter plots of eight cardinal direction prepositional phrases studied in this paper, and Table 3 provides figures to indicate the mean bearing (angle from north in a clockwise direction, ranging from 0° to 360°) and standard deviation of each direction. The circular nature of bearing values, where 0° is the same as 360° causes problems for predictive models, and thus following [15], we represent bearings as the (sinθ, cosθ), and the standard deviation figures are given using this representation.

Table 3 Comparison of Cardinal Directions.

Direction	**Mean bearing (°)**	**Standard deviation of cosine of bearing**	**Standard deviation of sine of bearing**	**Mean standard deviation**
north of	11.8	0.57	0.57	0.57
south of	192.6	0.54	0.44	0.49
east of	102.7	0.57	0.54	0.56
west of	263.8	0.47	0.51	0.49
north-east of	38.7	0.47	0.37	0.42
north-west of	307.8	0.50	0.46	0.48
south-east of	142.4	0.51	0.29	0.40
south-west of	223.7	0.44	0.27	0.36

The vague use of cardinal directions in natural language is well documented [14], and demonstrated by the range of locations clustered around the specified direction. However, the scatter plots demonstrate that in this data set, the presence of multiple extreme outliers is much greater than for data analysed in other work, such as [14]. This is likely to be due to inaccuracies in the data set, particularly resulting from methods used to georeference older historical data. Future work will derive and incorporate quality measures into approaches to georeference the collection, but here we work with the data as it is, accepting inaccuracies, as well as some evident gross errors, as part of the challenge that we address. Appendix A also illustrates the tendency for the main four cardinal directions (*north, south, east, west*) to be used for a wider range of directions than the other four (more specific) directions, and this can also been in the lower mean standard deviations for the more specific directions than for (*north, south, east, west*) in Table 3.

3.2 Pre-Processing

Spatial relations are commonly described with prepositions, and for the purposes of this paper, we focus on creating distance or direction models for a set of common spatial prepositions. Having replaced common abbreviations in the descriptions with their expanded versions (e.g. SE -> south-east), we used the spaCy[2] python library part of speech (POS) tagger to identify prepositional phrases as those that were tagged either as prepositions alone (e.g. *near*), or as nouns or adverbs followed by prepositions (e.g. *base of* and *north of* respectively). We then counted the frequency of each unique prepositional phrase to identify the most frequent. We first selected the most frequently appearing eight prepositions, and since some cardinal directions were among this set, we expanded the set to include all eight cardinal directions, and to also include the next most frequent non-directional prepositions to create a balanced set of eight prepositions that describe cardinal directions (which we describe as directional terms), and eight others for which distance is often an important defining characteristic. The final data set consisting of locality descriptions that use one of these sixteen terms formed 78.60% of the 15311 descriptions mentioned in Table 1.

The final set of spatial relation terms were:

- eight directional terms: *north, south, east, west, north-west, north-east, south-west, south-east* and
- eight other spatial relation terms for which distance may be an important component: *near, at, above, below, head of, mouth of, end of, and tributary of.*

While not all of the members of the second set of the terms are primarily distance-related, incorporating elements of elevation (*above, below*) or parthood (*head of, mouth of, end of, tributary of*), in this work we attempt to predict the distances associated with them. Even though *above* and *below* describe elevation, some distance association is implied, as they would not be used with locations that were a large distance from the reference object (e.g. *the hut above Lake Wakatipu*). Similarly, while the parthood terms refer to some specific component of an object (e.g. a river), the parthood relation also implies spatial coincidence. We acknowledge that there are many other semantic aspects of these terms than just the distance, but delay those aspects for future work.

We next used Named Entity Recognition (NER) to identify place names in the locality descriptions, testing several state of the art tools and selecting spaCy's NER tool as the most accurate after testing on a sample of 200 descriptions. We attempted to retrieve coordinates for all place names within a bounding box for New Zealand using three gazetteers: GeoNames[3], the New Zealand Geographic Board Place Names Gazetteer[4] and Nominatim[5], and selected the best result amongst multiple matches (disambiguated) as the place name

[2] `https://spacy.io/`
[3] `http://www.geonames.org/`
[4] `https://gazetteer.linz.govt.nz/`
[5] `https://nominatim.org/`

that was closest to the known location for the specimen. In addition to coordinates, the feature type (e.g. lake, river) was retrieved to support extraction of features for the machine learning model (see Section 3.3).

