Advancing explanatory and tonal dialectometry

Sung, H.W.M.

CHAPTER 5

Segmental Variation of Yue Dialects[1]

5.1 Introduction

In Chapter 2, several problems in the classifications of Yue and Pinghua dialects have been raised, namely problems related to subjectivity in the choice of features. One can reduce the biases introduced by selecting only a handful of features by using the dialectometric methods introduced in Chapter 3. In this chapter, Yue-Pinghua dialects will be examined using dialectometric techniques, namely with distance calculation, multidimensional scaling and cluster analysis. A comparison will be made between the dialectometrical analysis against the traditional classification. Moreover, a novel classification of the Yue-Pinghua dialects (according to the LAC) will be proposed based on the results of the dialectometric analyses.

In addition, dialectometry has applications other than dialect classification. For instance, one other central idea within dialectology is that dialect patterns can be explained by physical geographical features, such as mountains, and political boundaries acting as barriers. Rivers on the other hand can act as both a barrier and a medium of diffusion. These

[1]Section 5.2.1 and Section 5.3.2 are based on Sung et al. (2024) and Sung (forthcoming) respectively.

correlates of dialectal variation are linked to the "density" in communication (Bloomfield 1933). A dialectometric analysis based on a large digital dataset is more suitable to address these types of questions. Along these lines, two broader questions are raised: 1) do Yue dialects, which are genetically and typologically different from European languages, display a continuum pattern similar to that of European languages? and 2) Can the dialect landscape in the Yue region be explained by geography or socio-historical factors? These questions are addressed in Section 5.3 and finally, the chapter concludes in Section 5.4.

5.2 Dialectometric analysis of Yue segmental variation

Most existing classifications of Yue and Pinghua suffer from the subjectivity and opaqueness in the choice of features in the analysis. If each scholar uses a different set of features (sometimes even without providing the list of features used), just like in the debate over Yue and Pinghua, different conclusions can be made. This makes the replications and comparisons with the literature difficult. In addition, it is possible that extralinguistic factors played a role in the scholars'judgements when drafting a classification scheme. Therefore, there is a need to assess the Yue and Pinghua data more objectively (through the use of data-driven approaches), as well as making use of more data (features and dialects) than the existing studies.

Using the methods introduced in Chapter 3, one can process a much higher amount of data than manual analyses. This reduces the subjectivity and selection biases introduced in the analysis. Moreover, these methods are also reproducible and falsifiable, making the analyses more accessible to scholars who might be interested in the same problems and beyond.

The following subsections will present the results of the dialectometric analyses, which will address the Yue-Pinghua dichotomy, provide a comparison between the traditional classification and the dialectometric one, and lastly yield a novel classification of Yue.

5.2.1 The Yue-Pinghua dichotomy

The following analysis is done using classical Levenshtein distance.[2] The Cronbach's alpha of this dataset is 0.9772.[3] Cronbach's alpha is a method to measure consistency or reliability (Heeringa and Prokić 2018), based on the average inter-item correlation of the current dataset. It ranges from 0 to 1, and the higher the score, the more consistent the data is. The widely accepted threshold for Cronbach's alpha is 0.7 (Heeringa and Prokić 2018), and the score obtained by the current dataset indicates that it has "samples large enough to provide reliable signals" (Heeringa et al. 2006).

The dialectometric result on the segmental variation of Yue and Pinghua somewhat resembles the opponents of merging Yue and Pinghua together in the Yue-Pinghua dichotomy debate.

Figure 5.1 is a multidimensional scaling plot which approximates the distances calculated between all Yue and Pinghua dialects in a 2D plot. The axes, labelled as dimension 1 and 2, represent the underlying patterns which explains the most amount of variance found in the aggregated dialect distances. The first dimension captures the differences between Northern (Guibei) Pinghua and the rest of the dialects. Northern Pinghua (orange) dialects appear to be very different from the rest of the Yue (black) and Southern Pinghua (blue) dialects. We can see that most of the orange dots are located on the left, meaning they have high negative values in the first dimension. There is one exception, i.e. the Lingui dialect, which is located much closer to 0 in dimension 1, towards the rest of the dialects.

There seems to be no continuum between the Northern Pinghua dialects and the rest of the dialects in the data, which suggests an abrupt boundary between Northern Pinghua and the continuum consisting of Yue and Southern (Guinan) Pinghua dialects. This result resembles the

[2]The Levenshtein distances were calculated using *LED-A.org* (Heeringa et al. 2024). The parameters include: *Method*: plain cost = 1, *Extra allowed segment alignments*: 'i/j/u/w versus anything', *Normalization*: 'divided by alignment length' and *Measurements are based on*: 'whole words'. PMI Levenshtein, the version of Levenshtein distance with automatic inference of segmental distances based on co-occurrence statistics (Wieling et al. 2012), was also tested on this dataset. However, both the consonantal and vowel distances do not conform much to the natural classes based on phonetics (e.g. back vowels and front vowels cluster together; velar and labial consonants cluster together). Therefore, classical Levenshtein distance is used for this chapter.

[3]Calculated using *LED-A.org* (Heeringa et al. 2024).

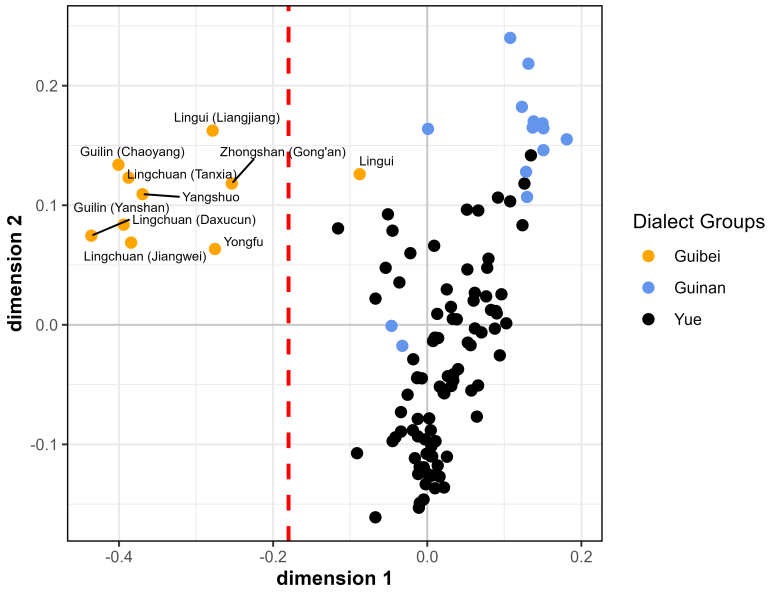


Figure 5.1: MDS plots of the segmental distances of 113 Yue and Pinghua dialects (Levenshtein distance), $r^2 = 0.59$

traditional analysis, according to the opponents to the separation of Guinan Pinghua and Yue dialects (Wu 2001; Tan 2000; Liang 1997, cited in Tan 2012).

Northern Pinghua, despite being very interesting with their own merits when it comes to its relationship with the Yue continuum, will be removed from the rest of the analyses throughout this dissertation. This decision is made following the outlier logic. Like removing outliers in other statistical analyses, the Yue continuum (including Southern Pinghua) can then become less skewed due to the Northern Pinghua as outliers (in other words, less squished in dimension one), so that tonal variation (which is a rather unexplored territory) can be interpreted more easily. The skewed picture of the Yue-Pinghua landscape is illustrated in the Figure 5.2.

In Figure 5.2, northern Pinghua dialects are represented by the circles in green, whilst other dialects can be found in a continuum ranging from pink-magenta to orange-pale yellow. Removing Guibei Pinghua from the data, as we will see in the following analyses, allows us to speculate the internal variation of Yue within the continuum more clearly.

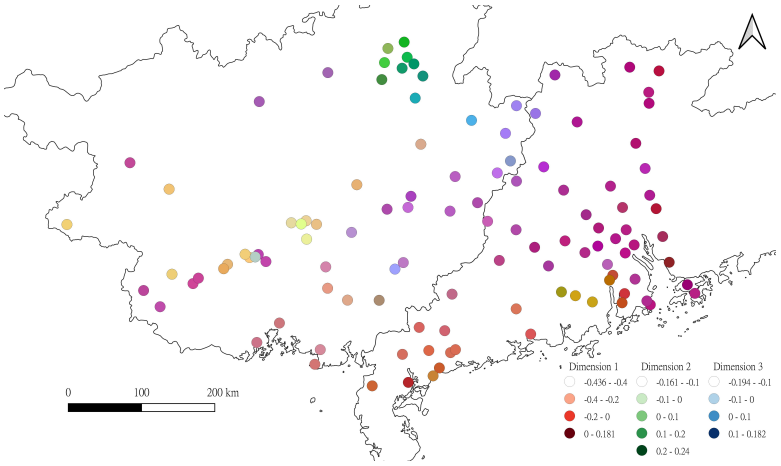


Figure 5.2: MDS map of the segmental distances of 113 Yue and Pinghua dialects (Levenshtein distance), $r^2 = 0.72$

Lastly, please note that Lingui[4] is now treated as part of Southern Pinghua based on its similarity with the rest of the dialects.

The Yue-Pinghua controversy has been hotly debated in Yue-Pinghua linguistics for several decades now. This section has provided a dialectometrical account of the problem for the first time, which compares the dialect distances between Yue, Southern (Guinan) and Northern (Guibei) Pinghua dialects, without removing certain dialects from the analysis based on manually chosen features, like in Carlyle (2020) (see Section 2.2.2). The result gives additional support to the previous literature which stated that Northern Pinghua is indeed linguistically quite distant from Yue, and could potentially be considered as a different dialect group (or Sinitic branch). The only exception is the Lingui dialect, which is linguistically closer to the rest of the dialects. However, this analysis only account for around 130 words. More research is required to confirm the current analysis by considering more lexical items, as well as finding which features are contributing to the boundary between the Yue continuum and the Northern Pinghua dialects, quantitatively.

