Advancing explanatory and tonal dialectometry

Sung, H.W.M.

CHAPTER 3

An Overview of Dialectometry[1]

The research presented in this dissertation relies on the quantitative methods used in dialectometry. In this chapter, I will give an introduction to dialectometry and provide the motivations for why I make use of dialectometric methods instead of using the isogloss approach to address issues in Yue dialect classification. The following section then introduces the methodology used in studying phonetic variation in dialectometry. It includes the explanations of the procedures and methods used in the following chapters.

3.1 Dialectometry

3.1.1 What is dialectometry?

Dialectometry refers to the measurement of dialect distances or similarities (Séguy 1973b). More broadly, it is the study of dialects through the means of computational and statistical approaches (Wieling and Nerbonne 2015).

Dialectometry offers a 'second life' to the linguistic atlases, due to its powerful ability to handle a vast amount of data. Linguistic atlases (and

[1]Section 3.1 is based on Sung (forthcoming).

dialect surveys) contain a huge amount of data, but there is a limit to the amount dialectologists can analyse manually, which led Séguy (1973b) to observe that "the rich collections that make up linguistic atlases remain underused". Dialectometry allows one to make use of more, if not all, data from these atlases and dialect surveys by determining "the dialect difference based not on a few arbitrarily chosen criteria, but on the integration of all the data" (Séguy 1973b).

3.1.2 Problems with the isogloss approach

For dialect classification, traditional dialectologists made extensive usage of isoglosses and they searched for bundles in order to locate possible dialect borders, as "the significance of a dialect area increases as more and more isoglosses are found which separate it from adjoining areas"(Chambers and Trudgill 1998:94). However, there are a number of problems with the isogloss approach. Kessler (1995) identified three: 1) isoglosses rarely coincide (see Figure 3.1); 2) there are exceptions to the isoglosses; and 3) dialects form a continuum.

A hidden problem of the isogloss method is potential cherry-picking. An isogloss map only shows the feature(s) which are considered significant. Furthermore, it is difficult to compare non-neighbouring, geographically distant varieties on isogloss maps. Additionally, an isogloss map quickly becomes chaotic when more features are added on the same map (for finding bundles). This can be illustrated with the example in Figure 3.1. This isogloss map is not easy to interpret, as there are many isoglosses present on the map, and they do not show areas which are relatively isogloss-free, despite the presence of bundles. Figure 3.1 only consists of 14 features (isoglosses) based on 40 dialects. If one increases the number of localities and features (which we have data for in the linguistic atlases), the interpretation of the map will only get more difficult. Hence, the isogloss approach does not appear to be suitable for such analyses.

Last but not least, an isogloss map is not entirely easy to replicate. Different researchers may make different decisions on, e.g. the inclusion of exceptions on either side of the isoglosses. This might cause inconsistencies in the replication of the analysis, especially if one wishes to add more data to the map.

Figure 3.1: Dutch isogloss map (digitised from Heeringa 2023)

3.1.3 An aggregate perspective

Because of the limits highlighted above, Nerbonne (2010a) argued for an aggregate perspective to view dialectal variation. An aggregate analysis "encompasses as much of the variation between language varieties as possible rather than concentrating on single linguistic features"(Nerbonne 2010a:476). This approach can obviate potential cherry-picked features based on scholar intuitions, which could lead to biases in the analysis, and can make the analysis more data-driven. Since different variants show different geographical patterns, the aggregate approach can strengthen their (geographic) signals by aggregating these differences (Nerbonne 2010b). This can be achieved through the calculation of dialect distances. Through these manoeuvres, "an aggregate characterisation"can then be obtained and it is then possible to "examine general tendencies in linguistic variation"(Nerbonne 2010a).

There are other advantages in using the aggregate approaches. Firstly, a substantial amount of (linguistic atlas/ dialect survey) data can be handled by computers. Moreover, in a relatively short period

of time, these qualitative data are converted into distances for further analysis. In addition, distances can be calculated not only between neighbours but between any pair of dialects in the data. Dialect comparison is no longer bounded by geographical proximity as would be the case in a classification based on isoglosses. The calculation of distances based on a substantial number of features is data-driven, meaning we can reduce the effect from selection biases, which could be present in the selection of features in a manual analysis. Moreover, dialect distances can be visualised using various techniques, such as dendrograms from cluster analysis and 2-dimensional (scatter) plots from multidimensional scaling, and their respective quantitative maps. Thus, the relationships (distances) between dialects can be interpreted much more easily and on a much larger scale. On top of these advantages, adding new data to the current dataset is not difficult. This is a huge advantage for scholars in situations where a subsequent volume of a linguistic atlas is being published. For instance, Volume 2 of the *Syntactic Atlas of the Dutch Dialects* (*SAND*, Barbiers et al. 2008) was published after Spruit's (2006) analysis, which was in SAND volume 1 only. Lastly, the consistency of the methodology and the aggregate characterisation encourage cross-linguistic comparison of dialectal variation (e.g. Nerbonne 2010b), which was previously uncommon.

The advantages of the aggregate approach are evident, in comparison to the isogloss approach. To address the research questions shown in Section 1.4, I make use of dialectometric approaches, for the arguments made for the aggregate perspective above.

3.2 A brief history of dialectometry

Dialectometry as it is known today is often associated with the enterprise of Jean Séguy's (1971; 1973b) pioneering turn from using isoglosses to look at dialectal variation to a distance-based approach. However, there have been applications of quantitative methods on dialect data before Séguy, and some studies apply similar methods independently from Europe. The following subsections will provide a brief overview of the development of the field.

