Advancing explanatory and tonal dialectometry

Sung, H.W.M.

CHAPTER 8

Tonal Variation of Yue Dialects[1]

8.1 Introduction

In the previous chapter, four existing tone distance calculation methods have been compared and evaluated. It was found that none of the existing methods are adequate for the Yue dataset, due to the fact that most of the tone representations used in these methods cannot differentiate the vast majority of the tones in the data. For the ones that can do so, they do not yield linguistically and perceptually coherent tone distances. In the discussion, I argued that Onset-Contour-Offset (OCO) and the Gandour-Harshman-Tang tone distance measurement (GH-T) should be considered further for the application to dialectometry, although with some modifications. OCO works hand-in-hand with Levenshtein distance, which is also used in the segmental analysis (see Chapter 3). Moreover, OCO being a string representation also makes it easier to be combined with the segmental transcriptions before distance calculation, considering the future possibilities with combining segments and tones in one single analysis. For the reasons above, OCO has been selected as the tone distance calculation method for the improvement and modifications in this chapter.

[1]This chapter is based on Sung and Prokić (2024b).

Another observation the previous chapter made is that tonal variation seems to behave differently from segments. However, this observation cannot be validated without an aggregate analysis on tones with a more adequate tone distance measurement.

Recently, Sung et al. (2024) proposed a modified version of Yang and Castro's (2008) tone distance calculation method (OCO), which made an improvement on the differentiation of tones in the Yue-Pinghua dialect dataset. This methodology now enables dialectologists to explore further issues in the tonal variation of (lesser-studied) tonal languages outside Europe in combination with the dialectometric methods.

The chapter is structured as follows: Section 8.2 explains how dialect tonometry works. The results of the analysis on tonal variation are presented in Section 8.3, and Section 8.4 compares tonal variation with segmental variation. Lastly, the discussion and conclusion can be found in Sections 8.5 and 8.6 respectively.

8.2 Dialect Tonometry

The main goal of this chapter is to explore how tones vary geographically within a given area, in this case, the Yue-speaking area in Guangdong and Guangxi provinces. As pointed out in Chapter 7, the majority of the previous tone distance measures are not adequate in dealing with a large variety of tones in a bigger dialect dataset. This chapter uses a modified version of OCO (Sung et al. 2024), which can account for more than 98% of the tones in the Yue dataset.

The measurement of tone distances can be called *tonometry*.[2] When one extends the application of a tone distance calculation method to the study of the dialectal variation of tonal languages in dialectometry, this approach is referred to as *dialect tonometry*.

8.2.1 The Modified OCO (mOCO) representation

OCO's biggest problem is not being able to distinguish enough tones for the Yue dataset. Sung et al. (2024) focused on this aspect and made some adjustments to the tone representation. Their Modified Onset-Contour-Offset (mOCO hereafter) representation still derives from Chao's (1930)

[2]The term 'tonometry' is also used for a medical test which measures the pressure inside an eye, but this meaning is unrelated to the usage in this dissertation. In linguistics, it has no prior usage.

tone letters. However, two aspects were modified, namely the number of contours for the onset and offset, as well as length being indicated. The adjustments are introduced below in more detail.

Firstly, the pitch levels are expanded from originally differentiating three levels (merging 1 and 2 and merging 4 and 5) to distinguishing all five levels, following Chao's (1930) tone letters. The modification creates a five-level contrast by having HH (5), H (4), M (3), L (2) and LL (1). The double-character representation 'HH' and 'LL', is designed to add weights to pitch levels that are more distant. For instance, 'H' and 'L' are immediate neighbours of 'HH' and 'LL' respectively. When the Levenshtein distance algorithm (see Section 3.3.1) is applied to the representation, the distance between 'HH' and 'H' or 'LL' and 'L' is 1; all other pitch levels cost a difference of 2. The other modification has to do with tone length. The differences in tone length are usually found between checked syllables and non-checked syllables. A superscript 'ʰ' is used to represent phonetically short tones, which indicates a difference of 0.5 with long tones in the distance calculation.[3]

Other than the tone representation, the pairwise tone distance calculation remains the same as Yang and Castro (2008), where Levenshtein distance is applied to the mOCO-represented tone directly (see Section 7.4.1). From now on, 'mOCO' refers to the tone calculation method which applies Levenshtein distance on the tones in the mOCO representation, unless specified (e.g. as a tone representation).

mOCO maintains the perceptual dimensions of 'direction' and 'average pitch', which have been identified as the two most important perceptual dimensions in Gandour and Harshman (1978). The two dimensions can be seen through the MDS plot in Figure 8.1, where the tone distances found in the Yue dataset were calculated with the mOCO representation, and projected on the plot (see Section 7.5 for the comparison with previous tone distance calculation methods). With the mOCO representation, we can now differentiate 72 out of 73 tones (98.6%) in the Yue dataset. The only pair of tones the mOCO representation cannot distinguish between is '232' and '242'. However, up to now, mOCO remains

[3]Superscripted characters are counted as a difference of 0.5 in the Levenshtein algorithm implemented in *Gabmap* (default settings), if the last character(s) of the *Offset* of both tones (but not the length) are identical. This implies that the tone length is only differentiated if the final character of the offset in the mOCO representation is identical. Please note that LED-A.org does not have the same implementation of the superscript 'ʰ'.

the best tone distance calculation method for the current exploration of tonal variation in Yue, following Sung et al. (2024). Moreover, being able to differentiate a high variety of tones in the current dataset should also make the tone distance method sufficient for most other tonal languages in Southeast Asia, and perhaps in other parts of the world (given that the same tone notation is used in the documentation, so that the conversion of the representation can be done).