We identified all instances of the 16 prepositions that were immediately followed by a tagged place name for which we could retrieve coordinates. We then calculated the distance and direction between the coordinates of the place name following the preposition and the ground-truth coordinates contained in the data set. These values represent the offset that describes the location of the specimen relative to the reference place name for simple preposition-place name pairs. For example, for the locality description *near Karangahake Gorge*, the distance between the coordinates of Karangahake Gorge retrieved from the gazetteer and the coordinates of the specimen contained in the collection reflects the quantitative meaning of the *near* preposition in this particular context. These, and their associated directions where prepositions are more direction-related, are the figures we aim to predict with our model.

For the eight distance-related prepositions, we use the geodetic distance between the reference object and the ground truth specimen coordinates as the dependent variable (the value we aim to predict) in our model. For the direction-related prepositions, we follow [15] and use the sine and cosine of the bearing ($\sin\theta$, $\cos\theta$) as the dependent variables for model training (and our regression model for directional prepositions thus has two dependent variables) and convert them back to bearings at the end. This approach is used to avoid problems caused by the circular nature of bearing measurements, in which $0°$ is the same as $360°$.

3.3 Regression Model

In order to predict the distance or direction corresponding to the prepositions in our locality descriptions, we incorporate a number of features in a machine learning regression model. The features included were as follows:

- The GloVe embedding of the **feature type of the reference object**. GloVe (Global Vector for Word Representation) generates multi-dimensional vector representations of words and was first introduced by a team at Stanford University to study the similarity index between the words. It is derived from word-word occurrences in a textual description by only considering the non-zero elements, which are used to calculate the embeddings based on probabilities. We used 200 dimension GloVe embeddings pre-trained on Wikipedia + Gigaword 5 [25]. The feature type for each place name (reference object) was retrieved from the relevant gazetteer along with the coordinates.
- The vector created by averaging the GloVe embeddings for the **feature types of all place names in the locality description** (excluding the reference object) using 200 dimension embeddings pre-trained on Wikipedia + Gigaword 5 [25]. The feature type for each place name was retrieved from the relevant gazetteer alongside the coordinates.
- One hot-encoding of the **geometry type** (point, line, polygon, volume) of the reference object feature type. The geometry type was retrieved from the Linguistically Augmented Geospatial Ontology (LAGO) [29] using WordNet [23] to match feature types retrieved from the gazetteer for our reference object to feature types contained in the LAGO if they did not already appear. This feature is included because geometry type has been shown to influence the use of geospatial prepositions (e.g. *the house beside the church* vs. *the road beside the river* - the latter implies alignment as well as proximity) [29].
- One hot-encoding of the **scale** of the reference object feature type (district scale, neighbourhood scale, immediate scale). The scale was retrieved from the LAGO as for geometry type, and is included because the influence of scale on the use of geospatial prepositions

has been demonstrated [29, 20]. Although not identical, this may be considered an approximate indicator of object size (for example, district scale may refer to objects such as mountain ranges, while immediate scale may refer to smaller objects such as houses).

- One hot-encoding of the census **Territorial Authority** of the reference object, out of a total of 85 districts that cover New Zealand. This set of features indicates whether two instances are in the same geographic area.
- The **area, population, population density and length (at the longest extent) of the meshblock**[6] that the reference object is in. These features indicate how urban/rural an area is, and are included to test whether this aspect influences the use of geospatial prepositions.
- The **area, population, population density and length (at the longest extent) of the Territorial Authority** that the reference object is in.
- The **year** that the specimen was collected. This is an approximate indicator of the accuracy of the coordinates, as recent records have GPS-level accuracy, while coordinates from 200 years ago may be very approximate (e.g. derived from map sheet or description).
- A boolean value indicating whether the location is **cultivated** from the collections data.
- The **altitude** of the specimen, providing an indication of the environment type (e.g. alpine).

We used ten-fold cross validation to test a number of different regression models including Support Vector Machine with polynomial (SVM-polynomial kernel) and Radial Basis Function kernel (SVM-rbf kernel), k-nearest neighbour, gradient boosting, support vector regression and decision tree. A number of other models including linear regression and multi-layer perceptrons were tested but did not perform well so were not pursued further.