3.2.1 Pre-modern dialectometric studies

Haag's (1898) study is often considered as one of the earliest quantitative dialectological study. Haag (1898) uses a *Kombinationskarte* ('combination map') to show the quantity of isoglosses (based on numerous variables) between the neighbours of each locality in the Alemannic and Swabian border zone. The use of Kombinationskarte could be found in later studies in the 20th Century, include Atwood (1955) and Glauser (1974).

In the 1950s, Reed and Spicer (1952) applied the 'correlation method' (a binary similarity measure of categorical feature data) to quantify dialect similarities.[2] A correlation coefficient is calculated between each pair of the lects in the dataset, which returns a correlation (similarity) matrix. In addition, the correlation matrix is divided into different sections based on a number of thresholds. For instance, pairs of dialects with a correlation coefficients higher than 0.58 are grouped together, yielding clusters of dialects.[3] Furthermore, Reed and Spicer (1952) have created 'isograde' maps, which shows the degree of similarity from a reference dialect in relation to the other varieties. This resembles the *reference point map* introduced by Goebl (1984).

Within Chinese dialectology, there were also some quantitative dialectological works before the beginning of modern dialectometry. Like in Haag (1898), Chao et al.'s (1948) *Report on a Survey of the Dialects of Hupeh* has a Kombinationskarte (綜合地圖 or 'aggregate map') which summarises the preceding 64 thematic dialect maps based on the dialect survey. Later in the 1990s, Cheng (1991) created a (Middle Chinese sound reflexes vs. dialect) data matrix and applied Pearson's correlation to calculate dialect 'affinities' (similarity scores). The returned scores are then fed into cluster analysis. By splitting the dataset into initials, finals and tones (and one aggregate analysis), Cheng has noticed very different sub-groupings, which might reflect different developments for different phonological units.

Before introducing the beginning of 'mainstream' dialectometry, there are two more notable early dialectometric studies, one from Japan and one from the Netherlands. Using data from the *Linguistic Atlas of Japan*

[2]The correlation method here is Kroeber and Chrótien's (1937) Q_6, which is a variant of Pearson's correlation (Phi) coefficient (they label it as V) that gives a more normal distribution.

[3]See O_2, VW_1, US_2, VW_2, and US_1 in Reed and Spicer (1952:352).

(LAJ), Inoue and Kasai (1982a,b) calculated the percentage of Standard Japanese forms that were recorded in each prefecture, and they performed factor analysis and cluster analysis on this so-called 'Kasai dataset'. When performing cluster analysis, Inoue and Kasai (1982a) have found that Japanese dialects can be divided into two large groups, namely Kanto (Eastern) and Kansai (Western). Each large group can be further divided into smaller subgroups. Following these earlier works in quantitative Japanese dialectology, Inoue also expanded his works beyond classification. For instance, Inoue (2007) investigates how railway distance is correlated to the usage of Standard forms (LAJ data) and whether high school students show a similar pattern as the LAJ or not. The second study made use of the 'Feature Frequency Method' (FFM) for the dialect comparison of Dutch dialects, based on the data from the *Reeks Nederlandse Dialectatlassen*. Based on the relative frequency of distinctive features (based on Chomsky and Halle's (1968) *Sound Patterns of English* system, plus additional features and modifications), Hoppenbrouwers and Hoppenbrouwers (1988) measured the Pearson's correlations (as a way to measure dialect similarities) between a reference variety (some local dialects and also standard Dutch) in order to find the most similar dialects to these reference varieties. Hoppenbrouwers and Hoppenbrouwers (2001) later applied the FFM technique and calculated the pairwise similarities and applied cluster analysis in order to classify 156 Dutch dialects.

3.2.2 Modern dialectometric approaches

The start of modern mainstream dialectometry is often credited to Jean Séguy (Wieling and Nerbonne 2015), the director of the *Atlas linguistique et ethnographique de la Gascogne*. At the end of the 6th volume of the atlas (Séguy 1973a), in addition to the display maps of the survey material, Séguy presented quantitative maps, including network maps (showing distances between neighbouring dialects through a network, on five linguistic levels lexicon, phonetics, phonology, verbs and morphosyntax), dialect frontier maps and a gradient map of Gasconity. Instead of displaying isoglosses to divide Gascony into different dialect areas, Séguy calculated the Hamming distance between the dialect localities and plotted them on network maps. This method is able to give an abstraction of the variation over a large quantity of variables (based on the previous volumes of the atlas), as well as demonstrating that there

is no such thing as an abrupt dialect border (Séguy 1973b).

In the 1980s, Goebl (1982, 1984) developed Séguy's idea further in Salzburg, by using computers to calculate dialect similarities[4], as well as introducing a number of computational and statistical techniques, visualisations and new map types. Firstly, Goebl (1982, 1984) automated similarity calculation by the means of *Relative Identity Value*. After obtaining the similarities, a rather new technique at the time, cluster analysis, was applied to the similarities, which returned groupings of dialects as clusters. Goebl has made extensive use of dendrograms from cluster analysis as well as plotting the cluster groupings on a map to show the numerical classification of dialects. In addition, new map types such as interpoint maps (network maps), honeycomb maps (maps that show intensity of similarities between neighbours), reference point maps and parameter maps were introduced. Goebl's methods have been applied to a variety of Romance languages, as well as English (Viereck 1985).