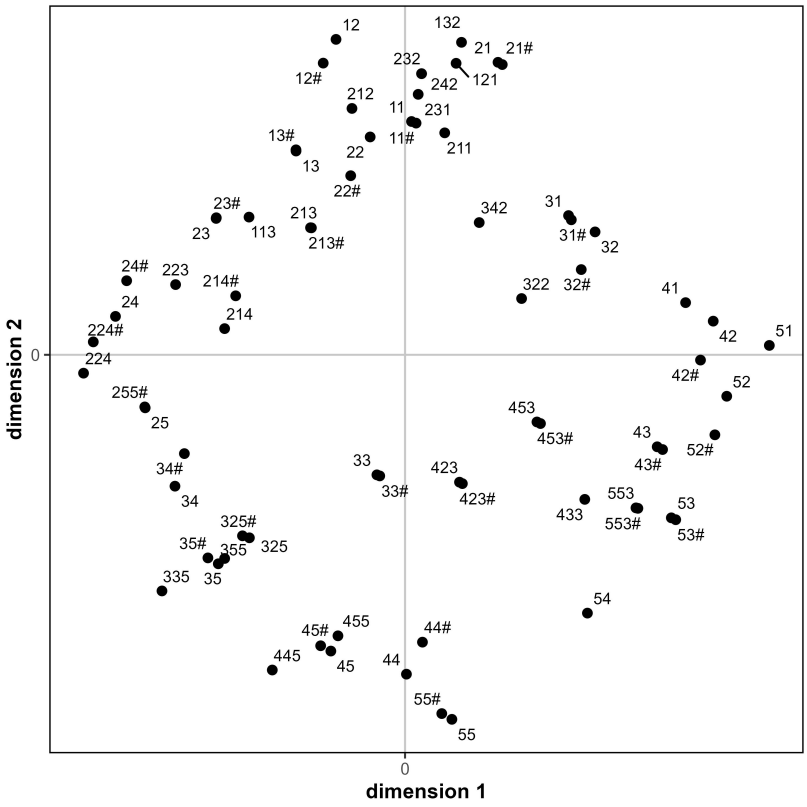


Figure 8.1: 2-dimensional MDS plot of tone distances using mOCO representation (from Sung et al. 2024, $r^2 = 0.47$)

8.2.2 Calculating aggregate tone distances

When tone distances can be calculated between each pair of items, aggregate tone distances can then be obtained to explore various research

questions in dialectometry, including dialect classification and comparisons of different linguistic levels. In dialect tonometry, an aggregate tone distance between each pair of dialects in the dataset can be obtained using the procedures described below. The steps in obtaining the aggregate distances are shared with the existing approaches used in dialectometry (e.g. in Heeringa 2004). The distances obtained from the Yue-Pinghua dataset are analysed in the following two Sections.

To get the aggregate (tone) distance between a pair of dialects, firstly, the tone distance between the tones of each lexical item is calculated using mOCO. Next, we divide the sum of the tone distances of all lexical items by the number of lexical items compared. These procedures are repeated for all the pairs of dialects in the dataset. All the aggregate distances are stored in a distance matrix. Lastly, the pairwise distances can be further analysed using cluster analysis and multidimensional scaling, and visualised through dendrograms, MDS plots as well as maps based on the results of these two techniques.[4] More details on Levenshtein distance, cluster analysis and multidimensional scaling can be found in Chapter 3. The procedures described above are illustrated in Figure 8.2.

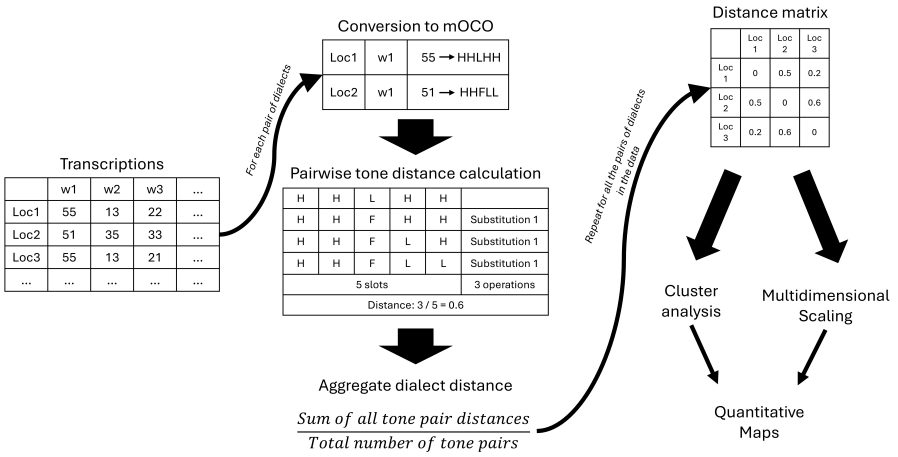


Figure 8.2: Dialect tonometry procedures

[4]The procedures from the aggregated distance calculation onwards are identical with calculating Segmental distances with Levenshtein distance described in Chapter 5.

8.3 Tonal variation under the scope of dialectometry

One of the classical debates in the history of dialectology is whether there are dialect areas (with abrupt boundaries) or a continuum of dialects. The existence of dialect boundaries was questioned before modern dialect geography (Meyer 1877, Paris 1888), and dialectologists have found evidence for the latter (Chambers and Trudgill 1998, Heeringa and Nerbonne 2001). While comparisons of different linguistic levels, namely lexis, morphology, phonetics and syntax, have been investigated before on a few European languages (Spruit et al. (2009) for Dutch, Montemagni (2008) for Tuscan dialects, Scherrer and Stoeckle (2016) for Swiss German), tone as a separate linguistic level has not been investigated under the dialectometric perspective. This is partially because European languages are not tonal[5] and dialectometry has not gained full popularity yet in dialectological traditions of tone languages.

In the following subsections, tonal variation of Yue dialects will be analysed under the aggregate perspective. This raises the question of whether tonal variation, like segmental variation, shows a continuum pattern. Next, a comparison between tonal and segmental variation (from Chapter 5) is made.

8.3.1 Tonal variation between dialects

Does tonal variation form a dialect continuum, like segments? It seems that dialects can indeed show gradual differences in terms of tones, when geography is not considered. Figure 8.3 is a multidimensional scaling (MDS) plot of the aggregate tone distances[6] between Yue dialects. There are a few small clusters, with the most obvious one in the lower right (circled in red, with Taishan, Enping, Doumen dialects etc., which correspond to the traditional Siyi dialects). However, there are no other clear tight clusters which are isolated from the rest of the dialects (based on the 2-dimensional MDS plot), which suggests a continuum.

Sung et al. (2024) and Sung and Prokić (2024b) have shown that dialects can vary gradually in terms of phonetics, tonemes as well as lexical distribution of the tones. These observations were based on the

[5]Pitch accent languages are not considered here.

[6]Cronbach's alpha is 0.98, calculated with *Gabmap* (Nerbonne et al. 2011; Leinonen et al. 2016).

patterns shown in six Siyi dialects (circled in red) and seven Guangfu dialects (circled in blue). However, the analyses above are solely based on the distances between dialects, without consideration of the geographical proximity of the locations where the dialects are spoken.