From now on in the dissertation, Guinan Pinghua dialects will be treated as a dialect group of Yue.

[4]This Lingui dialect comes from Wutong.

5.2.2 Comparison with the traditional classification of Yue dialects

For the classification of Yue dialects, it is important to find out to what extent a dialectometric analysis agrees with the traditional classification. The differences could be due to the way distances are measured; the use of quantitative classification methods and the absence of responsible features in the traditional classification. In addition, the traditional classification could also be based on extralinguistic factors. Lastly, it could also be possible that features that are not sound enough have been used in the classification. All of the above factors could lead to a disagreement between the traditional and the dialectometric classification (Prokić 2010:55). This subsection will compare with the LAC classification (see Section 2.2.1) against the patterns found on the MDS plot.

Comparison with the classification in LAC

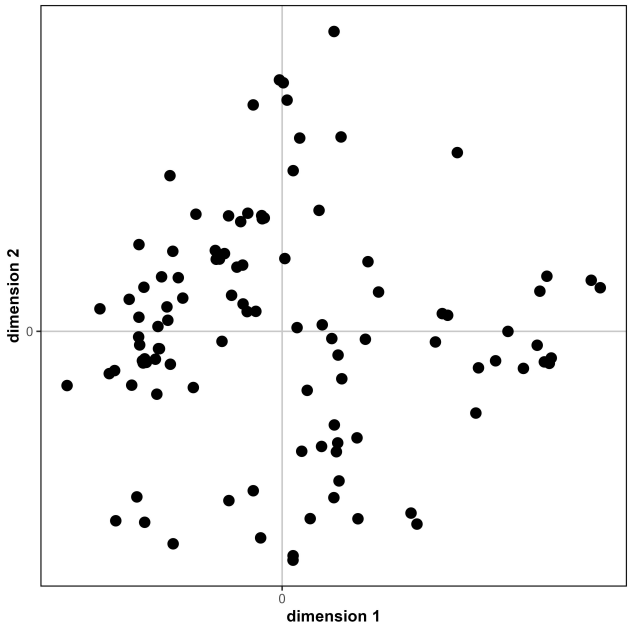


Figure 5.3: MDS plot of 104 Yue dialects (Levenshtein distance, r^2=0.50)

For the dialectometric classification of Yue dialects, Levenshtein distance has been computed between all the pairs of 104 Yue dialects (after removing Northern Pinghua).[5] The dimensions of the Levenshtein distances were reduced using multidimensional scaling, and they are visualised in a two-dimensional MDS plot (see Section 3.3.2 for the details on MDS) in Figure 5.3.[6] On this plot, there are no immediate tight clusters that are isolated from others. This is a typical image of a dialect continuum, where a (loose) cluster is connected with another cluster by some dialects in between. For instance, dialects in the cluster at the bottom (low dimension 2 region) are connected with the rest of the dialects through a string of dialects on the right of the x-axis around 0.

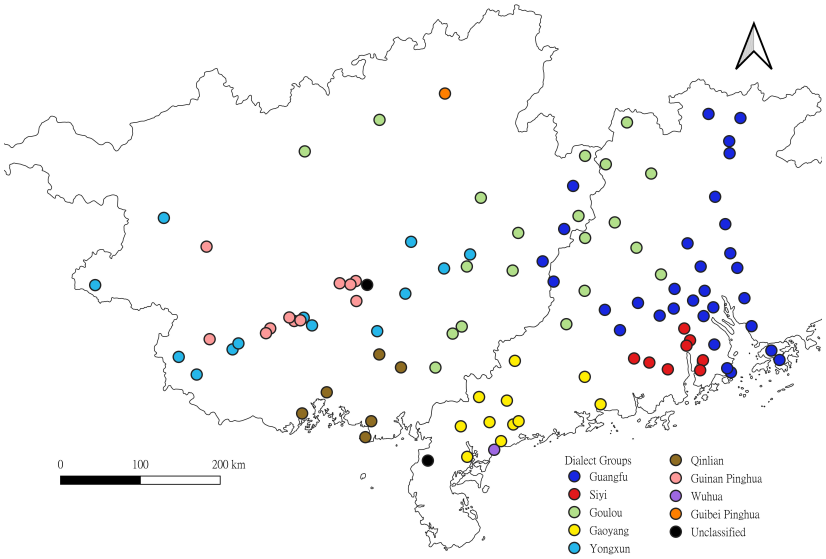


Figure 5.4: LAC classification map

The classification of the LAC is presented on a map in Figure 5.4. When we annotate the LAC dialect groups on the MDS plot, as shown in Figure 5.5, we can see a lot of overlaps between the dialect groups. For instance, the distribution of Guangfu (dark blue) dialects overlap with the distribution of Goulou (light green) and Yongxun (light blue) dialects. Adding the annotations of the LAC dialect groups does not

[5]The Cronbach's alpha of the Yue subset of the dataset is 0.9597.
[6]This plot was created using R.

delineate the clusters on this plot.

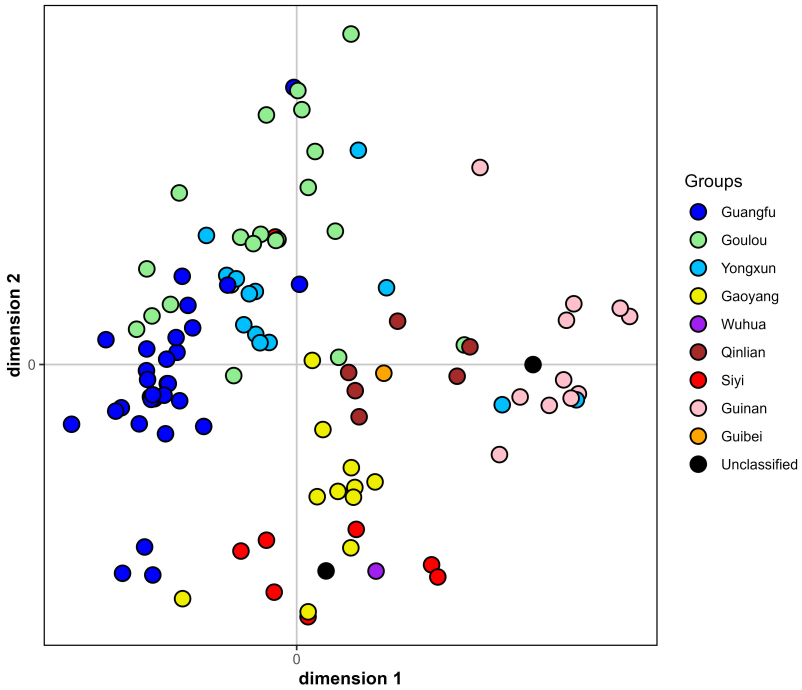


Figure 5.5: MDS plot of 104 Yue dialects (Figure 5.3) with LAC dialect group annotation (Dimension 1 & 2)

Figure 5.5 suggests a few things which could be adding some further insights to Yue dialect variation according to the traditional classification. First of all, the continuum consisting of Guangfu, Yongxun and Goulou dialects indicate that there are no strong linguistic borders between these dialect groups, as the traditional classification suggests. One possible reason why Yongxun dialects were classified as a separate dialect group could be related to an extra-linguistic factor. Yongxun dialects were formed from the emigration of Guangfu speakers around the time of the First Opium War, for trade (de Sousa 2020:268). This seems to be the basis of the division between Guangfu and Yongxun dialects, despite their high linguistic similarity. Goulou dialects, on the other hand, were classified as a separate dialect group based on their reflexes of Middle Chinese voiced obstruents. This criteria seems not representative enough to separate Goulou dialects from Guangfu and

Yongyun dialects.

Guinan (pink), as well as Siyi dialects (red), on the other hand, are quite distant from the tighter Guangfu continuum linguistically. Although Siyi dialects are situated with the Gaoyang dialects in Figure 5.5, the linguistic distances between the two dialect groups are actually bigger than it seems. Figure 5.5 only considers the first two dimensions from MDS. Some variation patterns, namely the distances of the Siyi dialects from the rest of the dialects, are not captured. When we turn to Figure 5.6, it gives a much clearer picture of Siyi dialects, although sharing the same space in dimension 2, are still distant from Gaoyang dialects, illustrated in dimension 3. This entails that (most) Siyi dialects are even more different from the Guangfu dialects than the Gaoyang dialects. Lastly, the Qinlian (brown) dialect group is linking the Gaoyang and Guinan dialects. This suggests another big continuum linking dialects in Guangdong and Guangxi, in addition to the Guangfu continuum discussed earlier.