4 Results

4.1 Distance Prediction

We evaluate the results of our methods against a simple baseline that relies only on the place name immediately following the preposition (the relatum), and like most current approaches ignores the spatial relation term. Hence it assumes that the distance and direction between the place name and the predicted location are zero. Table 4 shows the mean absolute error for the baseline and our best-performing machine learning model, together with the percentage improvement that our method provides over the baseline.

Overall, we see better performance (larger percentage increase) for cardinal directions than for the distance-related prepositions. While this is unexpected in that we would expect more consistency in distance for distance-oriented prepositions than direction-oriented, this may be explained by the larger average distances between relatum and locatum for the direction-oriented prepositions, so the baseline, which assumes a distance of zero, is a poorer estimate than for prepositions that are used for smaller distances between relatum and locatum. However, following this reasoning, we might expect that the poorer result for the *at* preposition could be equivalently explained by typically shorter distances between relatum and locatum, which are better predicted by our zero-distance baseline, but this is not supported by the data. Our goal is to predict the distance between relatum and locatum and, as the baseline predicts this distance to be zero, each baseline prediction is equal to the actual distance between relatum and locatum. Thus the mean absolute error (MAE)

[6] The smallest geographic unit for which New Zealand census data is recorded.

Table 4 Results of Machine Learning Regression - Distance Prediction.

Preposition	Best-performing Model	Count	Baseline	Regression	
			MAE (m)	MAE (m)	% improv
near	svm-rbf kernel	3478	6412	4491	30%
above	svm-rbf kernel	695	3581	2634	26%
head of	svm-rbf kernel	388	5298	3465	35%
below	svm-rbf kernel	278	4256	3462	19%
at	svm-polykernel	208	6075	5140	15%
end of	svm-rbf kernel	164	5678	4512	21%
mouth of	svm-rbf kernel	115	1630	967	41%
tributary of	svm-rbf kernel	112	5347	3934	26%
north of	svm-rbf kernel	1309	7277	4343	40%
south of	svm-rbf kernel	1211	8538	4509	47%
east of	svm-rbf kernel	959	7458	3862	48%
west of	svm-rbf kernel	879	8533	5054	41%
north-east of	svm-rbf kernel	187	10139	4701	54%
south-west of	svm-rbf kernel	169	9756	3892	60%
north-west of	svm-rbf kernel	147	6116	3212	47%
south-east of	svm-rbf kernel	185	9802	4523	54%

for the baseline is equal to the average distance between relatum and locatum across all instances of a particular preposition. The baseline MAE (and therefore the average distance between relatum and locatum) for the *at* preposition is in fact higher than for all other distance-related prepositions except *near*. Furthermore, the *mouth of* preposition has the shortest average distance (baseline MAE) between relatum and locatum, but is the best predicted of the distance-oriented prepositions using our method. The *mouth of* preposition describes a wide range of distances between 85 and 13200 metres.

Although the regression models show improvement relative to the baseline across all of the prepositions, and in many cases these are substantial, we consider that the MAE values are inflated by outliers that result from the low accuracy of some of the coordinates, and in some cases challenges in identifying accurate coordinates for the relatum place names due to their absence from, or duplication in, the gazetteers. For example, Table 5 shows that 80% of the error values (absolute value of predicted - actual distance) for the *near* preposition are below 5514.45m. Thus filtering out of the worst 20% of errors results in a MAE of 1672.09m.

Table 5 Errors for each Percentile for the *near* preposition.

Percentile	Error
10th	162.56
20th	422.68
30th	768.50
40th	1208.30
50th	1794.13
60th	2612.76
70th	3806.46
80th	5514.45
90th	10038.38
100th (all values)	108772.75

Table 6 Results of Machine Learning Regression – Direction Prediction.

Preposition	Best-performing Model	Count	Baseline	Regression	
			MAE (°)	MAE (°)	% improv
north of	gradient boosting regressor	1309	48.0	47.7	0.6%
south of	gradient boosting regressor	1211	35.3	30.4	13.8%
east of	k-nn	959	46.1	42.3	8.1%
west of	k-nn	879	43.4	41.8	3.5%
north-east of	decision tree	187	35.0	26.7	23.8%
south-west of	gradient boosting regressor	169	43.5	30.2	30.6%
north-west of	support vector regression	147	46.2	48.9	-5.9%
south-east of	support vector regression	185	48.3	40.0	17.2%

4.2 Direction Prediction

For the eight direction-oriented prepositions, we evaluated the ability of our machine learning model to predict direction using the features listed in Section 3.3.

