Dialect patterns and variation signs in southwest-central China- Calculational evidence based on Spatial Statistical Modeling and Topological Networks

Fang Yifei
College of Chinese Language and Culture, Nankai University, China
fangyifei@mail.nankai.edu.cn

Abstract
The southwestern-central China region is a critical linguistic contact zone shaped by historical migrations, yet its fine-scale dialectal structure remains poorly quantified. To systematically decode this geographic structure, we compiled a dataset of 510 lexical variables across 269 dialect localities. Integrating multidimensional scaling(MDS), fuzzy c-means (FCM) clustering, and topological network analysis, we reveal a pronounced bilayer spatial pattern: a macroscale dialectal continuum driven by massive Ming-Qing migrations, overlaying fragmented residual dialect islands preserved by micro-topographical isolation. Furthermore, multiple regression on distance matrices(MRM) demonstrates that terrain's independent barrier effect is negligible after correcting for spatial autocorrelation, indicating that intense population replacement effectively overrides static geographic resistance. Topological network analysis identifies Tianzhu County as a critical hub, illustrating how intense contact in transitional zones fosters stable, mixed linguistic systems like Suantang speech. These findings support a cultural diffusion model driven by demographic shifts. Ultimately, this computational approach provides fresh insights into the interwoven history of Chinese dialects and proposes a data-driven, geospatial paradigm for future dialects field-site selection.

Keywords: dialectometry, weighted Jaccard distance, fuzzy c-means clustering, multiple regression on distance matrices, topological network, cultural diffusion, dialect contact

1. Introduction

Chinese comprises a set of language varieties that constitute the Sinitic branch of the Sino-Tibetan family, spoken by approximately 16% of the world's population. Understanding the geographic structure of Chinese dialects and the evolutionary relationships among them is critical, as it bears on issues concerning the formation of the linguistic landscape of eastern Eurasia and, more generally, the mechanisms of language variation, while also contributing to the root-seeking project of the Chinese civilization. In the past, although the Biao of Ancient Chinese Phonology(*汉字古音表*)provided some regularities of dialectal evolution from a phonological perspective, our knowledge of Chinese dialects remains limited. Traditional phonology and experimental phonetic studies have tended to focus on qualitative descriptions and generalizations of features, relying heavily on scholars' empirical judgments based on specific representative points such as the Qieyun (*切韵*) phonological system (Hu, 2021). Such approaches are inadequate for precisely quantifying the genetic distances between different dialect groups, nor are they able to dynamically demonstrate the gradual clinal changes of dialectal features across continuous geographic space, leading to persistent disputes over the classification of marginal dialects, transitional dialects, and dialect islands.

The southwestern–central China region (here referring to Yunnan, Guizhou, Sichuan, Chongqing, Hubei, Hunan, Jiangxi and their adjacent border areas) has historically served as a convergence zone for multiple large-scale migration waves from north to south and from east to west, rendering its dialectal makeup considerably complex(Zhou & Lo, 1991), with both heterogeneity and isomorphism coexisting. The substrata of Southwestern Mandarin and the dialects of central China differ markedly; the *Language Atlas of China* notes that a major feature of Southwestern Mandarin is the merger of the ancient entering tone into the yangping(阳平) tone, which is fundamentally distinct from the complex retention of entering tone and voiced obstruent evolution observed in the Xiang–Gan area of central China as argued by Guo(2009) and others. Behind such differences, however, intensive contact has also promoted intermingling and mutual influence among dialects. Qiao(2025) points out that the colloquial reading of the character "耍" in Changsha dialect is [ɕya^{41}], while the literary reading is [sa^{41}]; the *Xungu Xieyin (训诂谐音,*1882) records only the literary reading [sa^{41}], whereas the *Xiangyin Jianzi(湘音检字*, 1937) documents both the colloquial [ɕya^{41}] and the literary [sa^{41}], and the Southwestern Mandarin of Changde dialect has [ɕya^{31}] today. Textual evidence from rhyme books demonstrates that this phenomenon did not arise from internal evolution within the Changsha dialect itself, but rather resulted from contact with Southwestern Mandarin as represented by the Changde dialect, constituting a change induced by language contact. Therefore, as the region where the influential Southwestern Mandarin comes into closest contact and competition with the non-Mandarin southern dialects (Xiang, Gan, Hakka), incorporating these specific dialect localities enables the selection of samples representing different historical layers of evolution, thereby revealing the complex hierarchical relationships among dialects. Furthermore, the southwestern–central region is topographically diverse, straddling the second and third steps of China's topographic ladder, representing a dramatic transition from the Qinghai–Tibet Plateau to the eastern hills and plains, and encompassing a crisscrossing network of high mountain ranges, hills, basins and plains(see Fig. 1). It is generally held that high mountain ranges segregate communities and thus promote differentiation, while gentle terrain facilitates population exchange and thus promotes convergence; dialects in geographic space often form a continuum, and spatial continuity reflects temporal continuity(Heeringa & Nerbonne, 2001) . These complex geographic factors are bound to exert significant influences on cultural boundaries and dialectal evolution. Hence, investigating the dialectal landscape and markers of variation in the southwestern–central China region is of considerable value.