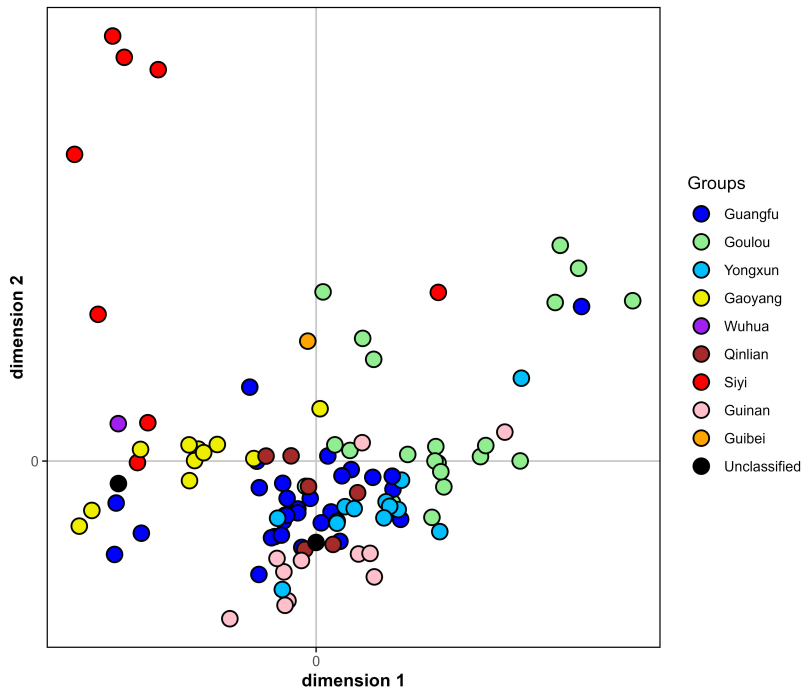


Figure 5.6: MDS plot of 104 Yue dialects (Figure 5.3) with LAC dialect group annotation (Dimension 2 & 3)

Following LAC's classification, we will find dialect groups which consist of only one member (due to the lack of dialect survey localities), namely the Wuhua dialect group (the Wuchuan-Wuyang dialect, in purple). Wuhua dialects were classified as a separate dialect group also due to their differences in the reflex of Middle Chinese voiced obstruents. This criteria also does not seem to be representative either, since overall, the dialect distances between Gaoyang and Wuhua dialects are not big.

Another insight which we might gain from the plot is that we can use the plot to estimate where the unclassified dialects (in black) lie in the continuum, and classify them. For instance, the Litang dialect on the right of Figure 5.5 is close to the Guinan dialects in general (in Guangxi, the black circle next to the purple circles). Based on the dialect distances, we can preliminary classify this dialect as a variety of the Guinan dialects, under the labels of the LAC. This is additionally supported by Litang's geographical location, surrounded by the Guinan dialects. The Suixi dialect, situated at the bottom of the plot, fits into the Gaoyang dialects and also the Siyi dialects. If we look at Figure 5.4, we will find that Suixi is also located between the Gaoyang and Qinlian dialects next to the Guangdong-Guangxi border, which does not contradict with what the MDS plot suggests. The classification of the previously unclassified dialects are further addressed in the next subsection.

Based on the observations of the differences between the LAC classification and the dialect distances, the discrepancies seem to come from several sources, namely possible usage of extralinguistic factors (Yongxun) and the usage of features that are not sound enough (e.g. Goulou and Wuhua). For Guinan Pinghua, we can see that they are distant to many dialect groups (although still linked to the others by e.g. the Qinlian dialect group), but there is an absence of responsible features in the description for the delineation of this dialect group.

5.2.3 A dialectometric classification of Yue dialects

The subsection above has shown that not all dialect groups which Wu (2007, 2012) came up with are linguistically sound enough. Although some of the groups seem to hold as a whole, there are also overlaps between the groups with a portion of their members, such as dialects within the Guangfu, Yongxun and Goulou dialect groups.

In order to obtain a more data-driven classification of the dialects based on the current dataset, this section will explore the partition of the Yue dialects using dialectometric approaches.

Cluster analysis

Cluster analysis can help us identify groups of dialects. However, to make use of cluster analysis to shed more insights to the dialect distances obtained with Levenshtein distance, a number of things should be considered, namely which cluster algorithm suits the current dataset better, and how many clusters should we consider.

Algorithm	Variance Explained
Complete likage	33.3%
UPGMA	72.4%
WPGMA	40.0%
Ward	29.1%

Table 5.1: Explained Variance of various cluster algorithms

To determine which cluster algorithm is most faithful to the original distance matrix, the explained variance of the four algorithms were calculated[7] and they can be found in Table 5.1. UPGMA has the highest explained variance with the original distance matrix, and Ward's method has the lowest. Based on this indicator, UPGMA shows the most faithful representation of the original distances.

In addition, Figure 5.7 plot shows the respective average silhouette scores[8] from the 2- to 10-cluster solutions in four cluster algorithms. UPGMA retains the highest average silhouette score in most of the cases, while WPGMA's scores remain high among the different cluster solutions until the 7-cluster solution. Ward's method starts off having the third highest scores until the 7-cluster solution, when it became the second highest and lastly, Complete Linkage starts off with a very low score, and become closer to Ward's method at the 4-cluster solution.

What this plot suggests is that UPGMA seems to be consistently outperforming the other cluster methods in terms of getting rather ho-

[7]Calculated using *LED-A.org* (Heeringa et al. 2024).

[8]The average silhouette scores were calculated using the *scikit-learn* library on Python.

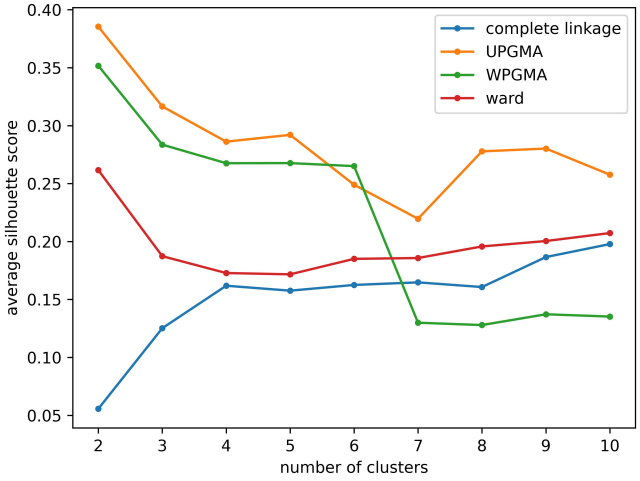


Figure 5.7: Average Silhouette Scores from different number of clusters obtained with four cluster algorithms

mogenous clusters (see Section 3.3.3). However, before getting to determining the number of clusters, I would argue that the explained variances and the average silhouette scores are not enough for the choosing which cluster algorithm suits best for the current dataset and for the purpose of dialect classification.

Taking a look at Figure 5.8. The left subplot is the silhouette plot for UPGMA, 5-cluster solution[9], and the right subplot is an MDS plot, annotated with the clusters of this cluster solution.

With the average silhouette score of 0.29 (the vertical dash line in red), we can see that the silhouette score of over half of the dialects in cluster 1 are over the average. This indicates that over half of the dialects in cluster 1 are fairly homogenous and distant from the nearest clusters. However, it is very apparent that cluster 1 is very big, cluster 2 is much smaller, and the rest of the clusters are invisible. This is because clusters 3 to 5 consist of only one member each. The dialects in clusters 3 to 5 are from Heshan, Xindu and Lingui. It is no coincidence that these three dialects are considered as its own cluster by the algorithm,

[9]The 5-cluster solution is chosen instead of the 3-cluster solution because it highlights the dialects which are island-like, see the following descriptions.

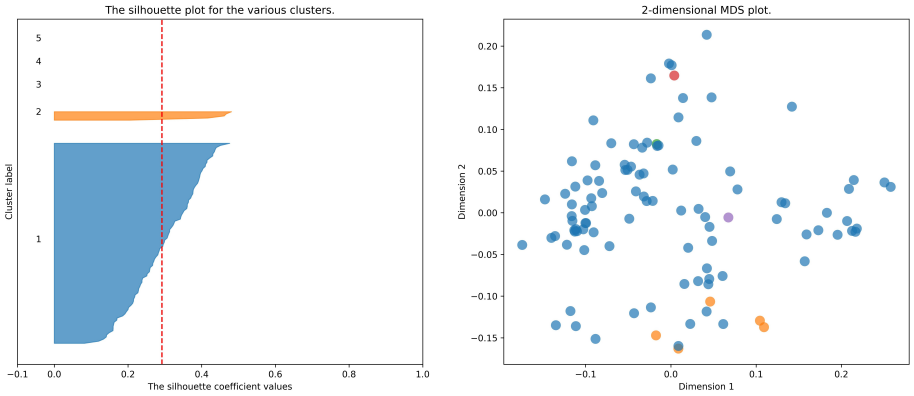


Figure 5.8: Silhouette Plot for UPGMA, 5-cluster solution

as they appear to be three dialect islands in their surrounding areas. These islands can be seen through the use of reference point maps in Figure 5.9, which show the dialect distances from one reference dialect against the entire dialect landscape. The island-like patterns confirms the clusters identified by UPGMA. The islands can also been seen from the distribution of the distances on the box plots next to the reference point maps. WPGMA also displays similar patterns, where many imbalanced clusters consisting of only islands are found.