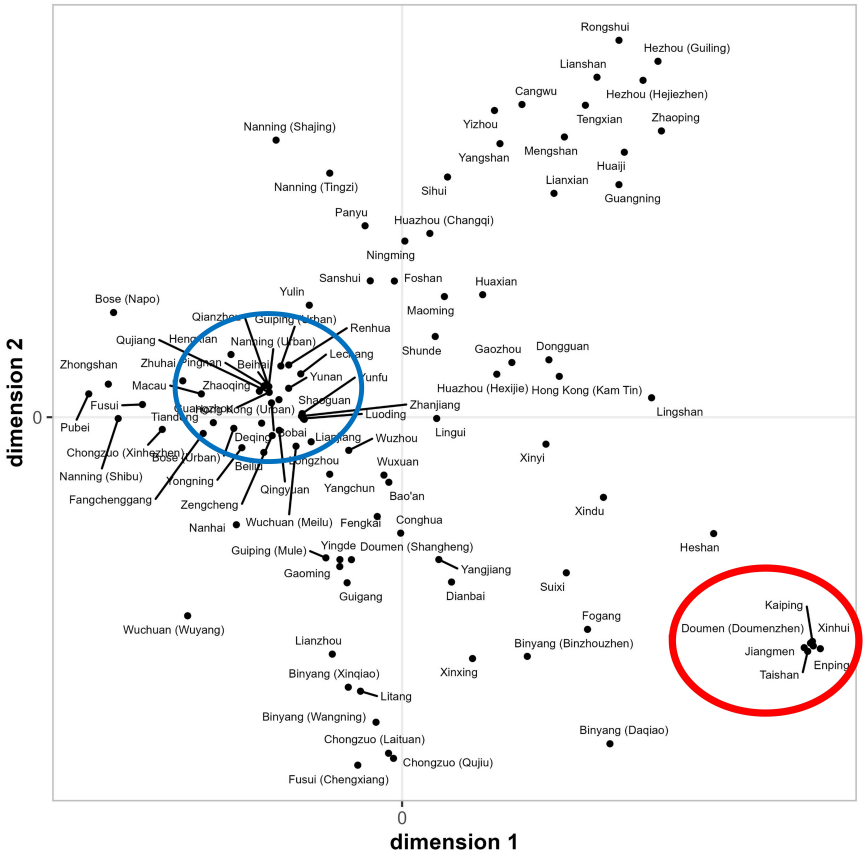


Figure 8.3: MDS plot of aggregate tone distances between Yue dialects, $r^2 = 0.49$

8.3.2 A geographical continuum?

Section 8.3.1 pointed out that dialects can form a dialect continuum. However, geography was not considered in an MDS analysis; an MDS plot does not depend on any geographical information in order to show dialect distances. A follow up question related to geography would be

whether a dialect continuum (i.e. geographically neighbouring dialects differing only a little at a time) is also found geographically.

An MDS map of the tonal distances of Yue can be found below in Figure 8.4.[7] This map shows several major geographical patterns. One of the areas that stands out is the Siyi dialect area. Unlike the surrounding dialects, it has a scarlet red colour, opposed to the surrounding green, purple-blue circles. This matches the MDS plot in Figure 8.3 (circled in red), as it is a tight cluster isolated from the rest of the dialects. Next, different shades of green can be found in the Eastern Inland area, and these circles extend towards the west to the Zhan-Mao area, i.e. south-western part of Guangdong (but is broken off by some brownish circles) as well as to western Guangxi. These dialects resemble the traditional Guangfu group or Inland dialect region (Guangfu sub-group) and the Coastal dialect region in the segmental analysis (see Chapter 5). A big group of purple-blue circles can be found in central Guangxi, namely around Binyang and Chongzuo. These dialects resemble the traditional Guinan group or Western dialect region in the segmental analysis. Lastly, a group of dialects clustered together in the north of the Yue area can be found with pale green circles. These dialects correspond to the Inland dialects (Central sub-dialect group) in the segmental analysis.

In order to quantitatively capture the areal patterns identified by manual inspection of the MDS map, cluster analysis is applied on the aggregate tonal distances. Out of the few agglomerative hierarchical cluster algorithms introduced in Chapter 3, they all show the divisions of the dialect regions identified above by visual inspection. The differences between these cluster solutions lie in 1) the priority of detecting outliers (UPGMA), 2) a higher cluster solution is required to identify the Siyi dialect group as a distinct dialect group (complete linkage, which requires at least a 9-cluster solution to see the Siyi dialect group), 3) some dialect regions have a different geographical extent (the purple cluster is detected by Ward's method, UPGMA and complete linkage, but the latter two classify the dialects with the light blue/ blue colour as part of the green cluster). Ward's method, which minimizes variance in the

[7]Details on the reprojection of the MDS coordinates to the RGB colour spectrum can be found in Section 3.3.2.

[8]This map and the following maps are zoomed more into the area of interest. The blank spaces, which include mostly Guangdong but also part of Guangxi consist of linguistic areas which are not Yue-speaking. These are omitted for a better view of the variation in the studied area.

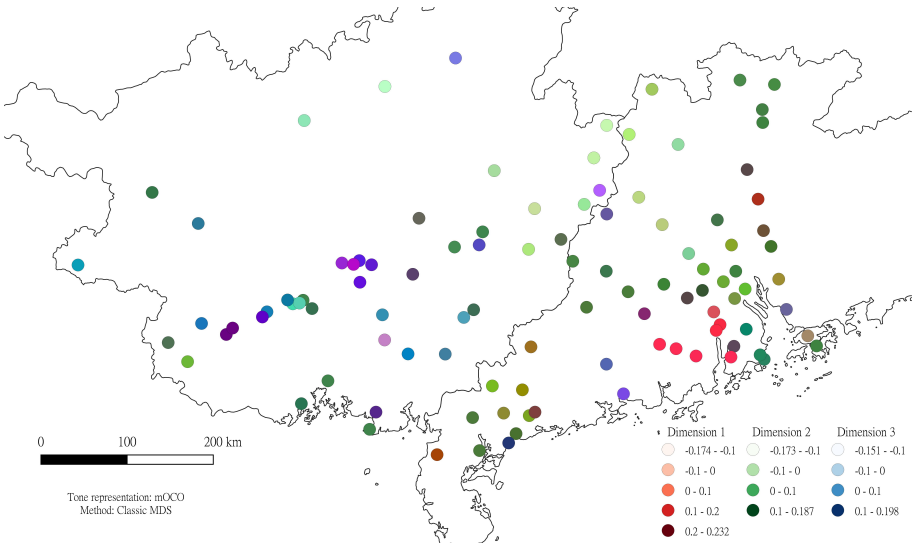


Figure 8.4: MDS map of tonal variation of Yue dialects, $r^2 = 67.2\%$[8]

groups which results in balanced clusters, shows closer resemblance to the visual inspection. In addition, it does not consider isolated dialects as clusters (in the lower cluster solutions) on their own (like UPGMA), and it can identify the outstanding dialect groups like the Siyi dialects early on. For the above reasons, Ward's method is used in complement with the visual interpretation of the MDS map.[9]