In the 1990s, Kessler (1995) made use of *Levenshtein Distance* for the calculation of phonetic distances of Irish dialects. Nerbonne and Heeringa (1997) continued this line of research in Groningen by experimenting, refining and elaborating the method on Dutch dialects, which then developed into Heeringa's (2004) dissertation on measuring dialect distances using Levenshtein Distance. Since then, Levenshtein distance has been applied to numerous languages, including Bulgarian (Prokić 2010) and German (Nerbonne and Siedle 2005). In addition, Nerbonne and colleagues have introduced more new map types such as the multidimensional scaling (RGB) map and beam map. Later, stochastic techniques were brought into dialectometry, namely noisy clustering and bootstrapping, respectively.

3.2.3 Recent developments

At the beginning, the field was mostly focused on developing methods to automatically identify dialect groups. However, this has changed in the last decade. Dialectometry has been applied to a wide range of topics beyond dialect classification, including diffusion of linguistic changes (e.g. Prokić and Cysouw 2013), dialect levelling (e.g. Leinonen 2010), regiolects (Nerbonne et al. 2013; Heeringa and Hinskens 2014), border

[4]Instead of using distances, Goebl (1982, 1984) uses similarities, which is the inverse value of the distances.

effects and dialect change (e.g. Heeringa et al. 2000; Wieling et al. 2018), socio-dialectology (e.g. Wieling 2012), extraction of important features (e.g. Prokić et al. 2012 and Sung and Prokić 2024a) and comparison of different linguistic levels (e.g. Spruit et al. 2009; Grieve 2013; Scherrer and Stoeckle 2016).

In addition to the new directions, there are also new methods being introduced to the field; many of these take corpus-based or natural language processing (NLP) approaches. These approaches do not require questionnaire data, i.e. word lists (or sentences) from dialect surveys. For instance, Huang et al. (2016:246) extracted 38 lexical alternations ("two or more different words with the same referential meaning") which have shown significant spatial autocorrelation from geo-tagged tweets on Twitter (data collected between 2013 and 2014), and they have identified dialect regions which match cultural geography and settlement history. Szmrecsanyi (2013) on the other hand uses data from the *Freiburg Corpus of English Dialects (FRED)* for his dialectometric study on grammatical variation. Szmrecsanyi used the frequency of a selection of grammatical variables as an input for his distance calculations, before applying methods from the dialectometric research groups in Salzburg and the Groningen.

More recently, the corpus dialectometry has been taken further by the application of NLP approaches to dialect corpora. Kuparinen and Scherrer (2023) applied dialect-to-standard normalisation, a neural machine translation technique, to generate embedding vectors for different dialect texts in two dialect corpora. This allows them to apply dimensionality reduction techniques and clustering in order to seek dialect groupings. In their later study, Kuparinen and Scherrer (2024) have applied a different approach in NLP, namely topic modelling, on three dialect corpora in order to perform dialect classification as well as extracting important features for each dialect group.

Last but not least, it should be mentioned that in recent years, there have been more studies on dialects outside Europe using methods from dialectometry. Just to name a few, these include Mayan (Blaha Pfeiler and Skopeteas 2022), Japanese (Heeringa and Inoue 2023) and Sinitic languages (Huang et al. 2024). By studying languages outside Europe, we can compare whether the patterns found elsewhere match with those in Europe, and if not, they invite us to ask further research questions. This practice can help us to gain a deeper understanding of language variation. In addition, analysing languages which are genetically and

typologically different from European languages might drive the needs to expand on the existing dialectometric methods. For instance, to study tonal languages, we need a method to calculate tonal distances. This issue is addressed in the current dissertation.

3.3 Methodology

To perform a dialectometric analysis, one has to first digitise existing data from the dialect surveys or linguistic atlases (if they were not already published in a digital format, which is the case for most of the dialect surveys/ linguistic atlases conducted in the 20th century, or even later). Details on the digitisation of the dataset used in this dissertation can be found in Chapter 4.

Dialectometry involves two major steps after data digitisation, namely distance calculation and recognition of dialect groups. Distance calculation concerns using a certain distance metric (depending on the type of input data) to measure the linguistic distance between a pair of dialects. For the purpose of dialect classification, dialect distances are measured between all the pairs of dialects in the dataset.

After obtaining all the pairwise dialect distances (which is stored in a distance matrix), one can analyse these distances (recognition of groups) through cluster analysis and multidimensional scaling. Cluster analysis seeks dialects (or any objects in a distance matrix) that are similar and group them together (visualised with dendrograms, if hierarchical cluster algorithms are used). Multidimensional scaling on the other hand reduces dialect distances from a higher dimension to a lower dimension, usually to two or three dimensions, in order to project the distances in the original distance matrix in a more interpretable way. The distances in the lower dimension can be visualised using 2D or 3D plots. In addition, the results from these techniques can also be visualised on maps.

The workflow of dialectometry (using Levenshtein distance) is illustrated in Figure 3.2.