Despite being a faithful representation of the original distances, I would argue that the cluster solutions we choose should fit the research goals, and not only based indicators alone. Different cluster solutions offer different ways to find structure in the dataset and none of them are particularly 'wrong'. UPGMA is able to identify outliers (the islands in the first few splits) very well. However, having heavily imbalanced clusters does not immediately give you the grand picture of where the bigger dialect groups are. This by no means the results from UPGMA is not useful at all, but I argue that using UPGMA alone is not enough for the purpose of finding dialect groups in the Yue dataset. A classification with numerous dialect groups consisting of one or two dialects is not very informative. For instance, we would consider the 2-cluster solution with the UPGMA algorithm to be the optimal cluster solution, indicated by the highest average silhouette coefficient out of the first ten cluster solutions. This 2-cluster solution consists of one group with one dialect, and another group with the remaining dialects in the data. I would argue that this classification does not contribute to the purpose of

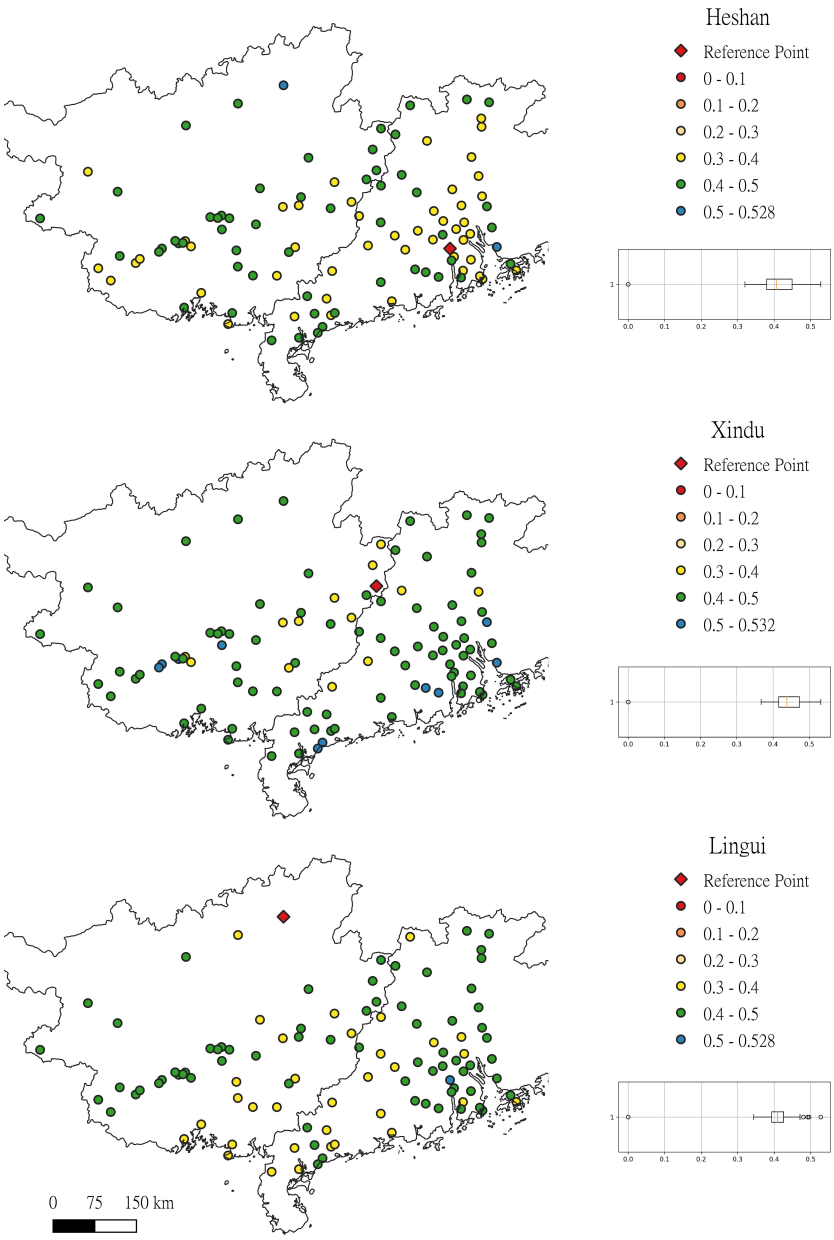
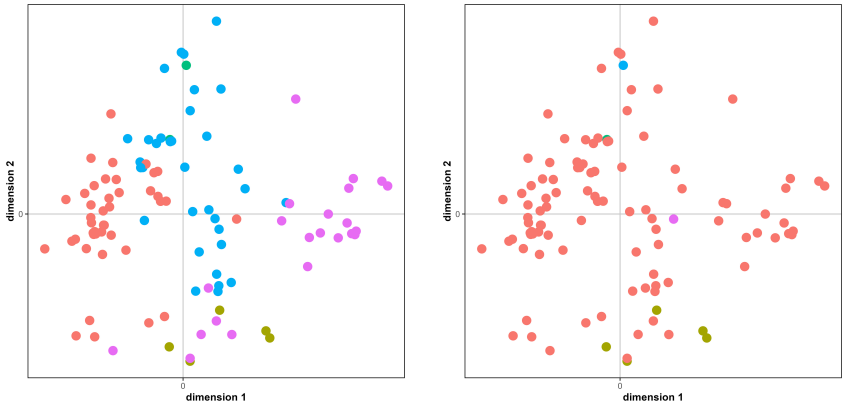


Figure 5.9: Reference point maps of Heshan, Xindu and Lingui. Diamonds represent the reference locations, and the rest of the symbols use colours to show their dialect distances from the reference locations.

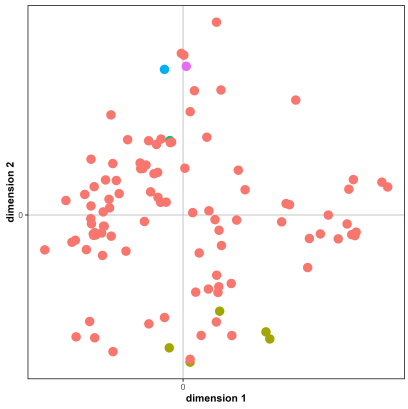
understanding the internal structure of the Yue continuum. Therefore, we need alternative cluster solutions to yield a more insightful internal classification of Yue.

In order to seek a better suited algorithm for the purpose of dialect classification of the Yue dataset, a visual inspection was performed over the MDS plot annotated with the cluster solutions (5-cluster solution, to be consistent with Figure 5.8) for all four algorithms, which can be found in Figures 5.10a to 5.10d.

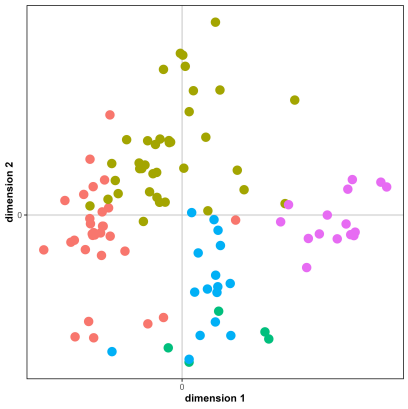


(a) *Complete Linkage*

(b) *UPGMA*



(c) *WPGMA*



(d) *Ward's Method*

Figure 5.10: MDS plots of the 5-cluster solution from four cluster algorithms ($r^2 = 49.7\%$)

As illustrated in Figure 5.10, Ward's method and complete linkage

yield much more balanced clusters than UPGMA and WPGMA. Out of the two, Ward's method have shown a higher average silhouette score (especially for 2- and 3-cluster solutions) than complete linkage. Ward's method seems to be a better choice. In addition, the silhouette plot in Figure 5.11 confirms that the clusters are more balanced than UPGMA. Ward's method is thus chosen on the basis of the clusters yielded are more balanced, which are more informative in a dialect classification.

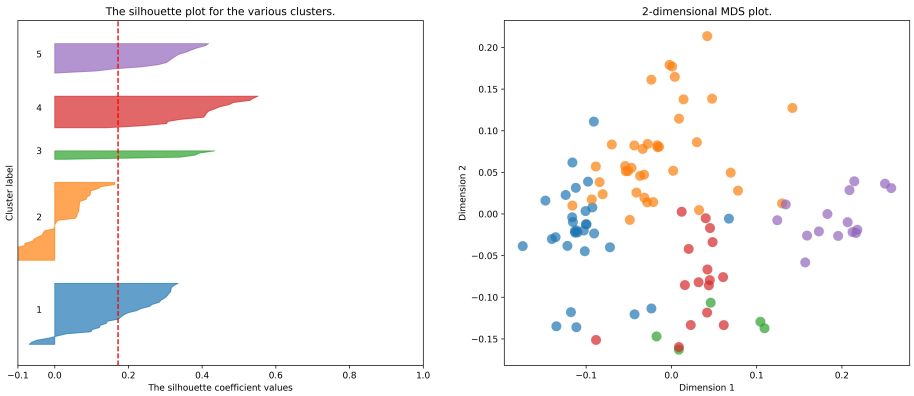


Figure 5.11: Silhouette Plot for Ward's method, 5-cluster solution

The dendrogram for Ward's method is shown in Figure 5.12. In terms of determining the number of clusters, we go back to Figure 5.7 for the average silhouette scores. For Ward's method, the 2-cluster solution has the highest score, and then the score drops at the 3-cluster solution. After that, as we increase the number of clusters, the score does not increase a lot up to the 10-cluster solution. This suggests that the major division between the dialect groups could lie somewhere between 2 and 3 clusters.

The cluster maps in Figure 5.13 show that the first group that splits from the bigger cluster is the Guinan dialect group (in green). This corresponds to the purple dialect group which dimension 1 captures in the MDS plots in Figure 5.10a (complete linkage) and 5.10d (Ward's method). This cluster is the most different dialect group from the rest of the dialects. The fact that Ward's method and complete linkage identified the same group seems to suggest that the Guinan dialect group is a stable dialect group.