The 4-cluster solution of Ward's method is shown through the cluster map in Figure 8.5. On this map, the Siyi dialects (Cluster 2) are identified in red; Cluster 3 corresponds to the Western dialects (both purple and blue dialects in Figure 8.4); Cluster 4 corresponds to the dialects in the light shade of green on the MDS map and lastly, Cluster 1 correspond to the dialect group with the Inland dialects (main in green) in Figure 8.4.

Cluster analysis returns a few geographical outliers, including Lingui, Yangjiang, Yangchun, Dongguan, Xindu and Fengkai in Cluster 3, as well as Fogang and Suixi in Cluster 2. Looking more closely to one example, Suixi (the western outlier of Cluster 2), it is clustered together with the Siyi dialects. If we look at the tone correspondences between Taishan and Suixi (see Appendix D), we can see that 60/129 (46.7%)

[9]Cluster analysis was performed in *R* (R Core Team 2024).

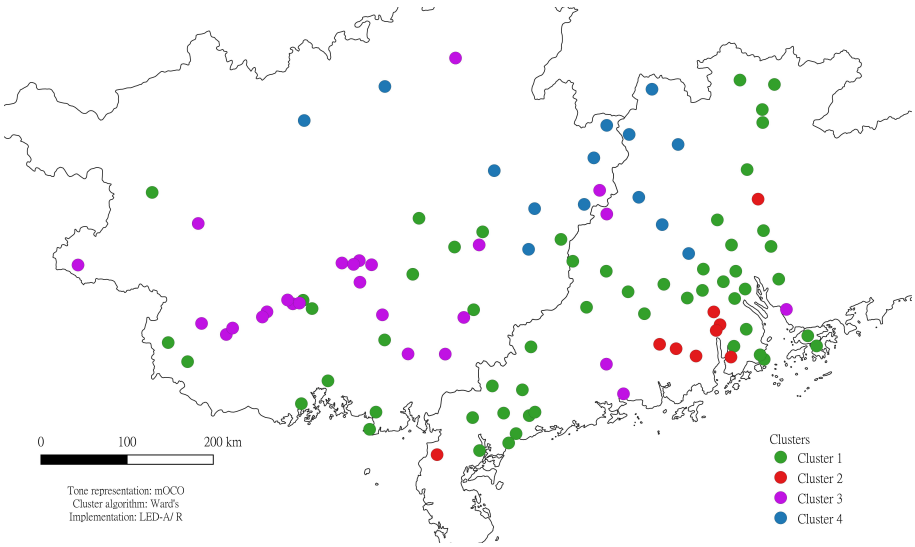


Figure 8.5: Cluster map of tonal variation of Yue dialects, (Ward's method, 4 clusters)

tone pairs are phonetically identical. These two dialects are not often associated together, because of their geographical proximity and their segmental differences (as stated in the classifications in the literature). Cluster analysis is able to highlight distant dialects which share a higher degree of similarity than we expected (based on the *Fundamental Dialectological Principle*, Nerbonne and Kleiweg 2007). The discovery of these dialects could potentially be useful in uncovering contact or historical formation information, which is outside the scope of this chapter.

It should be noted that Cluster 3 (corresponding to the purple-blue circles in Figure 8.4) can be split further into two sub-groups in a 5-cluster solution (with Ward's method). This is also a valid cluster solution for the following reasons. On the MDS map, although (different shades of) blue and purple indicate that these dialects are quite similar to each other when compared with the rest of the dialects, they are still giving signals that they share quite some differences. These differences are supported if we look at a 3D plot of the MDS analysis, which is shown in Figure 8.6. The colours of the dialects uses the exact same RGB projection based on the MDS coordinates as the MDS map in Figure 8.4. We can see that the blue cluster, despite forming a continuum with the purple cluster (e.g. Lingui and Yangchun dialects), is a smaller cluster

on its own. What the blue and purple clusters share is that they both have high dimension 3 values, which separate them from other dialects. However, because the blue dialects are not distributed continuously in terms of geographical proximity of the dialects, in the further analyses of the MDS map and the cluster analysis, the 4-cluster solution will be used.[10]

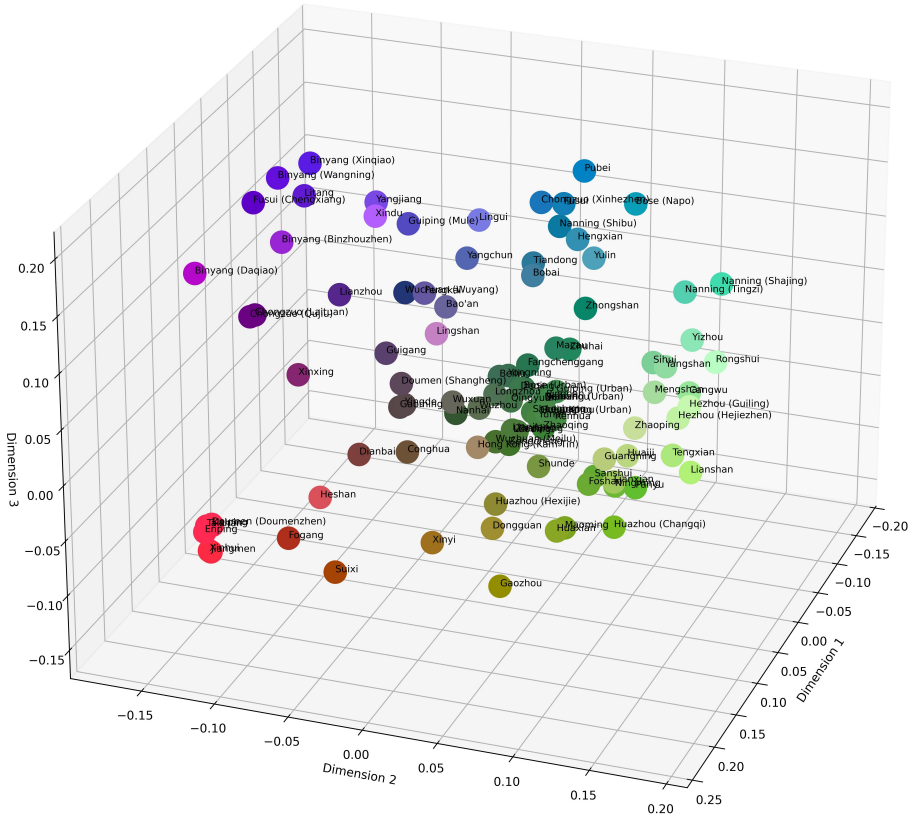


Figure 8.6: 3-dimensional MDS plot of tone distances using mOCO representation, (r^2=67.2%)