While the cardinal directions technically describe precise directions (e.g. *east of* specifies 90° using north as 0° and measuring angle in a clockwise direction, an angular measurement known as the *bearing*), research has shown that these directions are frequently used vaguely in natural language [14] to refer to a range of directions that are more or less in the direction. As a result of this tendency to use direction terms vaguely, rather than defining our baseline as the precise direction that corresponds to each term, we instead use the average deviation from the precise direction specified by the direction term. We thus use the average difference between the bearing of the actual line between relatum and locatum and 90° as the baseline for *east*, and evaluate the ability of our regression model to predict that difference.

As explained in Section 3.2, we represent these angles as two numbers: sinθ, cosθ, and perform a multivariate regression to predict both values simultaneously. Table 6 presents the results for the direction prediction, with the sinθ, cosθ values converted back into errors in degrees and compared to the baseline. This means, for example, that if we simply assumed that the preposition *south of* means a bearing of 180°, we would get a MAE of 35.3° using our dataset. However, the use of our regression model reduces this MAE to 30.4°, giving a 13.4% improvement. It must be acknowledged that the MAE for both the baseline and the regression model are relatively high. The high MAE for the baseline is an indication of the large spread of directions for which a given cardinal direction term is used, in some cases deviating substantially from the precise direction indicated by the term (e.g. 90° for *east*), as indicated in Appendix A, and while the regression model improves on the baseline by up to 30%, we anticipate that improvements could be achieved by the inclusion of additional features that focus on directional semantics. It is also of note that the regression model for *north-west of* predicts direction less well than if the precise direction were used (315°). This is most likely because the spread of data points for *north-west of* is relatively narrow compared to the other directions, with few outliers, and thus the regression models ability to model contextual variations in the use of the direction term is less effective.

4.3 Feature Importance

We analyse the importance of different features in the model by calculating the correlation coefficient between each feature and the three dependent variables (distance, sinθ and cosθ). Figure 1 shows the 25 features with the highest correlations. The light grey area indicates the importance of the GloVe embeddings for the relatum feature type across all of the distance and direction predictions. The dark grey squares represent the average embedding of the feature types of all other place names, as well as feature types mentioned explicitly in the descriptions (e.g. *pit in back paddock*), and are also important.

Geometry type is among the most highly correlated features with both distance and direction. For example, the boolean point geometry feature is negatively correlated with distance for *near* while the line geometry is negatively correlated with distance for *tributary of*. This means that expressions with point reference objects are more likely to be used for short distances than for other geometry types. The most common point reference object (which is also classified as a polygon reference object) is a small populated place or locality, and it is not unexpected that these would be referenced when closer to the specimen collection location than when further away, in contrast to non-point objects, which are likely to be larger in scale. Although they also appear among the 25 most correlated features with the two direction-related dependent variables, geometry type features do not exhibit a consistent pattern across multiple cardinal directions, with the line geometry boolean feature being positively correlated for *west of* and *south of*, negatively for *north of* and *south-east* of and not for the others.

The scale features are not strongly correlated with distance, but do appear in several of the cardinal directions, although the direction of the correlation (positive or negative) varies. The importance of territorial authorities (being the most highly correlated feature for distance for *head of* and *at* and for cosθ for *north-west* and *south-east*) indicates a geographic pattern in the way that geospatial prepositions are interpreted. There are 85 territorial authorities throughout New Zealand, with areas ranging from 19 to 29,552 square kilometres, and while some are very small and urban in nature, many cover widely varying terrains and environments including a single authority covering all of Fiordland and much of Southland. The meshblock geometry characteristics (length, area and to a lesser extent population) shown in yellow, and those of the territorial authorities are also important for some of the prepositions, as is altitude (shown in brown, along with year and cultivation).

5 Conclusion

In this paper we used regression to predict the distance and direction associated with 16 prepositions. We demonstrated that regression is a useful tool for predicting the distance associated with location descriptions, with improvements for distance-oriented prepositions of up to 41%, and for direction-oriented prepositions of up to 60%. We also showed the significant impact of outliers in this data set, highlighting the need to consider accuracy in these kinds of biological collections data sets that contain historical records. Results for prediction of direction (bearing) were less promising, with the best result showing an improvement of 31% for *south-west of* (the preposition that yielded the best direction prediction results). We also evaluated the importance of the features used in the model through correlation with the dependent variables, showing that relatum feature type is very important, but a range of other features also contribute, such as territorial authority, geometry type and scale.