The second group that splits off from the red group in the 3-cluster solution are distributed near the coastal area in the middle of the two

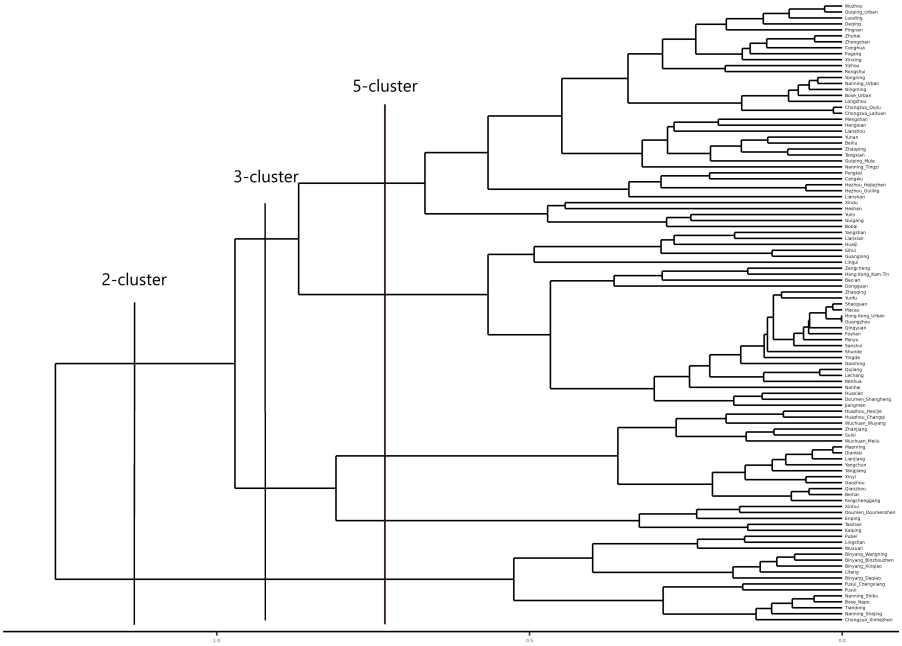
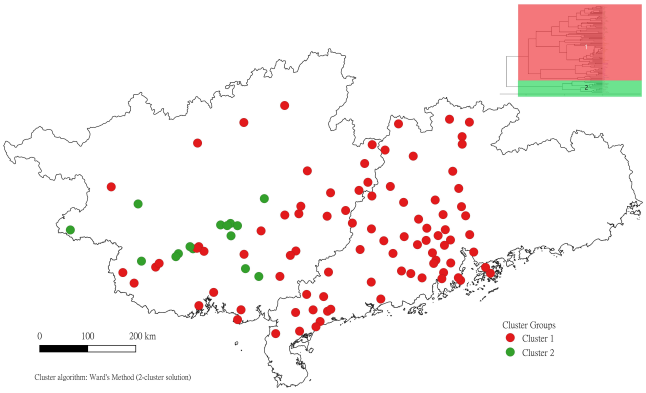


Figure 5.12: Dendrogram of 104 Yue segmental distances using Ward's method

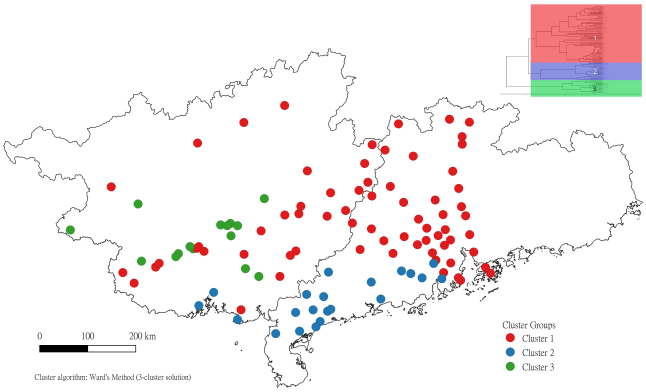
provinces. This cluster corresponds to the dialects distributed in the lower half of the MDS plots in Figure 5.10. Within this cluster of dialects, in Figure 5.10, we also see a group of dialects (the Siyi dialects) which are consistently identified by all four algorithms (in green for the Ward's method, and in brown for the rest). This suggests another stable cluster.

The observation above can be confirmed by the probabilistic dendrogram in Figure 5.14.[10] Siyi dialects (Doumen (Doumenzhen), Xinhui, Enping, Kaiping and Taishan at the bottom of the dendrogram) are grouped together 100% of the time; Guinan varieties (e.g. dialects around Binyang, dialects around Nanning and Chongzuo, and the other dialects in the magenta cluster) are clustered together 92% of the time. The probabilistic dendrogram also indicates that there are other smaller stable clusters in the data. For instance, Guangning, Sihui, Yangshan and Huaiji are clustered together 99% of the time; Nanning (Urban),

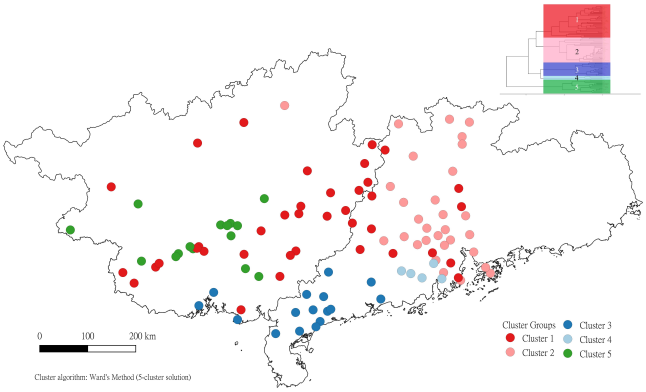
[10]This dendrogram was created using Gabmap, with the following settings: method: group + weighted average, noise: 0.2, exponent: 1.5, limit: 80%.



(a) *2-cluster solution*



(b) *3-cluster solution*



(c) *5-cluster solution*

Figure 5.13: Cluster maps based on the Ward's method

Ningming, Bose (Urban) and some other dialects in the blue cluster are together 97% of the time. Another notable, though slightly less stable cluster (86%) is found in the coastal area, near the Guangdong-Guangxi border (found at the top of the dendrogram), consisting of dialects such as the Yangjiang, Maoming and the Qinzhou dialects. Last but not least, Xindu, Lingui and Heshan are still identified as three very different dialects from the rest of the dialects in the dataset.

The Siyi cluster and the Guinan cluster both show a stable cluster pattern, and they are of interests to the analysis. Based on these results, I argue that they are well-supported to be reflected in the cluster analysis. Since the 3-cluster solution of the Ward's method does not show the Siyi cluster, the number of clusters have been increased until the Siyi cluster appears. This yields the 5-cluster solution. In this solution, the Siyi dialect group is separated from the coastal dialect group (cluster 2 in Figure 5.13b), but the big red cluster (cluster 1) also split into two. As shown in the silhouette plot in Figure 5.11, half of the members of this cluster do not have good separability with the neighbour clusters. This requires a complementary analysis of multidimensional scaling (see the next subsection) to decide whether the separation of the red cluster into two holds.

The 5-cluster map can be found in Figure 5.13c. On this map, Guinan dialects (cluster 5 in the dendrogram on the map) are still distinguished from the rest of the dialect cluster in green. Next, the other two big regions are still in blue (cluster 3 and 4) and red (cluster 1 and 2). The sub-clusters of these two dialect groups are indicated by different shades of blue and red respectively. In this way, several cluster solutions, as well as the hierarchy of the clusters can be easily perceived on the map.

Finally, to compensate for the balanced sizes of clusters as well as the existence of dialect islands, I propose the following solution. Firstly, Ward's method is still used to get an idea of the major (balanced) dialect groups in the data. This will form the basis of the new classification of Yue dialects. The dialect islands identified by UPGMA, on the other hand, complement the results from Ward's method. They are now added on the updated cluster map in Figure 5.15. While the colours reflect the clusters from Ward, the three islands are marked with triangles in the map.