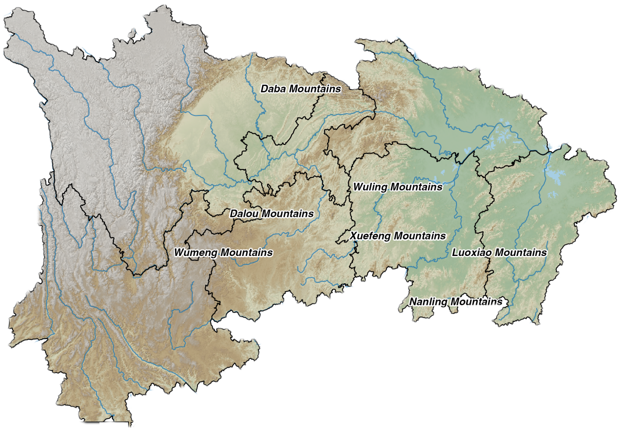


Fig. 1. Physical geography map of the Southwest-Central China region (including major mountain ranges)

Current dialectal research in the southwestern–central region predominantly focuses on isolated sub-areas and specific phonetic features, lacking comprehensive, macro-scale spatial comparisons. Consequently, the classification of central Chinese dialects remains highly contested. For instance, long-standing debates persist between proponents of "Hakka–Gan unity" (Liu, 2010)and advocates of "Hakka–Gan independence". What makes it more complicated is the evidence showing Min and Hakka share a common non-Sinitic substratum(Liu, 2004). Ultimately, these enduring controversies stem from methodological limitations: previous studies typically rely on small datasets and insufficiently justified sampling, hindering the fine-grained mapping of dialectal geographic patterns.

What model, then, can explain the historical processes of geographic dialect variation observed in the southwestern–central China region? What are the relationships, hierarchical structures, and exact degrees of divergence among these dialect groups? Furthermore, how can we accurately locate the dialect points that are most valuable for evolutionary research? Traditional Chinese dialectology investigations struggles to answer questions at such a macroscopic scale. The historical formation of Chinese dialects is generally framed by two opposing models. The horizontal transmission model, accepted by most Chinese scholars, posits that dialects evolved through continuous migration and contact with Mandarin, explaining the retention of archaic features along the southeastern coast. Conversely, the vertical transmission model assumes continuous, tree-like internal diversification over time with minimal inter-branch contact (M. Zhang et al., 2016).

To address the issues raised above and to extend our understanding of the dialects in the southwestern–central China region, this study adopts a quantitative approach to aggregate variation in Chinese dialects, employing the Weighted Jaccard Distance (WJD) and spatial metrics to account for the basic geographic structure of Chinese dialects from the perspective of dialect formation. Specifically, we draw upon unsupervised methods from machine learning, including multidimensional scaling and fuzzy C-means clustering, to identify the selected dialect areas on the basis of a relatively large and diversified feature set without any prior selection(Huang et al., 2024). To remedy the limited resolution of conventional clustering methods in handling micro-transitional zones, we further supplement our analyses with spatial modelling approaches such as multiple distance matrix regression (MRM) to assess the impact of specific geographic factors on dialect contact and to determine the diffusion patterns of dialects in the southwestern–central China region. Finally, we employ spatial topological network analysis, combined with network community detection (Louvain) and centrality measures, to locate nodes that play important roles in regional dialect contact and competition, thereby providing feasible reference points for future fieldwork on dialects.