What is striking in Figure 8.4 is that tonal variation seems not forming a geographical continuum, opposed to segmental variation (see Figure 5.17 for the MDS map of segmental variation). To demarcate the

[10]Having four clusters also makes it easier to draw the dialect boundaries with the 4-cluster solution in Figure 8.7.

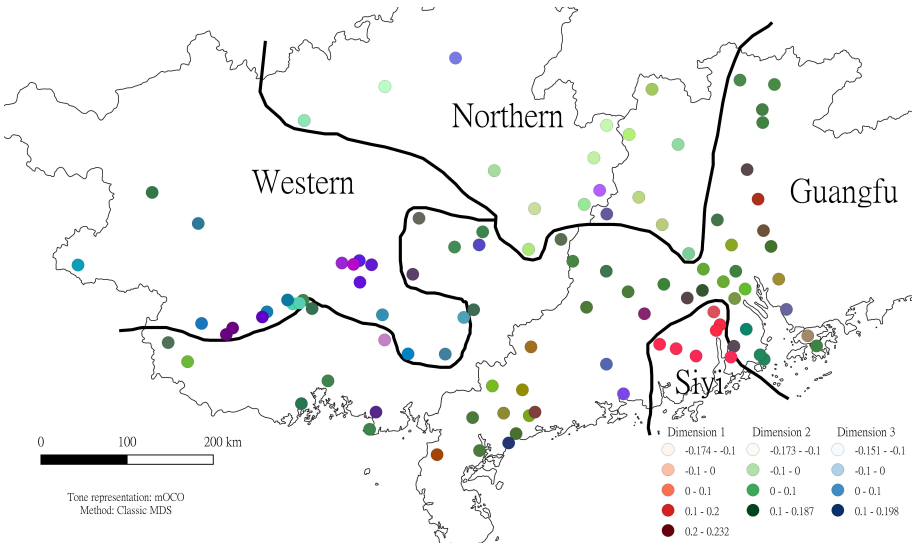


Figure 8.7: MDS map of tonal variation of Yue dialects with borders based on cluster analysis

four areas pointed out above, additional black lines were added to the MDS map (with outliers ignored), which can be found in Figure 8.7. It is not difficult to notice that there are rather sharp borders between the groups, based on the differences in the coloured circles (which represent dialect distances); red (Siyi) is sharply contrasted with the neighbouring green (Guangfu); light green (Northern) contrasts with the green circles in the south (Guangfu) and the blue-purple circles (Western) with the green. These patterns suggest that tonal variation forms dialect areas, and not a dialect continuum across the Yue-speaking area.[11]

8.4 Tonal vs. segmental variation

The comparison of different linguistic levels is not new, as it has been done in e.g. Dutch (Spruit et al. 2009), Swiss German (Scherrer and Stoeckle 2016), American English (Grieve 2013) and Norwegian (Gooskens and Heeringa 2006). These studies include comparisons between lexis, phonetics, morphology, syntax as well as prosody. However,

[11]The additional dialect boundaries do not imply that the dialects within any dialect region (with the possible exception of the Siyi dialects) are homogenous.

the relationship between the variation on the tonal and segmental levels has not been explored before.[12] It is intriguing to see how tones and segments, both of which are part of phonetics, will pattern.

To compare how much tonal variation correlates with segmental variation, the *Mantel test* (Mantel 1967) has been performed. Values between two distance matrices cannot be directly correlated. This is because values in distance matrices are often correlated, which violates "the usual assumption of independence between objects" in classical test approaches (Bonnet and Van de Peer 2002). A widely use method to account for such distance correlations is the Mantel test. It is a statistical test which uses permutation to assess the p-value of a 'normal' Pearson's or Spearman's correlation between two distance matrices. The null hypothesis is that distances in the one matrix are independent of the corresponding distances in the other matrix (Heeringa 2004).

This method was originally proposed to assess the spatial and temporal relationship of a pandemic (Mantel 1967), but it is widely used in other disciplines such as ecology (e.g. Legendre and Legendre 2012). Furthermore, the Mantel test has also been used in dialectometry, namely for investigating the relationship between geographical distance and linguistic distance (Nerbonne and Heeringa 2007; Huisman et al. 2021).

In the following analysis, the Mantel test has been applied to the tonal and segmental distance matrices using the R package *Vegan* (Oksanen et al. 2024). To perform the test, the *mantel()* function was used and the permutation was set to 9999. The Spearman's rank correlation coefficient[13] from the test is $r = 0.52$ with a p-value of 0.0001. This indicates that there is a significant positive correlation between tonal variation and segmental variation in the Yue-speaking area, but the strength of the correlation is neither weak nor strong.

To further understand how the two levels of variation are correlated with each other, a correlation analysis (Spearman's rank rho) was performed between the values from multidimensional scaling for the two distance matrices. This analysis consists of pairwise comparisons between six groups of values: values from the three dimensions for each linguistic level. Each pairwise comparison returns a Spearman's rank correlation coefficient[14], and a correlation plot for each pair, and they

[12]Pitch accent languages are not included.

[13]Spearman's rank correlation was used because linearity is not expected for the distance matrices.