3.3.1 Distance calculation

There are a number of ways to calculate dialect distances depending on what sort of data one is using. For feature (categorical) data, Goebl (1984) uses Relative Similarity Value (RIV) to find similarities (distance

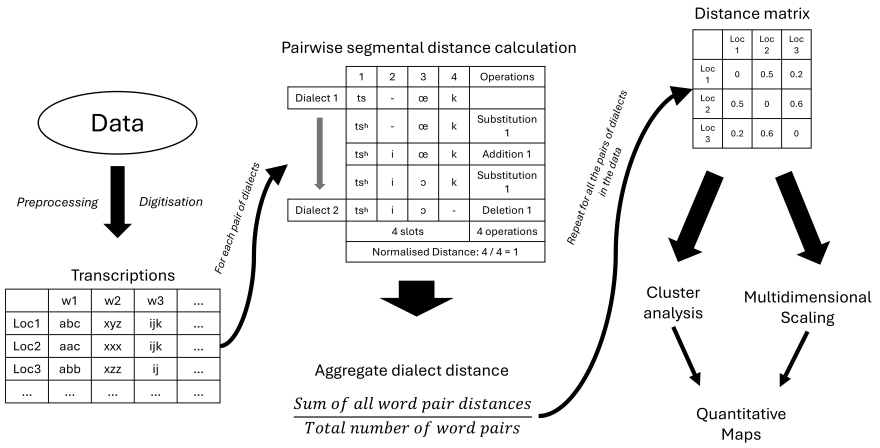

Figure 3.2: Workflow of Dialectometry using Levenshtein Distance

is 1 minus the similarity value) by comparing how many linguistic features[5] are shared by a pair of dialects out of all the features in a comparison. The formula for RIV is given in (3.1) below. j and k represent any two dialects; *COI* refers to *co-identity*, i.e. the number of shared items and *COD* refers to *co-difference*, i.e. the number of unshared items.

$$RIV_{jk} = \frac{\sum COI_{jk}}{\sum COI_{jk} + \sum COD_{jk}} \quad or \quad \frac{\text{no. of shared features in both dialects}}{\text{total number of features compared}} \tag{3.1}$$

Goebl's method, however, requires manual feature extraction (*taxation*, Goebl 2018) from transcription data, if the dialect survey/ linguistic atlas does not already provide categorical data of dialect features.

One common alternative approach in measuring distances for phonetic variation is the application of Levenshtein Distance (Kessler 1995, Heeringa 2004). Levenshtein distance counts the minimal number of operations (insertion, deletion and substitution) required to convert one string, i.e. the phonetic transcription of one word in one dialect, to another. This distance metric takes all the characters in a transcription into account, which may not be the case in Goebl's approach.

[5]These can be phonetic, phonological, morphological, syntactic, lexical or all of the above.

Dialects	Onset	Medial	Nucleus	Coda	Operations	Costs
Guangzhou	ts	-	œ	k		
	tsʰ	-	œ	k	Substitution from ts to tsʰ	1
	tsʰ	**i**	œ	k	Insertion of i	1
	tsʰ	i	**ɔ**	k	Substitution from œ to ɔ	1
	tsʰ	i	ɔ	**-**	Deletion of k	1
Bao'an	tsʰ	i	ɔ	-		4

Figure 3.3: Illustration of Levenshtein Distance

The first step in calculating Levenshtein distance (LD) is alignment. Alignment is an important step in LD, because if segments are not aligned in a linguistically coherent fashion, the distances we obtain will deviate from linguistic coherence. For instance, it makes more sense to use a variety of LD where consonants and vowels only align with themselves, and semivowels j and w are allowed to align with both consonants and vowels. Another important aspect of LD involves the gaps. When there is a difference in the number of consonants or vowels in an alignment, such as the example in Figure 3.3 where there is only one vowel in the Guangzhou dialect, but two vowels in the Bao'an dialect, then a gap is created to fit the alignment between the pair of strings.

The alignment of strings enables us to then calculated the minimal number of operations needed to transform one string to another, which yields the distance in the metric. The operations consist of insertion, deletion and substitution. An insertion is the addition of an element to an empty alignment slot, such as the Medial position in Figure 3.3. A deletion is the opposite of an insertion, where a previously occupied slot becomes an empty slot, as shown in the Coda position in Figure 3.3. Lastly, a substitution refers to a character being swapped with another character in the same slot, like the Onset position in Figure 3.3. Levenshtein distance calculates the minimal number of these operations required for the transformation of a string to another so that we can avoid alternative operations which can be applied to the same string transformation, but with an unnecessarily higher distance.

The last step is normalisation.[6] Normalisation converts the raw distances to a score which ranges from 0 to 1. In the example in Figure 3.3, the raw distances (number of operations) is 4. Since there are 4 slots in this alignment, the distance 4 is divided by the total number of alignment slots 4, which yields a distance of 1 for this word pair.

To obtain the aggregate distance for a pair of dialects, one sums all the word distances, and divide the sum by the total number of words compared in a dialect pair. The steps above are iterated for all the dialect pairs in the dataset. These pairwise distances are then stored in a distance matrix, which is a symmetrical table of pairwise distances. All the pairwise dialect distances can then be retrieved from the distance matrix for further analyses, such as cluster analysis and multidimensional scaling.

An example of a distance matrix is given in Table 3.1. In a distance matrix, each cell contains the distance for the corresponding dialects in the row and column. For instance, the distance between Foshan and Gaoming is 0.123, which can be found in the cells in row 2, column 1 or row 1, column 2. The distance from the same pair of dialects appear twice in this matrix, since it is symmetric. Lastly, the diagonal distances in the matrix are always 0, since these are the distances a dialect and its own, which always yield a distance of 0.