2. Method

2.1 Materials

The linguistic data of this study were drawn from the *Linguistic Atlas of Chinese Dialects* (LACD)(Cao, 2011). The LACD data were collected between 2000 and 2008, with NORMs (non-mobile, older, rural males) as informants. It is the most authoritative dialect atlas of Chinese to date, covering 930 localities across the country and comprising three volumes with 510 maps of linguistic variables, focusing on three levels of linguistic analysis: phonology (205 maps), lexicon (203 maps) and grammar (102 maps). In these maps, variants of the same alternating feature are classified by the cartographers into different types and subtypes according to specific linguistic criteria. Differences in colour symbols on the maps indicate type differences, while differences in shape symbols indicate subtype differences. The lexical classification used in LACD is generally based on roots (morphemes) rather than on phonetic forms. Since Chinese characters are morphosyllabic, marking morphemes with characters is more convenient and economical than using complex phonetic word forms. The appendix of LACD provides information on the informants and fieldworkers for the 930 localities, as well as the status of each locality according to the classification system of Li (1987). A limitation is that the appendix presents only two hierarchical levels: group and subgroup, while varieties such as Shehua(社话), Tuhua(土话), Xianghua(乡话) and Danzhouhua(儋州话) are grouped into a single category without further subgrouping, leaving their status ambiguous.

For the present study, we extracted from LACD a dataset comprising 269 dialect localities in the southwestern–central China region (the specific localities are listed in Supplemental Table 1). The dataset consists of a data matrix with 269 rows (localities) and 510 columns (linguistic variables). The coding scheme annotates the classificatory relationships of variants for each linguistic variable at both type and subtype levels. Specifically, we converted the type indicators represented by colours in LACD (following Huang et al., 2019) into uppercase Latin letters (A–Z), and the subtype indicators represented by different symbols of the same colour into digits (1, 2, 3). For instance, variants coded as A1 and A2 belong to the same type but different subtypes, while both differ from B1 in type. Moreover, additional variants are marked with "#" as a delimiter. For example, a code such as "A1#B2" indicates the co-occurrence of two distinct types of variants. In accordance with the consistent classificatory principles of the atlas, we further reclassified these variants into three levels: type, subtype, and micro-subtype, with the micro-subtype converted to lowercase Latin letters (a–z). Distance calculations in this study are based on these codings (see Supplemental table 2).

The raw geographic data used for the cartographic visualisation in this paper are derived from the publicly available GeoJSON provided by Alibaba Cloud https://geo.datav.aliyun.com/areas_v3/bound/100000_full.json.

2.2 Linguistic Distance Calculation

Similarity serves as the foundation for the calculation of linguistic distance. Given the pronounced taxonomic hierarchical attributes of the dialect feature data employed in this study, the conventional edit distance algorithm is not directly applicable. We therefore adopted the Weighted Jaccard Distance (WJD) proposed by Huang et al. (2024) as the core metric for measuring dialectal differences. This method represents a refinement of the classical Jaccard similarity coefficient, with the advantage of effectively handling the hierarchical structure of linguistic features and multiple phonetic transcriptions. The distance calculation follows two sets of weighting logic. The first is feature-level weighting, whereby different weights are assigned according to the granularity of the classification. Specifically, Type reflects the primary distinctions among dialects (e.g., some dialects refer to maize as baogu (苞谷) or zhenzhumi (珍珠米), while others call it yumi (玉米), with a weight of W=1.0; Subtype and micro-subtype reflect secondary distinctions (e.g., within the area where baogu is used, there may be regular tonal or rhyme differences for the term, or more subtle phonetic variations), with weights set to W=0.5 and W=0.25, respectively, as shown in (1).

$$\mathrm{DV_f} = W_t \cdot \mathrm{COD_t} + W_s \cdot \mathrm{COD_s} + W_m \cdot \mathrm{COD_m} \quad (1)$$

The second mechanism is multiple-response weighting, applied when a locality records multiple transcriptions for a single concept. Weights are assigned in descending order of prevalence: the primary variant receives W_1=1, while subsequent variants receive halved weights (W_2=0.5, W_3=0.25). For example, the term for "sun" is exclusively taiyang(太阳) in the Chengyu sub-group. Transitional counties, however, record both taiyang (W_1=1.0) and ritou(日头) (W_2=0.5), whereas southern Jiangxi primarily records ritou (W_1=1.0). Accommodating multiple readings prevents the misclassification of transitional dialects and preserves their continuum characteristics, as formulated in Eq. (2).