[14]Linearity is not satisfied for Pearson's correlation. See Figure 8.8 for the scatter

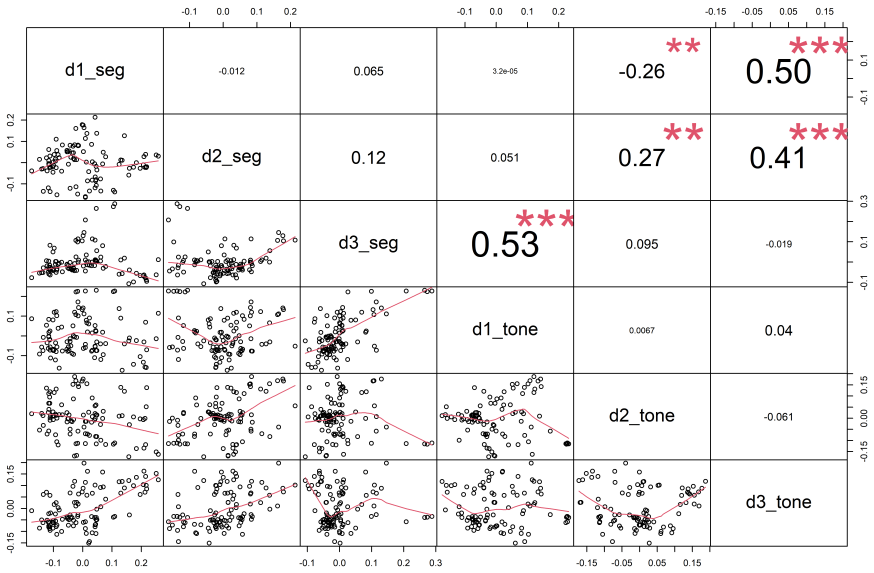


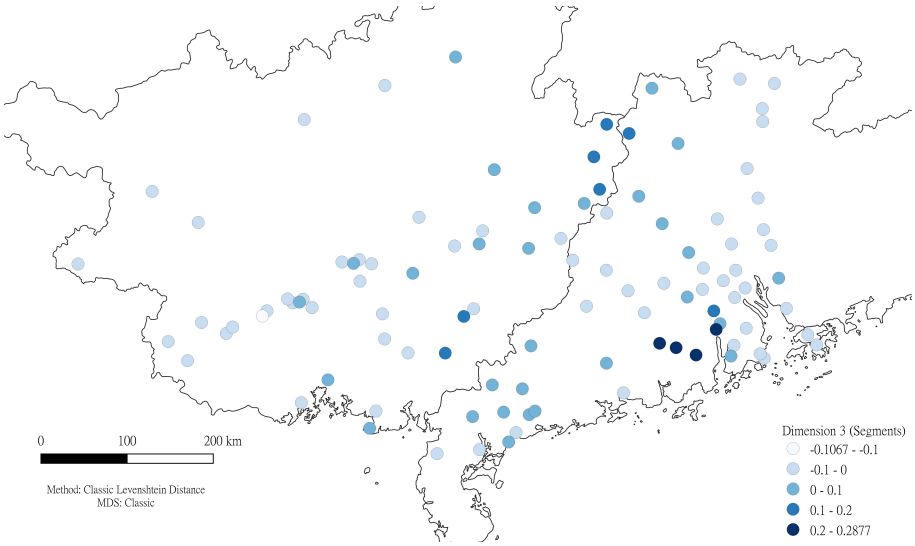
Figure 8.8: Correlation matrix between each dimension from MDS for tonal and segmental distance matrices

are presented in a correlation matrix, in Figure 8.8.

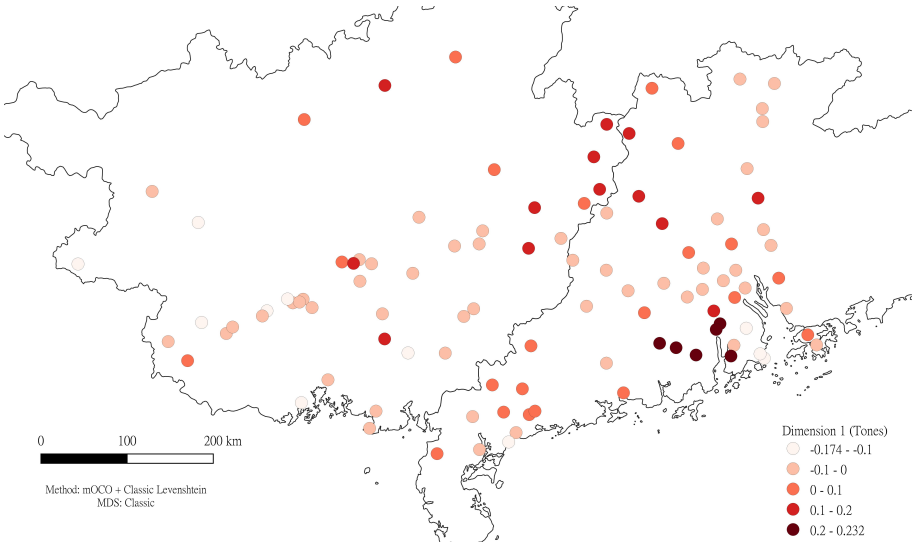
In Figure 8.8, we can see that the strongest significant correlation can be found between dimension 3 of the segmental matrix and dimension 1 of the tonal matrix, with a correlation coefficient r = 0.53. If we look at the maps in Figure 8.9, which shows the Dimension 3 values for segments (Figure 8.9a) and Dimension 1 values for tones (Figure 8.9b), we can see that these two dimensions mainly capture the Siyi dialect area. In addition, dialects to the north of the Guangdong-Guangxi border also show a higher value than the majority of the dialects in the dataset, although the number of dialects with a higher value differs. These two maps suggest that both on the tonal and segmental levels, Siyi dialects (and to a lesser extent, the He-Lian dialects) stand out from the rest of the dialects.

The second strongest correlation found in the correlation matrix lies between dimension 1 of the segments and dimension 3 of tones, with a correlation coefficient of r = 0.50. The dialects with the highest di-

plots.