In order to further improve the results of these models, future work will derive and incorporate spatial data accuracy measures so that greater weight is given to the coordinate data that is known to be accurate. In addition, we will explore more advanced methods for identifying place names that relate to specific prepositions, and for disambiguating place names. In future work we also plan to add further contextual features to the models, and to apply transformer-based neural network approaches such as BERT to the challenge.

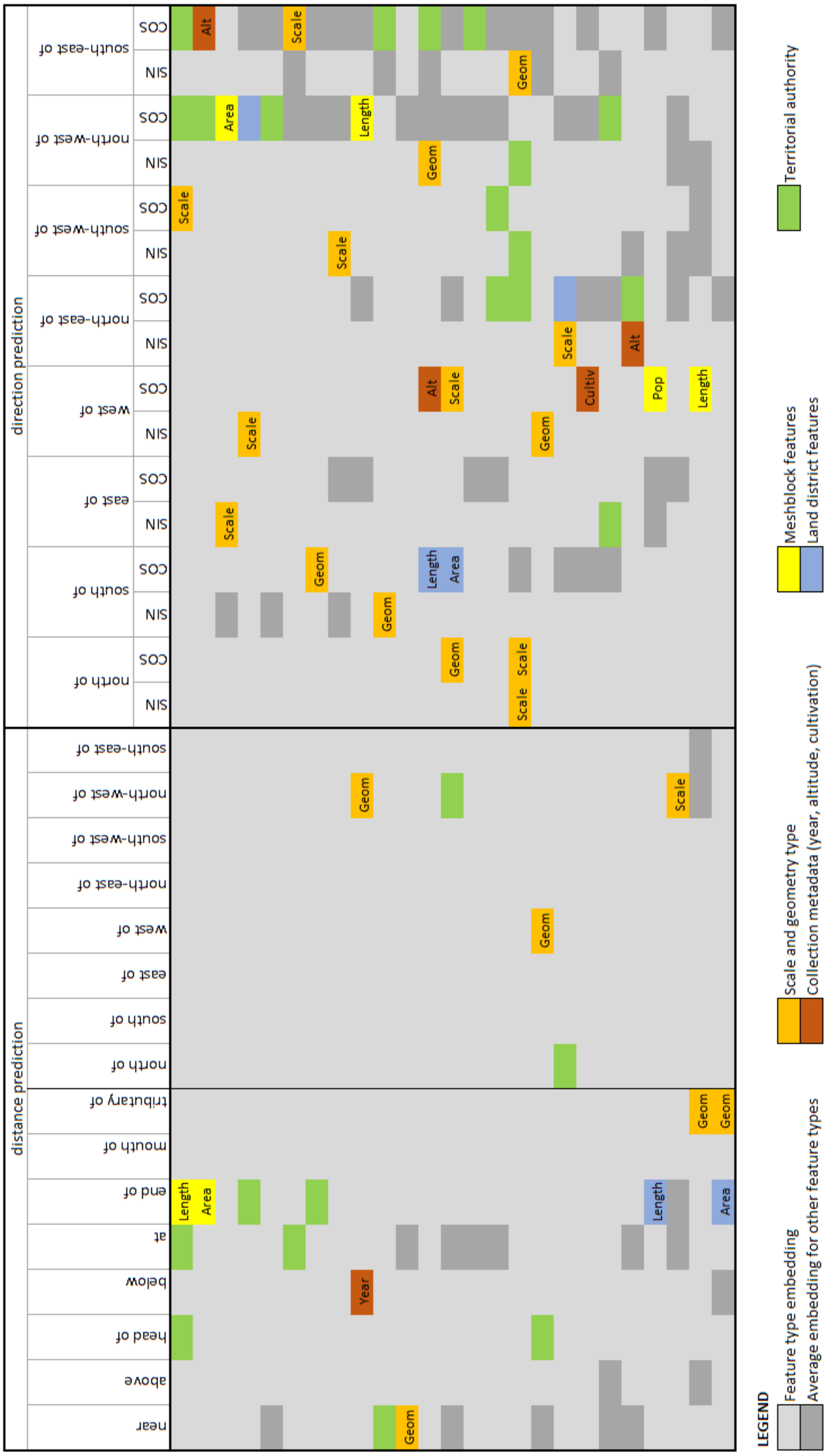


Figure 1 Feature Correlation with Dependent Variables.