	Foshan	Gaoming	Guangzhou	Nanhai	Nanning	Panyu	Sanshui	Shunde	Zhanjiang
Foshan	0	0.123	0.032	0.156	0.13	0.063	0.062	0.082	0.176
Gaoming	0.123	0	0.108	0.214	0.15	0.132	0.13	0.149	0.221
Guangzhou	0.032	0.108	0	0.174	0.115	0.038	0.058	0.09	0.144
Nanhai	0.156	0.214	0.174	0	0.209	0.207	0.198	0.156	0.261
Nanning	0.13	0.15	0.115	0.209	0	0.153	0.155	0.165	0.223
Panyu	0.063	0.132	0.038	0.207	0.153	0	0.068	0.068	0.169
Sanshui	0.062	0.13	0.058	0.198	0.155	0.068	0	0.094	0.169
Shunde	0.082	0.149	0.09	0.156	0.165	0.068	0.094	0	0.221
Zhanjiang	0.176	0.221	0.144	0.261	0.223	0.169	0.169	0.221	0

Table 3.1: A sample distance matrix

3.3.2 Multidimensional scaling

Multidimensional scaling (MDS hereafter) is a dimensionality reduction method which represents "measurements of similarity (or dissim-

[6]Normalisation is not a compulsory step, as Levenshtein distance with and without normalisation both correlate very similarly with perception Heeringa et al. 2006. However, normalisation makes the distances more interpretable in the form of proportions.

ilarity) among pairs of objects as the distances between points of a low-dimensional multidimensional space" (Borg and Groenen 2005:3). In our case, dialect distances, as represented and stored in the distance matrix (as n dimensions, n being the number of dialects in the data), are often brought down to two or three dimensions.[7] These 'dimensions' represent the major underlying patterns in the data, and they are ranked by their explained variance to the original distance matrix. This procedure allows the analysts to literally "look" at the data through a scatter plot and speculate the structure of the data visually.

There are two common algorithms for MDS, namely classical and non-metric MDS. Classical MDS positions objects in reduced dimensions while preserving the original distances as much as possible (Legendre and Legendre 2012:492). This algorithm requires a valid 'metric', i.e. actual distances, like euclidean or cosine distances, as an input. Non-metric MDS (nMDS hereafter), on the other hand, does not have the same requirement as classical MDS. In cases where the preservation of the distances is not the primary importance, nMDS can be considered (Legendre and Legendre 2012). nMDS ranks the distances in the matrix, and represents only the ordinal properties in the data (Borg and Groenen 2005:203). Hence, this algorithm is non-metric, as it no longer works with ratio data for the dimensionality reduction.

We can visualise the dialect distances in lower dimensions using an MDS plot. An MDS plot represents the dialects as points, and the further the points are from each other, the more different they are, approximating the distances in the distance matrix. Unlike cluster analysis (see Section 3.3.3), the points on an MDS plot are not partitioned into discrete groups. However, if natural clusters are present, they are also apparent in the plots. Hence, MDS serves as a complementary analysis to cluster analysis (Legendre and Legendre 2012:415) in identifying groups of similar objects (dialects). In addition, no geographical information is added to the plot, so the distances projected on the plot is simply based on the distance matrix generated in the distance calculation. Moreover, plotting the distances in a lower number of dimensions allows us to visu-

[7]Mathematically, it is possible to extract more underlying patterns (shared behaviours among a number of objects), i.e. dimensions, up to the number of dialects considered in the data. However, as the number of dimensions considered increases, the explained variance stops increasing much very quickly, meaning that adding more dimensions in the analysis does not help explaining the data further. Very often, 2 or 3 dimensions are presented, as they explain the most amount of variance.

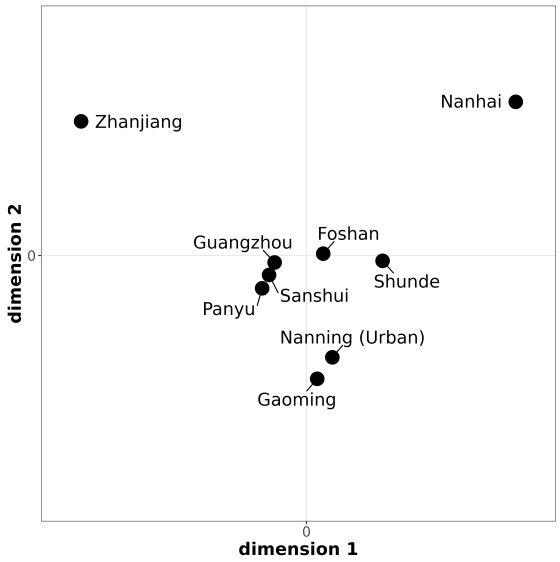

Figure 3.4: An example of an MDS plot (based on 9 Yue dialects)

alize continuum-like dialect relations (see Heeringa 2004; Leinonen 2010; Prokić 2010). An MDS plot can show 'transitional'dialects, often identified in the space of the plot between clusters, which is something cluster analysis cannot indicate. An example of an MDS plot is given in Figure 3.4.[8]

It is important to check how much the distances represented in an MDS plot represents the original distance matrix. This is indicated by the explained variance (r^2) or by the Stress value (Heeringa 2004). If the explained variance is too low, that means the MDS plot does not have a good representation of the distance matrix. However, there is no set threshold for when a MDS projection of the distances is too 'bad' to be discarded. In general, the higher the explained variance (or lower stress), the better. Although Heeringa (2004:161) notices that nMDS often has a higher explained variance than classical MDS, throughout the entire thesis, classical MDS is used unless specified. This is because in a dialectometry analysis, I argue that classical MDS is more suitable for the data. The preservation of distances is important, and the distances reflected in an MDS plot can be directly interpreted as their relative lin-

[8]The plot was created with LED-A.org.

guistic distance whereas with nMDS, this is not possible. An additional argument is that sometimes nMDS cannot process distance matrices if two (or more) varieties have the exact same distances with the rest of the dialects in the data.[9]

Another related matter to the explained variances is the number of dimensions needed for the interpretation of the data. By using a scree plot, we can look for a sudden drop of difference in the explained variance between two consecutive dimensions (known as the elbow method). The elbow method searches the point (dimension) where adding additional dimensions does not increase the explained variance a lot more than the existing dimensions combined. Figure 3.5[10] is a scree plot which shows the accumulative explained variance for each dimension, up to 4 dimensions. As illustrated in the figure, dimension 4 does not add more explained variance to dimension 1 to 3 combined.