$$\mathrm{DV_{jk}^{v}} = \frac{\left(W_1 \cdot \mathrm{DV_f^1} + W_2 \cdot \mathrm{DV_f^2} + \cdots + W_n \cdot \mathrm{DV_f^n}\right)}{\left(\sum_{i=1}^{n} W_i\right)}, i = 1,2, \ldots, n \quad (2)$$

The overall distance between localities is computed in three steps: measuring the distance between variants ($\mathrm{DV_f}$) via Eq. (1), computing the distance for a single linguistic variable, and aggregating these distances across all variables. The aggregate distance ($\mathrm{DV_{jk}}$) between localities *j* and *k* is the mean distance across all *p* variables (*p*=510), scaling from 0 (identical) to 1 (completely distinct), as shown in Eq. (3).

$$\mathrm{DV_{jk}} = \frac{\sum_{v=1}^{p} \mathrm{DV_{jk}^{v}}}{p}, \quad (p = 510) \quad (3)$$

To account for the distortion inherent in projecting 510-dimensional variation for 269 localities into a three-dimensional space, we report the stress value, where a stress below 0.2 indicates an optimal mathematical fit. The distance computation yields a 269 × 269 symmetric matrix containing unique pairwise values. From this distance matrix, the corresponding similarity value ($\mathrm{SV_{jk}}$) between localities is directly derived, as shown in Eq. (4).

$$\mathrm{SV_{jk}} = 1 - \mathrm{DV_{jk}} \quad (4)$$

For the distance calculation, we implemented the Weighted Jaccard Distance algorithm programmatically in the Pycharm 2.4 environment (primarily using NumPy and SciPy), and constructed the dialect feature distance matrix accordingly.

2.3 Multidimensional scaling and fuzzy c-means clustering

To quantify and visualize the linguistic distances, we applied multidimensional scaling (MDS) and fuzzy c-means (FCM) clustering. MDS was executed using the cmdscale function in R, retaining the first three dimensions to capture the primary geolinguistic variance. These dimensions were subsequently mapped to the red, green, and blue (RGB) channels, generating a continuous spatial visualization of the dialectal landscape. To delineate the internal structure of these dialect areas, FCM clustering was performed via the ppclust package. By calculating continuous membership coefficients, this algorithm quantitatively distinguishes stable dialect cores (high membership) from mixed transitional boundaries (distributed membership). The FCM results were projected using two distinct configurations: a discrete map (Type 1) assigning each locality to its maximum-membership cluster, and a continuous heatmap (Type 2) illustrating membership intensity. Finally, the 269 sampling points were transformed into contiguous spatial surfaces using Voronoi tessellation (ggvoronoi). This geometric partitioning serves strictly as a visualization heuristic, extending regional attributes to unsampled or non-Sinitic areas based on nearest-neighbour proximity.

2.4 Spatial statistical modelling and topological network analysis

To mitigate the micro-scale clustering distortions of topographic distances, we utilized multiple regression on distance matrices (MRM) to quantify geographic effects on linguistic variation (Goebl, 2006; Quilodrán et al., 2025) . Using the WJD matrix as the dependent variable, we constructed two independent matrices to simulate environmental resistance: a Euclidean matrix (absolute straight-line distance) and a least-cost path (LCP) matrix. The LCP matrix models actual travel routes, accounting for the rugged topography that historically constrained movement and intensified spatial isolation. Consistent with Séguy's curve, which dictates a sub-linear relationship between absolute distance and lexical variation (Huisman et al., 2021). Euclidean distances were log-transformed prior to regression, as formulated in Eq. (5).

$$WJD_{ij} = \beta_0 + \beta_1 \ln\left(D_{eucl(ij)}\right) + \beta_2 D_{topo(ij)} + \epsilon_{ij} \quad (5)$$

Subsequently, to quantify the driving forces behind dialectal differences, a standard MRM regression design was employed to assess the effects of geographic factors.