References

1 S. Atdağ and V. Labatut. A comparison of named entity recognition tools applied to biographical texts. In *2nd International Conference on Systems and Computer Science*, pages 228–233, August 2013.

2 Arthur D Chapman and John R Wieczorek. *Georeferencing Best Practices.* GBIF Secretariat, Copenhagen, 2020. doi:10.15468/doc-gg7h-s853.

3 Hao Chen, Stephan Winter, and Maria Vasardani. Georeferencing places from collective human descriptions using place graphs. *Journal of Spatial Information Science*, 0(17):31–62, 2018.

4 Guillem Collell, Luc Van Gool, and Marie-Francine Moens. Acquiring common sense spatial knowledge through implicit spatial templates. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

5 Curdin Derungs and Ross Purves. Mining nearness relations from an n-grams web corpus in geographical space. *Spatial Cognition and Computation*, 16, October 2016.

6 André Dittrich, Maria Vasardani, Stephan Winter, Timothy Baldwin, and Fei Liu. A classification schema for fast disambiguation of spatial prepositions. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on GeoStreaming*, pages 78–86. ACM, 2015.

7 M.J. Egenhofer. Reasoning about binary topological relations. In *Second Symposium on Large Spatial Databases*, volume 525 of *Lecture Notes in Computer Science*, pages 143–160. Springer-Verlag, 1991.

8 M. Gahegan. Proximity operators for qualitative spatial reasoning. In *Spatial Information Theory A Theoretical Basis for GIS*, pages 31–44. Springer Berlin / Heidelberg, 1995.

9 K.P. Gapp. Angle, distance, shape and their relationship to projective relations. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 112–117, 1995.

10 Milan Gritta, Mohammad Taher Pilevar, and Nigel Collier. A pragmatic guide to geoparsing evaluation: Toponyms, named entity recognition and pragmatics. *Language Resources and Evaluation*, 54, September 2019.

11 Hans W. Guesgen. Reasoning about distance based on fuzzy sets. *Applied Intelligence*, 17:265–270, 2002.

12 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(10):1067–1090, 2008.

13 Mark Hall, Philip Smart, and Christopher B. Jones. Interpreting spatial language in image captions. *Cognitive processing*, 12(1):67–94, 2011.

14 Mark M. Hall and Christopher B. Jones. Generating geographical location descriptions with spatial templates: a salient toponym driven approach. *International Journal of Geographical Information Science*, 36(1):55–85, 2021.

15 Kota Hara, Raviteja Vemulapalli, and Rama Chellappa. Designing deep convolutional neural networks for continuous object orientation estimation. *arXiv preprint*, 2017. arXiv:1702.01499.

16 Annette Herskovits. Semantics and pragmatics of locative expressions. *Cognitive science*, 9(3):341–378, 1985.

17 Morteza Karimzadeh. Performance evaluation measures for toponym resolution. In *Proceedings of the 10th workshop on geographic information retrieval*, pages 1–2, 2016.

18 Morteza Karimzadeh, Scott Pezanowski, Alan MacEachren, and Jan Oliver Wallgrün. Geotxt: A scalable geoparsing system for unstructured text geolocation: Geotxt: A scalable geoparsing system. *Transactions in GIS*, 23, January 2019.

19 J.D. Kelleher and F.J. Costello. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306, 2009.

20 Anna-Katharina Lautenschütz, Clare Davies, Martin Raubal, Angela Schwering, and Eric Pederson. The influence of scale, context and spatial preposition in linguistic topology. In *International Conference on Spatial Cognition*, pages 439–452. Springer, 2006.

21 G.D. Logan and D.D. Sadler. A computational analysis of the apprehension of spatial relations. *Language and space*, pages 493–529, 1996.

22 Mateusz Malinowski and Mario Fritz. A pooling approach to modelling spatial relations for image retrieval and annotation. *arXiv preprint*, 2014. arXiv:1411.5190.

23 George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

24 Reinhard Moratz and Thora Tenbrink. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial cognition and computation*, 6(1):63–107, 2006.