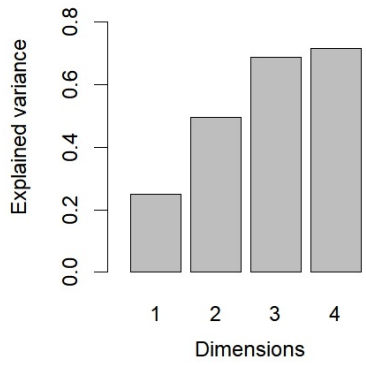

Figure 3.5: An example of an MDS scree plot

The MDS dimensions can also be projected on an MDS map. On an MDS map, the coordinates of the first three dimensions will be used, and each dimension is assigned a colour in the RGB colour spectrum. The resulting map will show points (individual dialects) with different content of the three dimensions, reflected in different amount of Red, Green and Blue combined (Heeringa 2004:161-163). The three-dimensional RGB spectrum is shown in Figure 3.6 (based on Leinonen 2010). Using this technique, linguistically similar dialects are represented

[9]This is the case for the aggregate tonal distances used in Chapter 8, using the isoMDS() function from the MASS library in R.

[10]This figure is borrowed from the analysis from the next chapter.

by similar colours, and linguistically different dialects are marked with sharp colour differences. If a certain dialect has a high intensity of either red, green or blue, it means the dialect shows a particularly strong pattern towards a particular dimension, but not others.

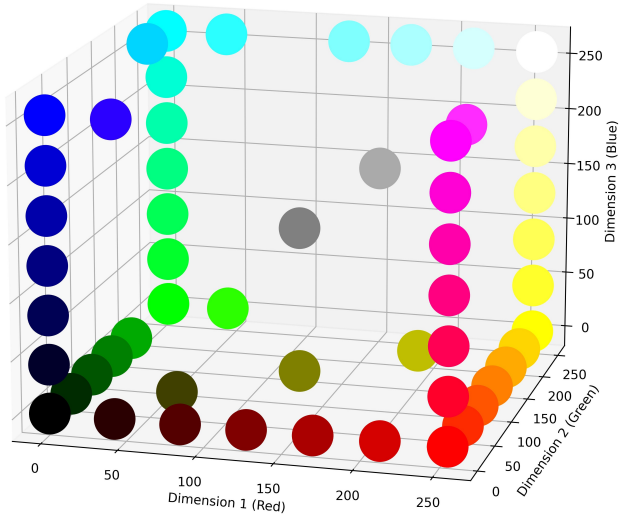

Figure 3.6: Three-dimensional RGB spectrum (based on Leinonen 2010)

3.3.3 Cluster analysis

Cluster analysis refers to the partition of objects (dialects in our case) into groups (Manning and Schütze 1999; Everitt et al. 2011). The application of cluster analysis helps us to identify dialects which are similar enough to be considered as the same group. There are many algorithms used in cluster analysis, but the most commonly used algorithms in dialectometry are the so-called agglomerative hierarchical clustering algorithms. These algorithms find successive clusters based on previously established clusters, creating a hierarchical representation of the clusters in a dataset (often visualised in a dendrogram). There are also other cluster algorithms available, including divisive methods, which separates a number of elements successively into separate groupings (Manning and Schütze 1999) as well as partitional clustering (sometimes called flat clustering), where a single partition is given to the data without any hi-

erarchical structure (Jain and Dubes 1988). However, divisive clustering algorithms are not much utilised in the dialectometric literature. The partitional algorithms show more presence in dialectometry than the divisive methods, though agglomerative clustering is still the preferred method for discovering clusters in the dialect data.

Many hierarchical cluster algorithms used in dialectometry are often referred to as *hard* clustering techniques. This is because any element (dialect) in the data is only allowed to be in one cluster, and not others. This returns a crisp cluster membership for all the elements (Everitt et al. 2011). There are two problems with hard clustering. Firstly, based on our current understanding of dialect variation, dialects often form a continuum, and do not have an abrupt boundary. Rather than forcing a dialect into a particular cluster, it would be more useful to find the probability of a dialect belonging to a certain dialect group (cluster). Secondly, hard clustering is known to suffer from instability: small differences in the distance matrix can lead to big differences in the groupings (Nerbonne et al. 2008). Probabilistic clustering algorithms are introduced to dialectometry as a complementary technique to hard clustering. This allows us to speculate to what extent the hard clusters are justified.

In the following sub-sections, agglomerative hierarchical clustering and probabilistic clustering will be introduced, since they are the more common approaches in dialectometry, which have been applied to a number of languages.