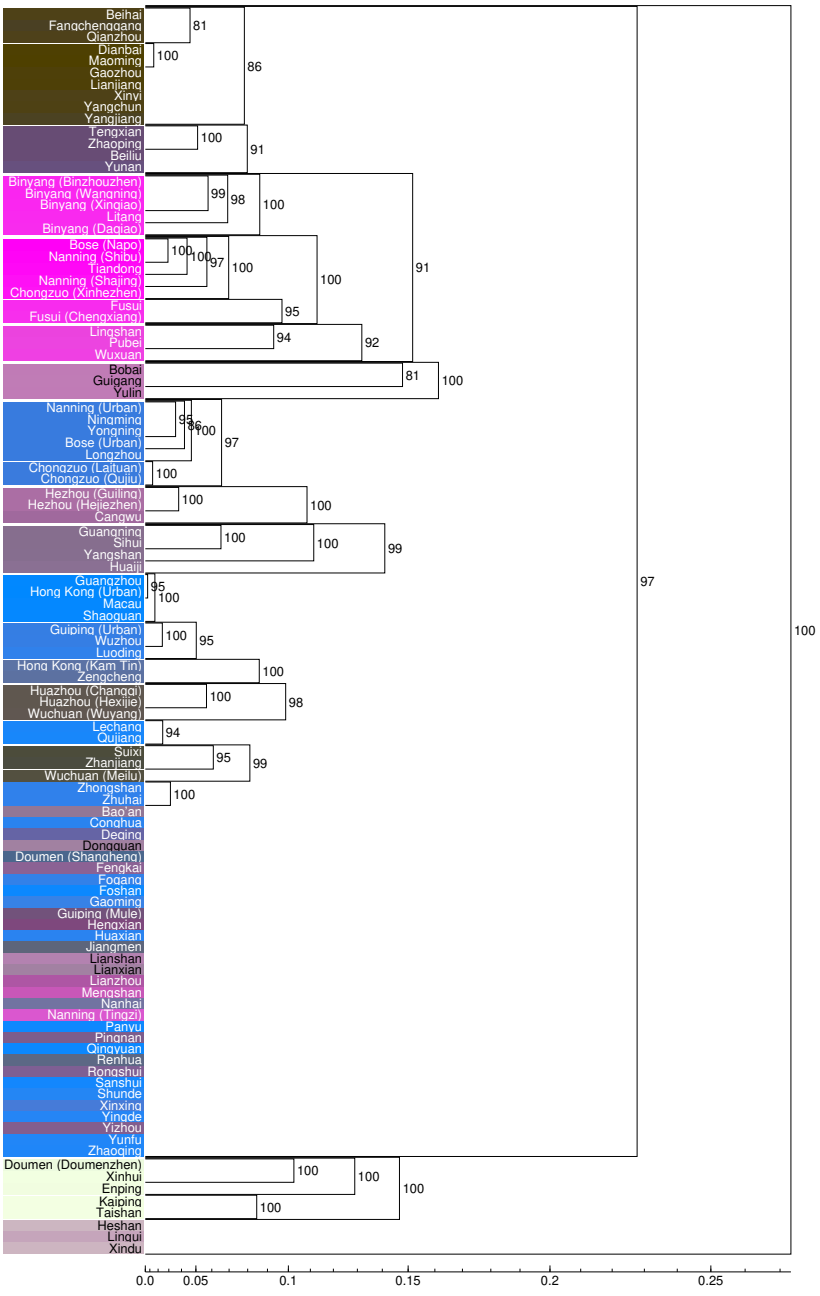


Figure 5.14: Probabilistic dendrogram of 104 Yue dialects

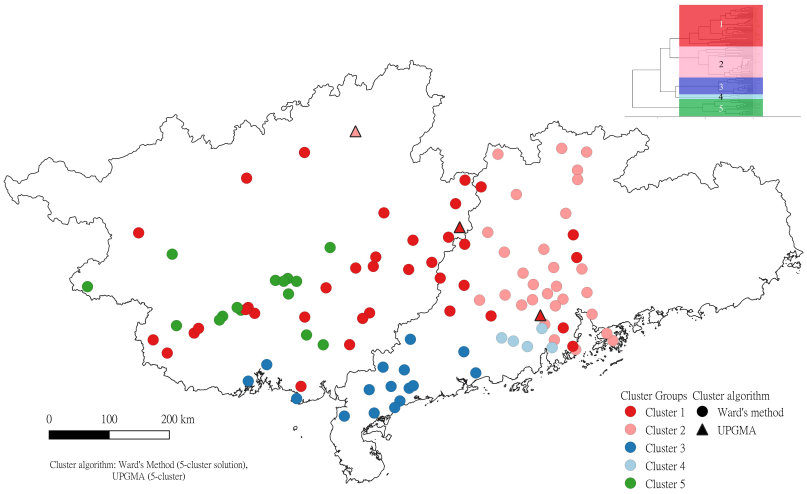


Figure 5.15: Cluster map based on the Ward's method with dialect islands

Multidimensional Scaling

The hard clustering algorithms used in the previous subsection are known for their instability, though through the probabilistic dendrogram and a comparison of different cluster algorithms, two stable dialect groups, namely Siyi and Guinan, can still be identified. However, due to the conflicting strength in the algorithms, we cannot identify both balanced clusters and dialect islands simultaneously. In addition, the algorithms above also return hard clusters, meaning each dialect can only belong to one group and they suggest hard and abrupt dialect boundaries, which is often not the case for dialect variation. To avoid relying on the cluster analysis completely when classifying dialects, multidimensional scaling (MDS) is used as a complementary technique, namely to complement the gradual nature in dialectal variation.

The scree plot in Figure 5.16 shows the accumulated explained variance from dimension 1 to dimension 10. The plot suggests three is the optimal number of dimensions.

Two MDS plots have already been shown earlier in Figure 5.5 and 5.6 to visualise dialect distances in 3 dimensions. However, it is difficult to visualise a 3- (or higher) dimensional space in 2D plots. Now, we turn to projecting the MDS dimensions on a map, which allows us to validate

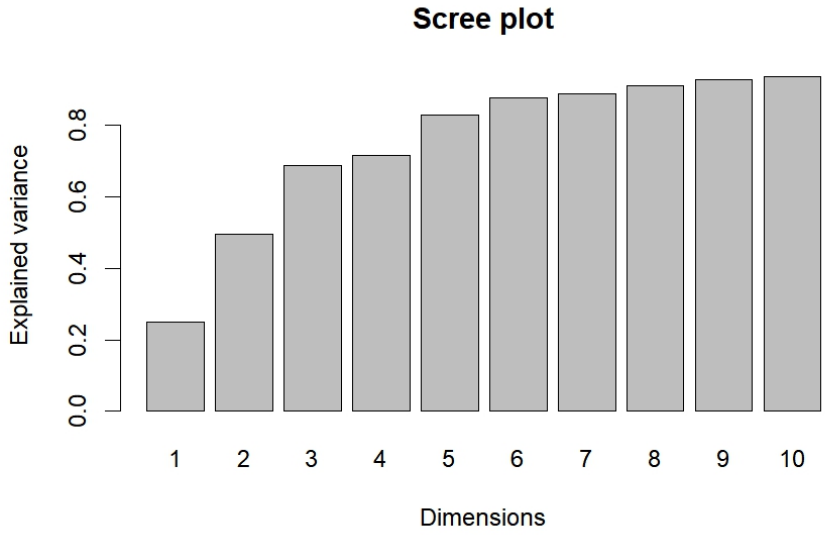


Figure 5.16: Scree Plot for MDS (Segmental variation)

the clusters we observe in a dendrogram.

In terms of the dialect groups reflected in Figure 5.17, three colours stand out, namely green, purple and orange. These colours correspond to cluster 1 and 2 together, cluster 4 and cluster 5 respectively in Figure 5.15. That leaves cluster 3. The area where dialects of cluster 3 are spoken seems to form a continuum with cluster 5 which goes from a maroon/ ruby colour (in southwestern Guangdong) towards a moss/ cherry wood colour, which then connects to the orange cluster (cluster 5). Cluster 3 is a textbook example of showing that cluster analysis returns abrupt boundaries between dialects, even though the dialect transitions are more gradual than cluster analysis suggests.

The green dialects show another continuum pattern. The two-way division from the cluster analysis (Ward's clusters 1 and 2) is not as obvious on the MDS map. It can be seen that dialects in Guangxi tend to have lighter green circles compared to the dialects in Guangdong, but the differences are more subtle (compared to the transition between cluster 3 and 5). Additionally, the green continuum exhibits a great diversity of heterogeneous dialects in the green dialect area, reflected by different shades of green.

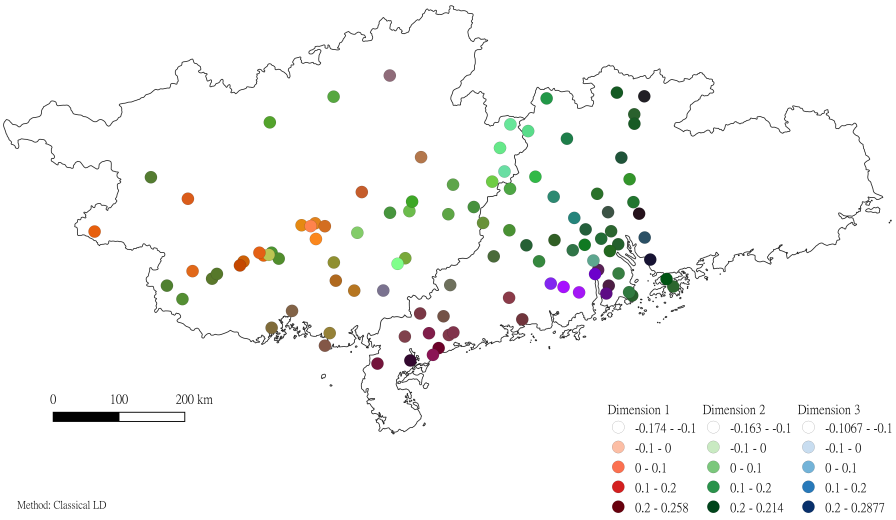


Figure 5.17: MDS map of Yue segmental distances, r^2=0.69

The Siyi dialects can be found in the purple area. The core dialects are in bright purple, but as we move eastward towards the Guangfu area, the colour gets darker and blends into the Guangfu area. This suggests a transition from Siyi to Guangfu. The continua in Yue will be discussed further in the next section.

Last but not least, although Lingui, Xindu and Heshan dialects were identified as being quite different by UPGMA and WPGMA, the MDS map does not particularly show their isolation as dialect islands, unlike the Yulin dialect, which is represented by a bright green circle near the southern end of the Guangdong-Guangxi border.

Classification	Dialect Groups							
LAC	Siyi	Guangfu	Goulou	Yongxun	Gaoyang	Wuhua	Qinlian	Southern Pinghua
Cluster analysis	Siyi (Cluster 4)	Guangfu (Cluster 2)	Central (Cluster 1)		Coastal (Cluster 3)			Western (Cluster 5)
MDS map	Siyi (Purple)	Inland (Green)			Coastal (Red/brown)			Western (Orange)

Table 5.2: Summary of Yue segmental classifications

To sum up, the classification of Yue dialects based on the dialec-

tometric results are presented in Table 5.2.[11] The cluster analysis has merged several groups from the LAC based on the LD analysis. This is on the basis that the dialect distances do not support a 8-group division. Guangfu and Siyi dialect groups retain their traditional names as their distributions are largely the same as the traditional classification. Goulou and Yongxun dialects (mostly spoken in Guangxi) have been merged, now labelled as Central Yue dialects based on its geographical location. Guangfu and Central dialects can be further grouped together as Inland dialects. Gaoyang, Wuhua and Qinlian dialects is the second group of dialects which were merged together and now labelled as Coastal Yue dialects. Guinan Pinghua dialects are now labelled as Western dialects based on the locations of the dialects. Furthermore, the MDS map allows us to see more gradual transitions between the clusters from the cluster analysis. Therefore, the segmental classification of Yue dialects can be concluded to have 4-5 groups, namely Inland (Guangfu and Central), Siyi, Coastal and Western dialects.