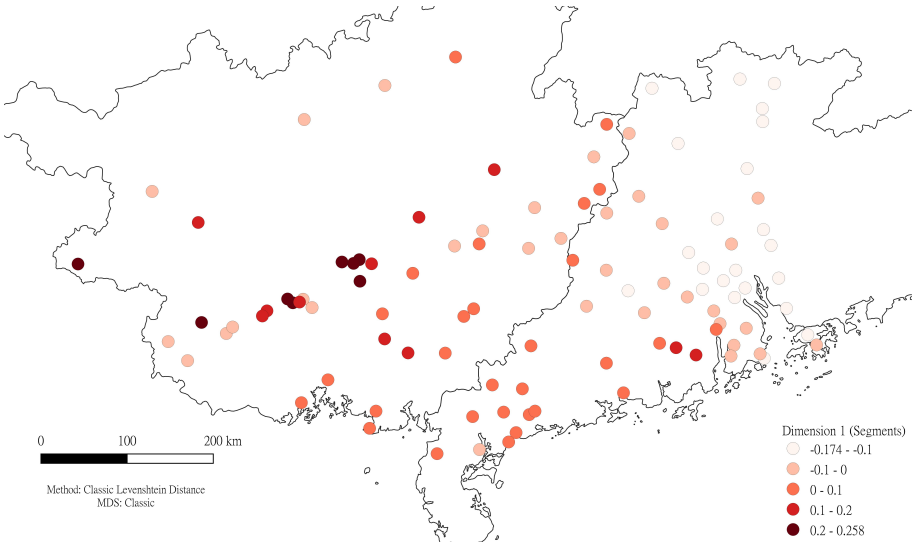


(a)

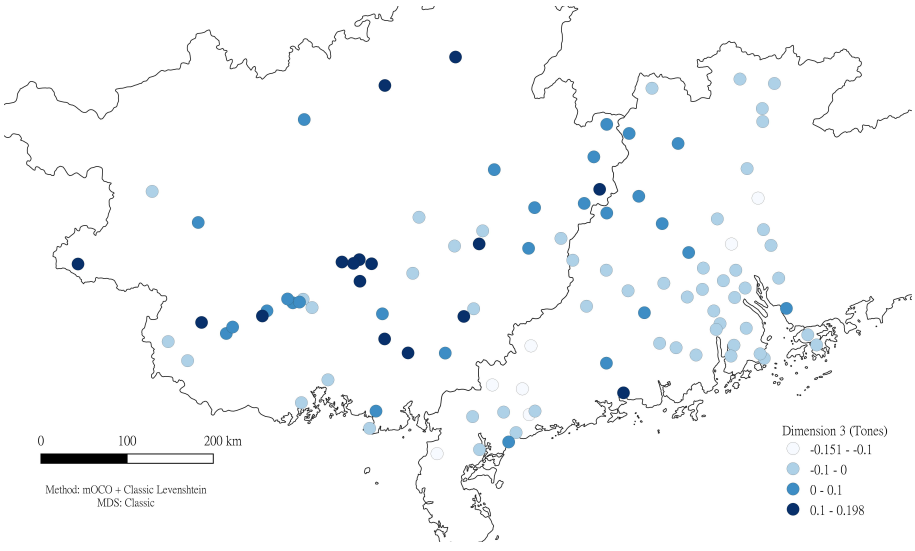


(b)

Figure 8.9: Individual dimension maps for (**a**) segments (Dimension 3, $r^2 = 34.8\%$) and (**b**) tones (Dimension 1, $r^2 = 26.0\%$)



(a)



(b)

Figure 8.10: Individual dimension maps for (**a**) segments (Dimension 1, $r^2 = 25.0\%$) and (**b**) tones (Dimension 3, $r^2 = 21.2\%$)

mension values on both levels are found in the Western dialect area (see Figure 8.10), which correspond to the dialects in the traditional Guinan Pinghua area. Based on the maps, we can also see that there are more Western dialects which differ from the rest of the Yue dialects on the tonal level than the segmental level, but Western dialects are distinctive from the rest of the dialects on both tonal and segmental levels.

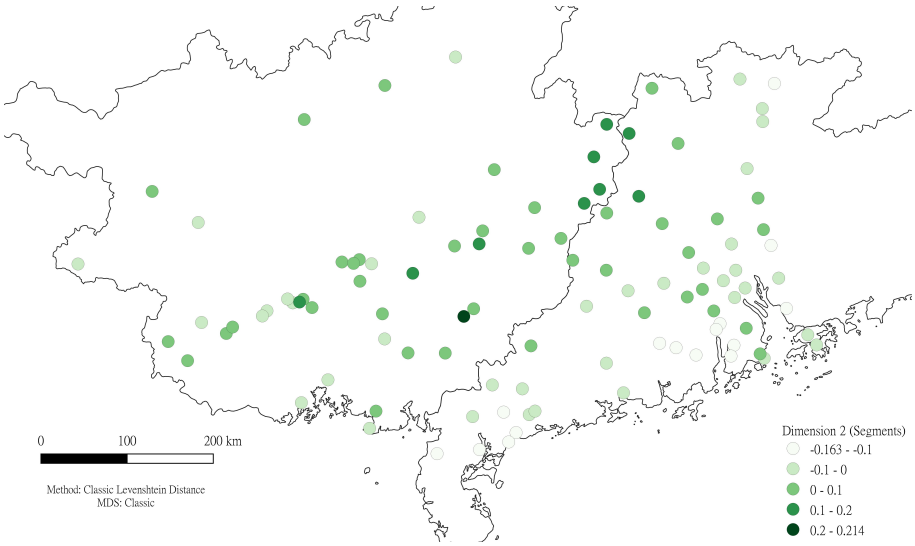
The 3rd strongest pairwise correlation goes to dimension 2 of the segmental level (Figure 8.11a) and dimension 3 of the tonal level (Figure 8.10b) and the 4th pair goes to dimension 2 of segments (Figure 8.11a) and dimension 2 of tones (Figure 8.11b). However, based on the maps shown in Figure 8.11 and Figure 8.10, it is unclear which parts of the maps contribute to these correlations.

In general, we can conclude that Siyi and Western dialect groups are distinct from the rest of the dialects on both linguistic levels. However, for the rest of the dialects, we do not have any immediate evidence that they pattern in a similar way, as shown Figure 8.11. It should also be noted that in Dimension 2 of the tonal analysis, the central northern area stands out from the rest of the dialects, which is not seen in the map for Dimension 2 of the segmental analysis. This might explain why the correlation between the two matrices is only r = 0.52 and not any higher.