However, regional linguistic data exhibit strong spatial autocorrelation, which inflates Type I errors in traditional Mantel tests. To diagnose this, we performed principal coordinate analysis (PCoA) on the linguistic distance matrix via the cmdscale function, followed by Moran's I test using a k-nearest neighbours spatial weight matrix (k=4, i.e., for each dialect locality, the four nearest neighbouring localities were identified). After confirming highly significant spatial autocorrelation, Moran Spectral Randomization (MSR) with 9,999 permutations was conducted. This rigorous null-model correction effectively isolates the true independent effect of topography from the confounding baseline of spatial proximity, ensuring the robustness and reproducibility of geographic inferences.

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} \quad (6)$$

To enhance the interpretability of the results, to emphasise their practical implications for dialect research, and to avoid forcibly assigning data to discrete traditional clusters, we finally introduced spatially embedded topological networks. The WJD similarity matrix was transformed into a mathematically meaningful network; threshold sensitivity analyses were performed at cut-off values of 0.85, 0.90 and 0.95, with 0.90 ultimately selected as the threshold, retaining only the top 10% strongest linguistic similarities to construct the underlying undirected weighted graph, thereby effectively filtering out weak and noisy connections. The Louvain community detection algorithm was then applied to this filtered graph to objectively identify tight dialect clusters purely on the basis of modularity optimisation. Using the sf and ggraph packages in R, the resulting topological network was embedded directly into geographic space. Two complementary topological metrics were used to quantify the structural roles of specific dialect localities within the regional continuum: (1) Betweenness centrality, reflecting the capacity of a node to serve as a “bridge” or “intermediary” in the network topology. In spatial networks, nodes with high betweenness centrality tend to lie at the junctions of multiple dialect communities and serve as essential pathways for the cross-regional transmission of dialect features(Barthelemy, 2011). (2) Node degree, reflecting the local connectivity of a node. In densely connected contexts, regions with high node degree reveal core areas of strong internal homogeneity, shaped jointly by frequent local interactions and historical population stability (Curtin, 2007). If a dialect node shows prominent results across the selected network centrality metrics, it provides evidence that this locality plays a significant role in shaping the regional dialectal landscape, acting as a core node of contact and competition in dialect transition zones.

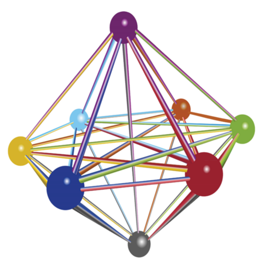


Fig. 2. In these quantitative models, data are structured using two types of variables. As illustrated, nodes represent individual objects with specific measurable attributes (point variables), while the connecting edges quantify the interactions, similarities, or distances between pairs of objects (pairwise distance variables).

By locking the colour palette to the global Louvain community assignment scheme across the two maps, this network-based approach provides a visually consistent comparison between frequently interacting transitional hubs and deeply entrenched homogeneous cores, thereby offering a dynamic complement to the static Thiessen (Voronoi) polygon maps.

3. Results

3.1 MDS and RGB Mapping

Based on the dialect distance matrix constructed from lexical and phonetic features, this study employed multidimensional scaling (MDS) to reduce the high-dimensional linguistic differences into a three-dimensional space. Figure 3 presents the MDS maps for each of the three dimensions. The computation results show that the cumulative variance explained by the first three dimensions is 34.47%, with the first dimension (Dim1) accounting for 21.47%, the second (Dim2) for 7.17%, and the third (Dim3) for 5.83%. The model's stress value (Stress-1) is 0.453, contributed mainly by the eastern region; the goodness of fit (GOF=0.345) does not reach an acceptable level, indicating that more advanced spatial statistical modelling is required to aid in explaining the variation (see Fig. 3). We do not assume that all Chinese dialects can be adequately characterised by the three global dimensions of regional variation; nevertheless, for the first dimension, which has a relatively higher explanatory power, it does reveal a certain degree of the high nonlinearity and heterogeneity of linguistic features in the southeastern dialect transitional zone.