5.3 Patterns in Yue variation

Yue is a genetically and typologically very different language from European languages. Yet, we are seeing that Yue also displays a dialect continuum with the MDS analysis. The differences in terms of Yue's different linguistic and socio-historical make up make Yue a great laboratory to explore further what correlates can be found within the Yue continuum, and how it can enrich our understanding of the relationship between language variation and human communication and contact, as well as history and geography.

To explore these topics, the following sections will focus on: 1) a closer inspection of the continua within the Yue-continuum (on the segmental level) and 2) whether (some of the) Yue variation can be explained by geographical and socio-historical correlates. It should be noted that this section only examines patterns of variation and their relatedness to external factors. Details of the linguistic features are looked into in the next chapter.

[11]The LAC classification has been modified due to space. Suixi has been merged with Gaoyang, and Lingui has been merged with Southern Pinghua.

5.3.1 Dialect continua in Yue?

The MDS map in Figure 5.17 gives a general impression that Yue dialects form a continuum. This aligns with what had been found in Europe more than a century ago. Paris (1888) and Meyer (1877) discovered in the late 19th century that there are no clear boundaries between dialect areas.

One can speak of Yue as having four continua within a continuum. This is because if we start the continuum from the right (green dialects in Guangdong), the first continuum can be found going from the Inland (green) dialects to the Siyi (purple) dialects. There are transitional dialects between the two dialect groups, namely Doumen (Doumenzhen) and Xinhui (see also the red dots in Figure 5.6, Dimension 3 on the left). The second continuum can be found going from the Guangdong Coastal dialects (ruby circles) towards the Coastal dialects (in mossy colour) in Guangxi, which then forms a continuum with the orange circles in the north, which is the Western dialect area. The third continuum is the entire green dialect area stretching from the Guangdong to the western edge of the Inland dialect group. For the last continuum, there is a transition going from the Inland area to the He-Lian (Hezhou-Lianshan) area. The descriptions above are visualised on a map in Figure 5.18.

Continuum 1: Inland to Siyi

The Siyi dialects are very different from the rest of the Yue dialects, indicated by its unique purple colour in the MDS map. However, the presence of transitional dialects (indicated by the in-between 3rd dimension values, see Figure 5.6) between Guangfu and Siyi suggests some contacts between the two dialect groups. The transition of segmental features have indeed been found in Sung (2025). The case of Siyi shows that of the transitional dialects are able to blur the dialect boundaries between the two very distinct groups of dialects.

Continuum 2: Coastal to Western

de Sousa (2020:267) stated historically, Yue had expanded westward several hundred years ago towards the Pinghua region, and during this expansion, Yue "absorbed linguistic elements from the pre-existing Sinitic languages (most of which probably resembled Pinghua) and the indigenous languages (most of which probably resembled Zhuang)". Figure 5.18 seems to partially support this description. A continuum can be

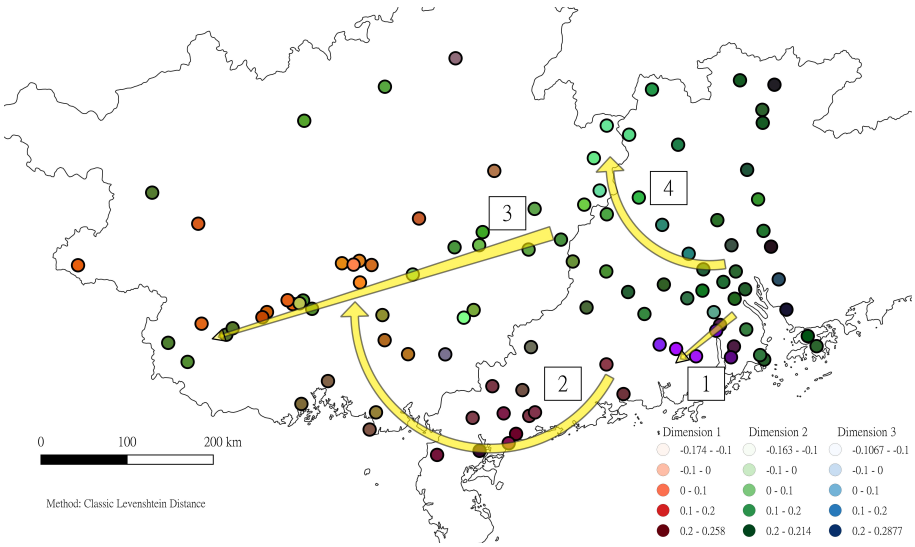


Figure 5.18: Continua within a continuum in the Yue-speaking area. (1) Inland-Siyi continuum, (2) Coastal-Western continuum, (3) Inland continuum, (4) Inland-He-Lian continuum. The arrows indicate each of the continuum with a starting point from eastern Guangdong (the heartland of the Guangfu dialect area).

found from the Coastal dialect area towards the Western dialect area. However, the gradual expansion does not seem to start from the Pearl River Delta, as de Sousa described. Perhaps this is due to the gaps in the localities from the existing dialect surveys. For instance, there is a clear gap (near the '2' in Figure 5.18) between the Guangfu dialects and the Coastal dialects (it seems to be an understudied dialect area of Yue for unknown reasons). Even within the Coastal dialects, we can see a big gap between Yangjiang and Yangchun and the rest of the Guangdong Coastal dialects. Generally, more data are needed to verify and examine the Yue expansion stated by de Sousa, and the transitions between the Inland, Coastal and Western dialect areas.

Continuum 3: Inland

The historical formation of Continuum 2 does not apply to Continuum 3, even though it resembles the description. This is because the formation of Continuum 3 happened much later, around 150 years ago during the

First Opium War (de Sousa 2020:268). The link between Continuum 3 and external factors is discussed in Section 5.3.2 below.

Continuum 4: Inland to He-Lian

The last continuum goes from the Guangfu area to the He-Lian (Hezhou-Lianshan) area. Although both areas share the main colour green on the MDS map, they clearly show different shades. While the Guangfu area shows darker green, the He-Lian area (and Yulin) has a much brighter green than most of the Inland dialects. This suggests that dialects in the He-Lian area have their own unique features, although it is unclear what unique features these dialect exhibit at the moment, as this area is a rather understudied area.

5.3.2 Correlates with Yue variation

Another interesting aspect of Yue variation which is worth exploring is whether some of the variation we see can be explained through geographical and historical, socio-political correlates. Bloomfield (1933:328) proposed that "weakness in the network of oral communication" can be the limit of the spread of an innovation. These weaknesses include physical barriers like mountains as well as political boundaries. On the other hand, diffusion can occur along a traffic route. This can be a major road or river (c.f. Figure 27 in Niebaum and Macha 2014). Yue also exhibits these traits, as we will see in the following examples.

Political border effect or physical barrier?

The northern part of the Yue-speaking area near the border between Guangdong and Guangxi consists of Inland dialects. This area is mountainous, which is an interesting area to explore the effect of political border vs. physical barrier.

The Lianshan dialect is located in northern Guangdong (see the map in Figure 5.19). We might expect that Lianshan would be linguistically closer to other nearby dialects around the area, namely Yangshan, Lianxian and Huaiji. However, this is not the case. The linguistic distances between the Lianshan and Hezhou dialects, which are spoken in Guangxi, are much lower than the other dialects spoken in northern Guangdong (indicated by the different shades of green).

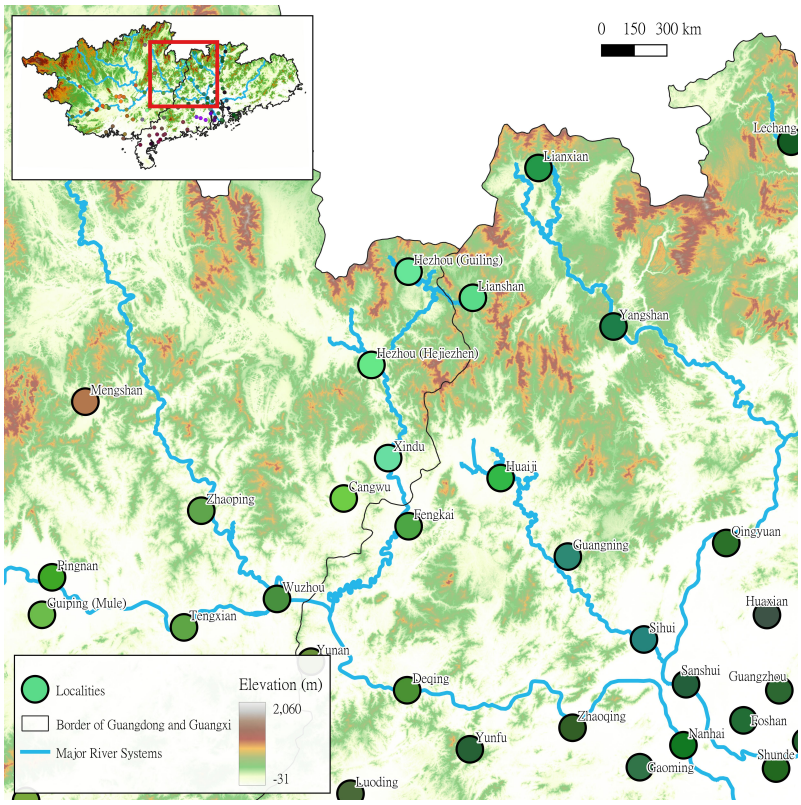


Figure 5.19: MDS map of the northern border between Guangdong and Guangxi overlaid with elevation and major river systems

The differences between Lianshan and, for example, Yangshan can be explained by physical geography. Firstly, Lianshan and Yangshan have a mountain situated in between, separating the two locations. Furthermore, for people from Lianshan to travel to Yangshan, they either have to climb the mountains or travel along the river through Guangxi, then back to Guangdong. This is related to what Bloomfield (1933:46) calls "density of communication". Speakers accommodate to people around them all the time. For mobile speakers of Lianshan, they are more likely to encounter people from Hezhou than Yangshan, simply because the accessibility (travel distance) from Yangshan is lower.