25 Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

26 V.B. Robinson. Individual and multipersonal fuzzy spatial relations acquired using human–machine interaction. *Fuzzy Sets and Systems*, 113(1):133–145, 2000.

27 J.R.J. Schirra. A contribution to reference semantics of spatial prepositions: The visualization problem and its solution in VITRA. *The Semantics of prepositions: from mental processing to natural language processing*, page 471, 1993.

28 Michael Spranger and Luc Steels. Co-acquisition of syntax and semantics: An investigation in spatial language. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1909–1915. AAAI Press, 2015.

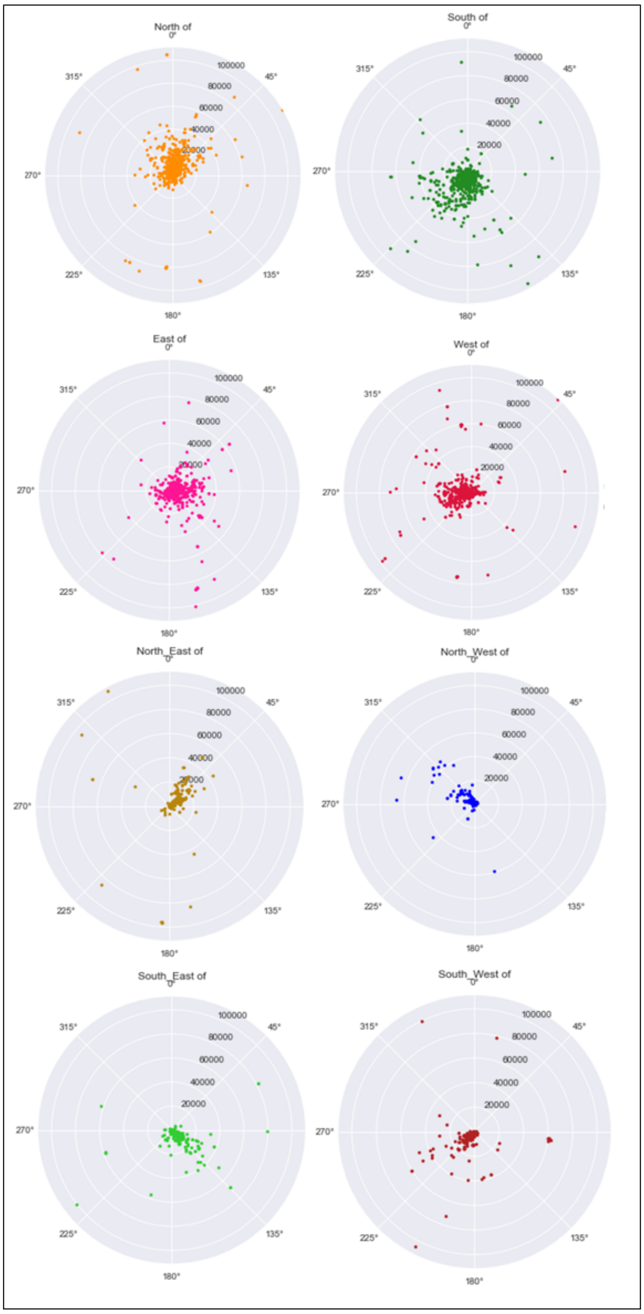
29 Kristin Stock and Javid Yousaf. Context-aware automated interpretation of elaborate natural language descriptions of location through learning from empirical data. *International Journal of Geographical Information Science*, 32(6):1087–1116, 2018.

30 Jan Oliver Wallgrün, Alexander Klippel, and Timothy Baldwin. Building a corpus of spatial relational expressions extracted from web documents. In *Proceedings of the 8th workshop on geographic information retrieval*, GIR '14, New York, NY, USA, 2014. Association for Computing Machinery. doi:10.1145/2675354.2675702.

31 M. Worboys. Nearness relations in environmental space. *International Journal of Geographic Information Science*, 15(7):633–651, 2001.

32 Xiaobai Yao and Jean-Claude Thill. How far is too far? – A statistical approach to context-contingent proximity modeling. *Transactions in GIS*, 9(2):157–178, 2005.

A Radial scatter plots for cardinal direction prepositional phrases



North of

South of

East of

West of

North_East of

North_West of

South_East of

South_West of