Agglomerative Clustering

Agglomerative hierarchical clustering algorithms seek partitions of objects by finding successive clusters based on previously established clusters. There are many algorithms in this family, but the most common ones are Single Linkage, Complete Linkage, Unweighted Pair Group Method using Arithmetic Averages (UPGMA), Weighted Pair Group Method using Arithmetic Averages (WPGMA) and Ward's Method. The resulting nested clusters can be visualised with a dendrogram.

All the algorithms mentioned above proceed from a distance matrix (see Table 3.1), followed by iterations of the fusion between the closest elements, i.e. the elements with the smallest distance in the entire distance matrix. What makes the algorithms different from each other is how the distances are recalculated from the newly fused elements. For

instance, Single Linkage (also known as Nearest neighbour) uses the smallest distance between the newly fused elements, [*ij*] with the rest of the elements *k*. Contrastively, Complete Linkage (also known as Furthest neighbour) updates the matrix by using the biggest distance between the fused elements [*ij*] with the rest of the elements *k*. UPGMA takes the average (arithmetic mean) of the distances of the fused elements [*ij*] as their new distances to *k*, whereas WPGMA does the same, except the average distances are weighted by dividing the sum of the distances of the fused elements by 2, instead of the number of elements in the fused group (hence it is weighted). Lastly, Ward's method (Ward 1963), also known as the minimal variance method, merges elements which increase the sum of squared distance of the mean from each cluster the least.

Different algorithms have their own strength in identifying clusters in different distributions of data points. For instance, single linkage tends to chain nearby objects together, resulting in clusters which are unbalanced and elongated. (known as the chaining effect, Everitt et al. 2011:89). On the other hand, complete linkage clusters tends to find "compact clusters with equal diameters apart" (Everitt et al. 2011:89). Next, UPGMA tends to be more faithful to the distances found in the original distance matrix (higher cophenetic correlation coefficients, Heeringa 2004), whereas Ward's method tend to give balanced clusters (Prokić and Nerbonne 2008).

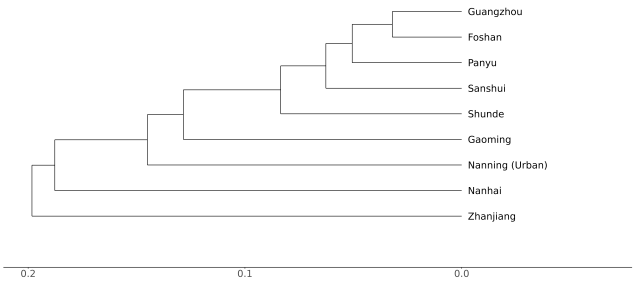

Figure 3.7: Dendrogram of 9 Yue dialects (UPGMA, $r^2 = 0.85$)

The results of cluster analysis can be visualised on a dendrogram (tree diagram), a cluster map or an MDS plot (see Section 3.3.2). A dendrogram (see Figure 3.7[11]) shows the history of the fusion during the cluster analysis. Furthermore, the branch lengths indicate the distances

[11]This figure is created using LED-a.org (Heeringa et al. 2024).

between clusters or dialects (Everitt et al. 2011:88). A cluster map on the other hand takes the cluster grouping from the dendrogram and assigns colours to each group, showing where each cluster is distributed in the region.

Choosing a suitable cluster algorithm

Since different algorithms have different strengths in finding clusters depending on the distribution of the objects, it is necessary to assess which algorithm is most suitable for a given dataset.

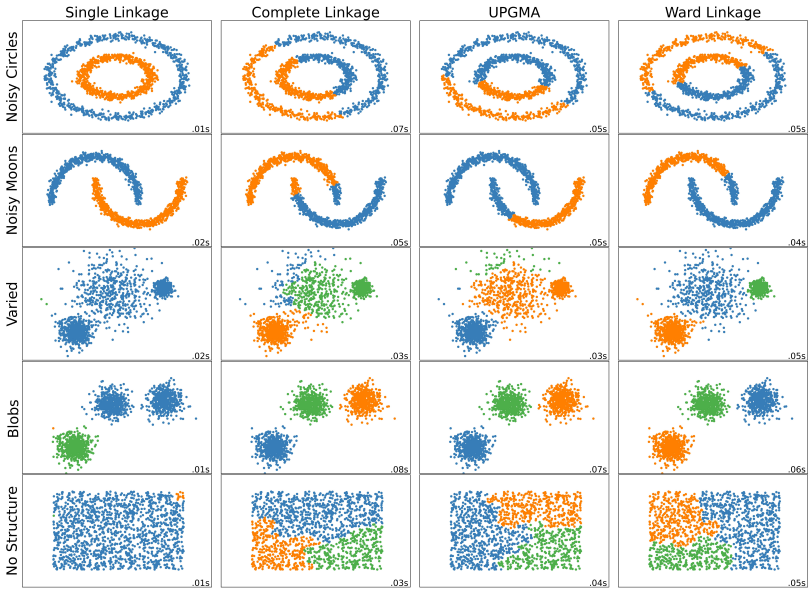

Figure 3.8: Comparison of different cluster solutions based on various cluster algorithms (Rows: datasets, Columns: algorithms)

In Figure 3.8[12], we can see that Single Linkage is suitable to identify clusters which are chain-like (both the 'Noisy Circles' and 'Noisy Moons' datasets). For tight clusters, as illustrated by the 'Blobs' dataset, Complete linkage, UPGMA and Ward's method are both good at identifying

[12]This figure was created based on the code provided by Pedregosa et al. (2011), available at `https://scikit-learn.org/stable/auto_examples/cluster/plot_linkage_comparison.html#sphx-glr-auto-examples-cluster-plot-linkage-comparison-py`. Weighted average is not available as a cluster algorithm here.

the clusters. This is not the case, however, for Single linkage. For the 'Varied' dataset, we see a different picture. Only the Ward's method is able to give reasonable clusters of this dataset. Lastly, it should be pointed out that all these algorithms will always return cluster solutions, even if there are no natural clusters in the dataset. This is illustrated by the 'No Structure' dataset.