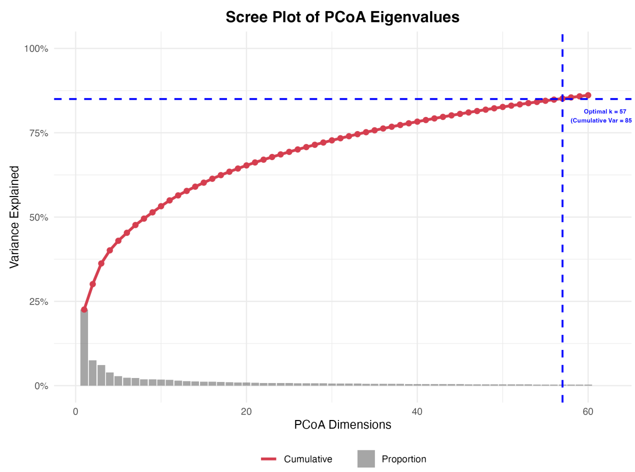


Fig. 3. MDS score maps

The first dimension reveals a dialectal pattern extending from the southwest to central China, exhibiting a pronounced binary oppositional structure, in which the green area broadly coincides with the traditional classification of Southwestern Mandarin, while the orange area encompasses the central and eastern peripheries (the Fujian–Jiangxi border). In the first dimension, a clear transitional zone marked in white separates the southwestern and central regions, forming a continuum between the two rather than a discrete area (see Fig. 4a). As the dominant variable, it contributes more than two-fifths of the explanatory power, accounting for the largest share of variance in the data, and generally corresponds to a longitudinal gradient dominated by east–west geography. The negative end of the data, represented by the green area, is densely clustered and highly convergent within an extremely narrow range (–0.19 to –0.20), as exemplified by Lincang (–0.201), Fuyuan (–0.199), Beichuan (–0.198), Jianshui (–0.196) and others, indicating the very low internal variability and high homogeneity of Southwestern Mandarin. The positive end, the orange area, exhibits significant dispersion in its distribution, with localities such as Wuyishan (0.217), Yushan (0.197), Guangfeng (0.195) and Dingnan (0.191), all located mainly in Jiangxi Province and its margins. The second and third dimensions primarily reveal the internal structure of the high-variation regions, further corroborating the dialectal geographic pattern of "homogeneous southwest, fragmented southeast" (see Fig. 4b,c). These two dimensions differentiate the subgroup relationships within the Gan dialect area. The high-value areas represented by Guangfeng (0.165) and Yushan (0.144) in northeastern Gan exhibit a pronounced north–south or areal divergence from the northern Min features represented by Wuyishan (–0.091). In contrast, the projections of southwestern cities on the maps generally converge towards zero (e.g., Fuyuan –0.0009), once again indicating the lack of structural sub-differentiation within Southwestern Mandarin.

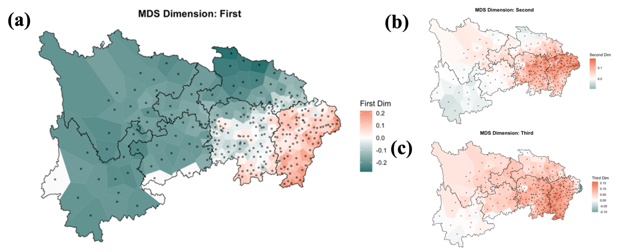


Fig. 4. MDS maps of the southwestern–central China region. Panels (a), (b) and (c) display the images of the three dimensions

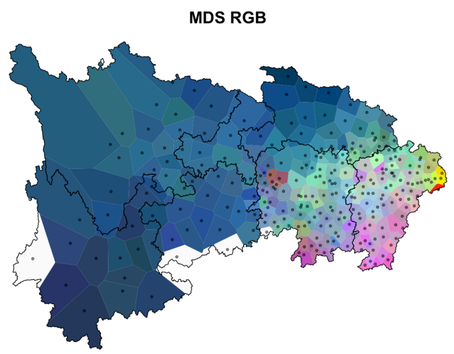


Fig. 5. RGB composite map of the southwestern–central China region

In addition to mapping each dimension individually, we present in Fig. 5 a single RGB map that combines the three dimensions. This map clearly shows that the dialectal landscape of the central China region is highly complex, consisting neither of a single dialect continuum nor of a collection of discrete dialect areas separated by sharp boundaries. Rather, it comprises relatively homogeneous, clearly discernible dialect areas of varying sizes, separated by both relatively sharp boundaries and more gradual transitional zones. Apart from the southwestern region, which forms a distinct area on the RGB map (blueish hues), the central region surrounding Hunan, including Hubei, Jiangxi and other areas, presents a less clear picture. Central China appears to be a vast transitional zone at the junction of four major dialect areas in south-central China, characterised by a mixed makeup drawing from different dialects.