Adding elements such as elevation and major river systems can help us explain some of the variations we see on an MDS map. They can

tell us indirectly what contact and travel distance would be like for the speakers, which are useful in explaining why the Lianshan dialect is more similar to Hezhou dialects than to Yangshan, despite both Lianshan and Yangshan look close to each other on a (blank) map and are both spoken in northern Guangdong, near the border.

Geographical and socio-historical correlate

Continuum 3 in Section 5.3.1 shows a stream of green dialects piercing through the orange dialects in Guangxi, coming from Guangdong. As shown in Figure 5.20, if the MDS map is once again overlaid on a physical geography map consisting of elevation and major river systems, we will see that the green dialects in Guangxi are largely distributed along the major rivers.

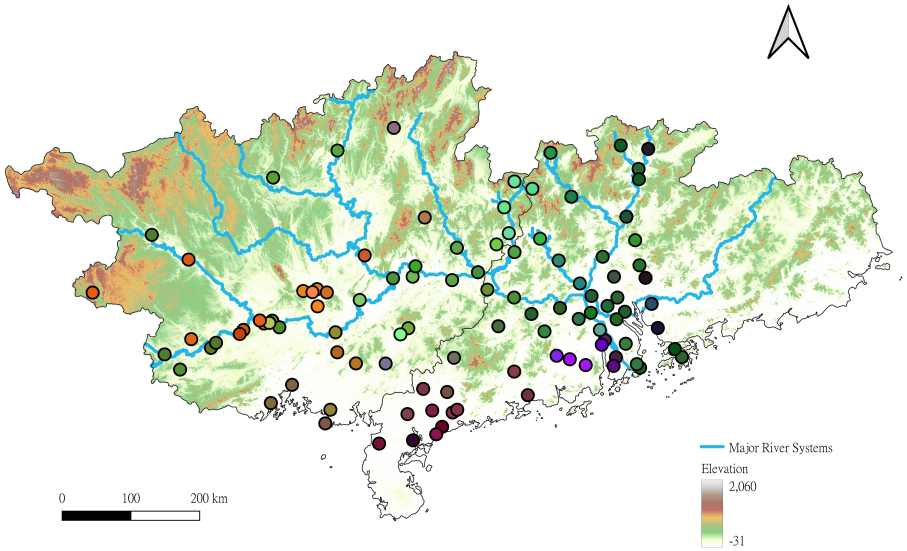


Figure 5.20: MDS map overlaid with elevation and major river systems

The rivers seem to provide an explanation for the distribution of the green dialects in Guangxi, but the real cause of this distribution is a socio-historical factor. de Sousa (2020:267-268) and Yu (2016:100) speak of two types of Yue spoken in Guangxi, namely the 'native' Guangxi Yue

and Guangxi Cantonese. The 'native' Guangxi Yue varieties include traditional Goulou and Qinlian dialects. These varieties are spoken in large areas, and they spread to Guangxi from the East much earlier.[12] These varieties have very marked differences from each other (Yu 2016:100). On the other hand, the second group of Yue dialects (traditional Guangfu and Yongxun dialects) are labelled as 'Cantonese' by de Sousa, which came to Guangxi later. They are situated as enclaves along the Xijiang river in the Guangxi Yue continuum (especially obvious when they appear next to the Western dialects in Figure 5.20). The traditional Guangfu varieties in Guangxi (situated near the Guangdong-Guangxi border, near Wuzhou and Hezhou) are not so different from Cantonese in Guangdong, since they retained close ties with the Guangfu region in Guangdong. Traditional Yongxun dialects, on the other hand, are the dialects which are distributed along the river all the way to the west of the province. Another interesting property of these dialects is that they have recognisable phonology of Cantonese, but they also received noticeable influence from Zhuang (de Sousa 2020:268).

The reason why the Cantonese varieties (varieties in green) are distributed along the river is due to migration. The traditional Yongxun dialects were formed around the end of the First Opium War (1839–1842), and more speakers migrated after the Second Sino-Japanese War (1937-1945). The ancestors of the Yongxun dialect speakers emigrated from the Pearl River Delta (heartland of the Guangfu region) and moved along the Xijiang river. The reasons for emigration include trade and relieving population pressure in the Pearl River Delta, as well as escaping from war (Yu 2016:100). This population movement is reflected on the MDS map.

Reflection of older political border

The similarity between traditional Gaoyang and Qinlian dialects can be explained by the older provincial border between Guangdong and Guangxi. In Figure 5.21, the MDS map is now overlaid on an administrative map which reflects the provincial border before 1952. The pre-1952 borders have been stable since the Ming dynasty (Tan 1982; Shi 1981). The major differences between this map and the present-day

[12] According to Yu (2016), 'native' Yue varieties spread to Guangxi during the Tang and Song dynasties, while de Sousa (2020) suggested during the Ming and early Qing dynasties.

Guangdong-Guangxi borders are: 1) Huaiji used to belong to Guangxi and the Qinlian area used to be part of Guangdong. The current provincial borders were the results of the redefinition of provincial borders in 1965 (Shi 1981).[13]

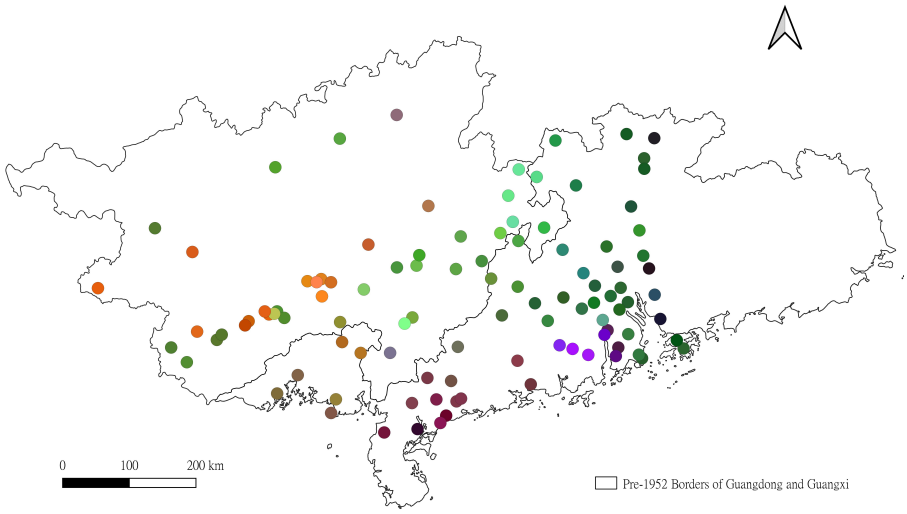


Figure 5.21: MDS map with Pre-1952 Provincial Borders

Haag (1898) observed that only 50 years are needed to notice the border effect from a newly established political border. The Coastal dialects had been spoken in the same province for several hundred years, until more recently. Haag's observation might explain why, for instance, the Coastal dialects share a number of traits, and thus identified as part of the same cluster, despite being spoken in two difference provinces nowadays. However, there could be recent changes in these dialects which are reflected in the colour differences on the MDS map within the Coastal dialects spoken on different sides of the Guangdong-Guangxi border.

[13]There have been changes that went back and forth between 1952 and 1965, namely with the membership of the Qinlian area being part of Guangxi or Guangdong. The Qinlian area eventually became part of Guangxi and remained there after 1965, which is what we see in the Guangxi administrative boundary in the present-day.

5.4 Conclusion

In Chapter 2, a number of classifications of Yue dialects have been reviewed, and most of them are suffering from the 1) lack of justification in the subgrouping process, 2) lack of completeness (often focused on one province) and 3) lack of data. For the evaluation of the traditional classification, the LAC has been chosen, as it is the most widely acknowledged classification, and there are maps to help determining which dialect group a dialect belongs to in both provinces.

The comparison between the LAC classification and the dialectometric analysis shows that not all dialect groups in the LAC are justified. With the help of cluster analysis and multidimensional scaling, 5 dialect groups and 4 continua have been identified in the Yue-speaking region.

Although Yue is a genetically and typologically different language from European languages, they show a lot of similarities in terms of how dialects vary. For instance, Yue forms a continuum, even for some of the most distinctive dialect groups. In addition, some of the variation can be explained by extralinguistic correlates, such physical barriers, population movements, and changing provincial borders.

Dialect classification is not the end goal, though. Although now there is a more data-driven classification for Yue dialects, we still do not know what features each dialect group possesses. Furthermore, Zhan (1988) proposed to use distinctive, characteristic features to separate each dialect group as the criteria for the groupings. To what extent is this the case for the Yue classification in e.g. the LAC? The following chapter discusses a novel methodology for extracting features after a dialectometric analysis, followed by a detailed discussion on the features each dialect group possess.