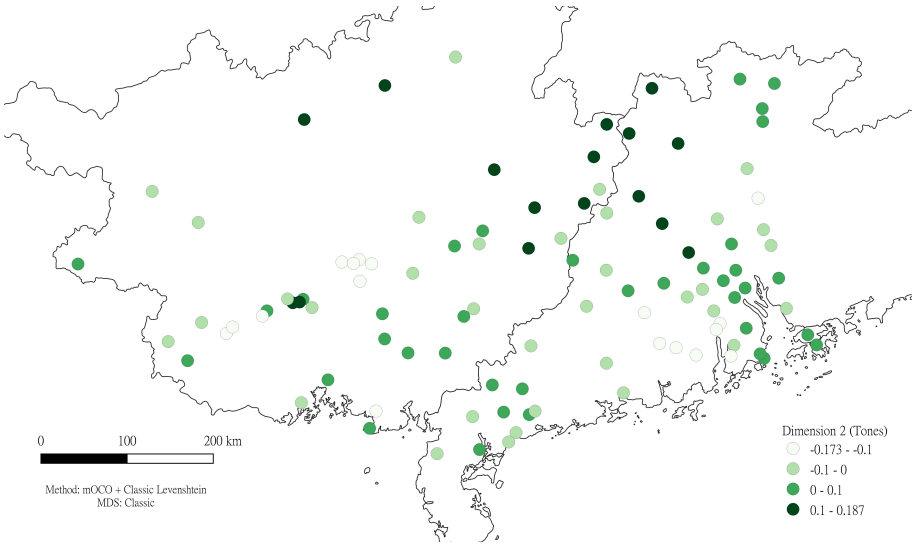
8.5 Discussion

Based on the MDS map in Figure 8.4, which reflects the tone distances, we can see that tonal variation is not homogenous. Within a certain dialect area (with the exception of Siyi dialects), there are different shades of green or blue/purple. This is a rather different picture compared to segments (see Figure 5.17 for the MDS map of segmental variation). This discovery can serve as a reminder for dialectologists that one should not always assume a continuum pattern in dialectal variation.

In Section 8.4, it has been observed that while two dialect areas, Siyi and Western, correspond to each other in terms of the tonal and segmental aggregate analyses, this is not the case for the other areas. It is interesting to see that although segments and tones are both essential parts of the pronunciation of words, they have different degrees of resemblance depending on the dialect area. Guangfu dialects, which have shown an expansion pattern (due to emigration) on the segmental level,



(a)



(b)

Figure 8.11: Individual dimension maps for (**a**) segments (Dimension 2, $r^2 = 24.0\%$) and (**b**) tones (Dimension 2, $r^2 = 27.0\%$)

appear to exhibit a mismatch with the patterns observed on the tonal level. One possible explanation could be that tones could change at a different rate from segments in terms of contact-induced changes. The motivation of this hypothesis comes from the fact that in the segmental variation, the MDS map (Figure 5.17) shows a very clear green continuum going from the east (Guangdong) to the west (Guangxi). Many dialects located in Guangxi, which are spoken near the orange (Western) dialects, are traditionally classified as Yong-Xun dialects, and they were formed around 150 years ago as a result of Guangfu speakers emigrating from the Guangfu region in the east along the Western River (de Sousa 2022:268). However, on the tonal MDS map (Figure 8.4), this emigration pattern is not very apparent. Could it be possible that the transplanted varieties were in contact with the local Yue dialects, and segments were somehow transferred quicker than tones? There might be an example of this in the formation of 'Plastic' Mandarin in Changsha (Hunan). Plastic Mandarin is a levelled variety between Standard Mandarin and the local Changsha (Xiang) dialect. It has a highly similar segmental inventory to Standard Mandarin (which makes the variety mutually intelligible to Mandarin speakers, while the local Changsha dialect has a distinct phonemic inventory which makes it not intelligible to Mandarin speakers), while the phonetic tones were mainly acquired and adapted from the Changsha dialect (Xu 2022:151-152). Could this also be the case for some of the Yue dialects in Guangxi? Alternatively, it could also be the other way round, i.e. tones from local dialects influence the transplanted dialects quicker than segments. At this point, no conclusions can be made, since this requires more research on segmental versus tonal changes.

The current chapter compares tonal variation vs. segmental variation. The next step would be to explore ways to combine the two levels into a single phonetic level. There are some things to consider: (1) Should we assign weights to the two levels? If so, how? and (2) Should we really combine them simply because they are often grouped as phonetics, even though they have very different properties, and they show very different behaviours?

For (1), Heeringa and Nerbonne (2006) have shown a way to combine distances from two linguistic levels (lexis and segments) by first converting the distances into z-scores, then take an average between the two distances. However, they have acknowledged that this method may not necessarily be optimal, since (in their case) pronunciation plays a

more important role than lexis (Gooskens and Heeringa 2006). Another potential method is to concatenate both the segmental transcription and the mOCO representation together before the calculation of Levenshtein distance. Since Yue (and many Southeast Asian tone languages) have rather short number of segments within a syllable, the implied weighting of the tone elements in the mOCO representation when combined with the segmental transcription becomes questionable. Do and Lai (2021) have found that in monosyllables, onsets and nuclei weigh more than codas and tones in phonological distance judgments. The findings in the weighting of different levels from Gooskens and Heeringa (2006) and Do and Lai (2021) point to the direction that more careful and sophisticated methods are needed when combining two linguistic levels together. However, right now, there are simply not enough studies for us to know how exactly we should combine and weight different linguistic levels. In addition, it is also unclear whether different languages may show different weighting patterns. Perhaps different (tone) languages require different ways to combine tones and segments in a combined aggregate analysis. This awaits for further research.

Regarding (2), I argue that the debate on whether we should combine tones and segments in one single analysis is somewhat parallel to arguing whether we should merge morphological variation and syntactic variation together, or more relatedly, separating consonants and vowels or not. For the latter case, often an aggregate analysis would not separate the consonants and vowels. But there are cases, for instance in Heeringa and de Wet (2008), where separating consonants and vowels can give us more detailed insights (in Heeringa and de Wet's case, it is in finding the origin of Afrikaans). Therefore, both separate and combined analyses of different linguistic levels have their own merits, depending on what the research questions are.

Lastly, a remark shall be made on the tone representation. mOCO is able to differentiate 72 out of 73 tones in the Yue dataset, and it is the representation that can differentiate the greatest number of tones converted from Chao's (1930) tone letters at the moment. However, it still requires further refinements before it can fully differentiate all of the convex (and concave) tones.

8.6 Conclusion

Through using a modification of Yang and Castro's (2008) OCO tone representation with Levenshtein distance, this chapter explored the patterns of tonal variation across over 100 dialects. Unlike previous findings, tones show a pattern which resembles dialect areas in space, rather than a continuum. Moreover, it has been found that there is a correlation between tones and segments, but the correlation is not strong. By looking more closely at the underlying dimensions between the two matrices, it has been discovered that the dimensions for the Western dialects and Siyi dialects correlate with each other, but not for the rest of the dialect areas. The mismatch in the segmental and tonal variation in the other two dialect groups awaits future investigation.