3.2 FCM Maps

Based on the computed raw WJD values, Fig. 6 presents the FCM maps showing hard clusters (Type 1) and fuzzy clusters (Type 2) for cluster numbers ranging from 2 to 10. In constructing the fuzzy c-means (FCM) model, the initial number of clusters was set to cluster=10. After iterative convergence and defuzzification (i.e., assigning hard classifications according to the maximum membership principle), the results showed that the effective number of clusters was 8. Subsequent analyses were therefore conducted with this effective cluster count of 8[1].

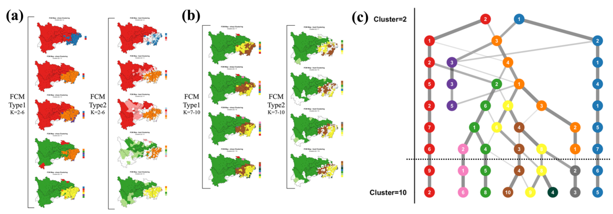


Fig. 6. FCM maps of the southwestern–central China region. (a)represents Type 1 (sharp), (b)represents Type 2 (heat). From top to bottom, the clustering results for k=2 to k=10 are shown. (c) Hierarchical variation tree of dialectal relationships in the southwestern–central China region. The dashed line indicates the optimal number of clusters (max) determined by the algorithm

The FCM Type 1 dialect partitioning demonstrates that the Southwestern Mandarin core encompassing Yunnan, Guizhou, Sichuan, and Chongqing, maintains strict internal cohesion and a 100.0% inheritance rate during early clustering stages, retaining high structural stability across all subsequent runs. This highly stable lineage sharply contrasts with the eastern and central regions, which exhibit pronounced fission-recombination. At the early macro-scale clustering stages, the Hakka and Gan dialects are deeply embedded within the same broad transitional

[1] Since FCM is an algorithm based on objective function optimisation, during the iteration process, if the natural structure of the data is insufficient to support ten significantly separated categories, some cluster centroids will shift and ultimately coincide with certain dominant centres or migrate to regions with sparse data. When no sample point has a membership degree to a given centre higher than to any other centre, that category becomes an "empty set" at the stage of hard partitioning (sharp, i.e., Type 1). For example, when cluster=10, Clusters 1 and 7 have zero sample assignments, indicating that these two categories failed to capture valid samples and were therefore not included in the subsequent dialectal hierarchical tree classification (see Fig. 5). At the same time, owing to the compactness of the feature space and the inherent adaptability of the algorithm, the redundant category settings were automatically eliminated, revealing the true dimensionality of the data.

complex. As clustering resolution increases (from K=4 to K=6), the Hakka nodes progressively separate from the Gan lineages to form an independent community with high purity. Concurrently, south-central Jiangxi and the Hunan–Jiangxi border consistently act as intense dialect contact zones, forming fusion corridors that deeply mix regional lineages. Hierarchical variation analyses (Supplemental Fig. 1) yield three primary observations regarding these spatial dynamics. First, boxplots across different hierarchical levels confirm that transitional zones persist across finer resolutions without disappearing, reflecting a continuous dialectal spectrum. Second, there is no linear correlation between transition scores and cluster splitting, suggesting that dialects undergoing continuous fission still retain stable, independently transmitted lineages. Third, the global weighted average transition score remains almost horizontal through the eighth level but drops abruptly at the ninth, indicating that it is possible that sudden separation occurs during dialectal evolution.

3.3 MRM and Mantel test

Owing to the distortion introduced by conventional MDS, the incorporation of spatial statistical inference and network topological analysis becomes essential. In the conventional regression, both the Euclidean geographic distance (β = 0.0358, p = 0.0001) and the topographic resistance distance derived from the least-cost path algorithm (β = 0.0175, p = 0.0075) showed statistical significance. However, data from regional dialect continua often exhibit strong autocorrelation, and the two spatial distance metrics are highly collinear. The Pearson correlation coefficient between Euclidean distance and topographic resistance distance is as high as 0.9187. More critically, the Moran's I test based on a k=4 nearest-neighbour weight matrix yielded a value of 0.9391 (p = 0.0001), confirming the presence of strong spatial autocorrelation in the linguistic distance matrix, implying that the significance in the conventional MRM may contain false positives. Therefore, Moran Spectral Randomization (MSR) with 9,999 permutations was applied for null-model correction. The results show that before removing the confounding baseline of spatial proximity, the raw Mantel effect size of topographic resistance was substantially overestimated (r = 0.3982). After MSR correction to isolate the autocorrelation effect of geographic location, the true independent effect size of topographic resistance dropped to r = 0.0370 and was no longer statistically significant (p = 0.4067). This finding confirms that at the current meso-scale level, the diffusion of dialects is primarily governed by wave-like contact driven by pure geographic proximity, and that the independent barrier effect of natural topography (i.e., mountain ranges) is not significant.