By visual inspection, one can make judgements for the performance of the cluster algorithms with intuition, using the cluster-annotated MDS plot (Legendre and Legendre 2012). Another method for evaluating the suitability of the algorithm is the cophenetic correlation coefficient. This coefficient is computed through correlating the distances (branch length) represented in the dendrogram with the original matrix. The higher the correlation between the two, the more the cluster solution corresponds to the original matrix (Prokić and Nerbonne 2008).

In addition to the cophenetic correlation coefficient, the silhouette method can also indicate how balanced and separate the clusters are. A silhouette score of a dialect indicates a dialect's "separation from its cluster against the heterogeneity of the cluster" (Everitt et al. 2011:128), ranging from -1 to 1. A score of 1 indicates the dialect is 'well classified' in a particular cluster, i.e. having small distances with the other dialects in the same cluster, while being distant from the nearest cluster. Taking the average of all the silhouette scores can give us an indication of the overall separability and heterogeneity of a cluster solution using a certain algorithm. By comparing the average silhouette scores for a range of cluster solutions between different cluster algorithms, we get an idea which algorithm can generate better clusters overall.

Determining the number of clusters

When applying cluster analysis, one has to determine the number of clusters. Since cluster analysis is used in an exploratory fashion, i.e. there is no ground truth to how many groups there are and which dialect groups dialects are part of, internal validation methods can be used to determine how many clusters are appropriate in an analysis.

The silhouette method also has applications in determining the number of clusters in an analysis. A silhouette plot, which displays the silhouette scores of all the dialects (see the description of the silhouette method above) can be accompanied with an MDS plot in showing how fitting the cluster solutions are to the distribution of the dialects in the

data (Legendre and Legendre 2012). The silhouette plot is particular useful in showing the proportion of dialects in a cluster which have low separability (mixing with another cluster) as well as how balanced the clusters are.

Probabilistic Clustering

Cluster analysis is known to generate a 'hard' partition to the dialect data, meaning an element (a dialect) is only allowed to be in one cluster, but not two or more clusters nor does it show gradual membership for certain clusters. For instance, cluster analysis will always return a number of clusters, despite the data may not show a clustered pattern. Prokić and Nerbonne (2008:163) found that although a multidimensional scaling plot (see Section 3.3.2) shows a non-binary distribution of dialects, cluster analyses would still give a two-way partition. In addition, cluster analysis is known to be unstable, meaning that small differences in the distance matrix (e.g. caused by typos, small changes in the dataset) can lead to big differences in the results (Jain and Dubes 1988:79), i.e. changing the cluster membership of dialects. To identify stable clusters, several probabilistic approaches have been used in the dialectometrical literature, namely bootstrapping (Nerbonne et al. 2008), noisy clustering (Nerbonne et al. 2008).

The central idea of bootstrapped clustering is resampling the data through replacement (Nerbonne et al. 2008). Instead of using the cluster solution from one distance matrix, multiple distance matrices are created by resampling different sets of words and replacing them with some existing words. This means that some words would be repeated (weighted more) in these resampled distance matrices, while other word-distances are removed. By iterating clustering from many resampled distance matrices, we get a probability of the dialects being a member of certain clusters.

Noisy clustering on the other hand is performed by adding random noise (specified by a noise ceiling) to the original distance matrix, and iterate this process for 100 times or more (Nerbonne et al. 2008). Like bootstrapping, each iteration will yield a different partition, and based on these different solutions, a probabilistic dendrogram can be produced to show how stable the clusters are, and which dialects are part of these clusters. A probabilistic dendrogram is shown in Figure 3.9.[13] Both boot-

[13]This figure was created using *Gabmap* (Nerbonne et al. 2011; Leinonen et al. 2016).

strapping and noisy clustering can identify stable clustering. However, neither of them are able to remove the biases of a particular cluster algorithm used in the iterations (Nerbonne et al. 2008:652).

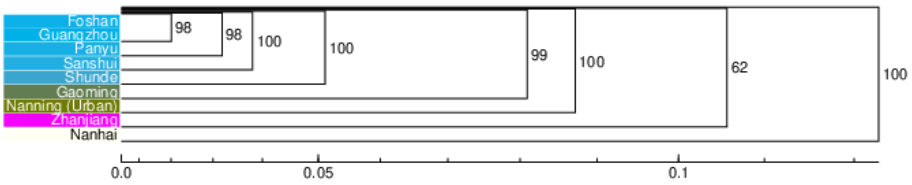

Figure 3.9: Probabilistic dendrogram of 9 Yue dialects (noise: 0.2, exponent 1.5, method: UPGMA and WPGMA)

3.3.4 Quantitative maps

Finally, the results from multidimensional scaling and cluster analysis can also be visualised in the form of maps. In addition, these results can also be overlaid on maps with other extralinguistic features, such as elevation, river systems and (present or historical) political borders. These can be done using a Geographical Information System (GIS). The maps in the current thesis were created using *QGIS* (QGIS Development Team 2022).