3.4 Topological network analysis

To transform the dense Jaccard similarity matrix into a mathematically meaningful network, we performed threshold sensitivity analyses at cut-off values of 0.85, 0.90 and 0.95. The data show that at a threshold of 0.85, 15% of the edges are retained, yielding 14 communities (modularity 0.3871); at a threshold of 0.95, excessive filtering retains only 5% of the edges, resulting in over-fragmentation into 63 communities (modularity 0.3948). Therefore, we selected 0.90 as the optimal threshold, which precisely retains the top 10% strongest linguistic similarity connections and objectively partitions the network into 26 tight Louvain dialect communities (modularity 0.3720),see Fig. 7.

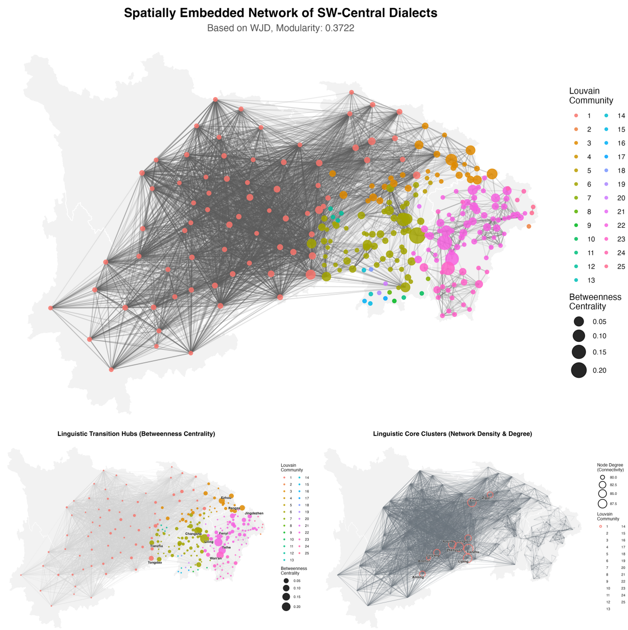


Fig. 7. Results of the topological network visualisation. The first panel shows the dialect localities divided into 25 communities, along with the edges representing the WJD distances between them. The second and third panels display, respectively, the top ten nodes with the highest betweenness centrality extracted from the first panel, and the top ten nodes with the highest degree centrality based on edge connections.

3.4.1 Dialect similarity node pairs

Among the retained strongest connections, certain dialect pairs exhibit extremely high underlying linguistic homogeneity. The top ten city pairs in terms of similarity are: Guiyang–Qinglong (similarity 0.9258), Zhenyuan–Tongren (0.9222), Yiyang–Yuanjiang (0.9060), Tianzhu–Zhenyuan (0.8853), Yujiang–Yingtan (0.8828), Zhengan–Zunyi County (0.8815), Anlong–Qinglong (0.8797), Changsha County (Futian)–Wangcheng (0.8747), Hengdong–Hengshan (0.8737) and Huayuan–Jishou (0.8737).

3.4.2 Nodes with high degree centrality

According to the node degree results, the top ten hub cities with the highest degree centrality are: Tianzhu (connections 89, total similarity 66.05), Zhenyuan (89, 65.39), Enshi (87, 62.58), Liping (87, 61.80), Yuqing (84, 61.83), Guiyang (82, 60.79), Tongren (82, 60.08), Zigui (81, 57.25) and Anshun (80, 58.67). The most densely connected nodes in the network are all concentrated in Community 1 identified by the Louvain algorithm, forming the core component of the dialectal network. Geographically, this core is basically distributed across the contiguous area of central-eastern Guizhou and southwestern Hubei, where the eastern slope of the Yunnan–Guizhou Plateau meets the hinterland of the Wuling Mountains (roughly between 105°E and 111°E, and 26°N and 31°N). In terms of physical geography, this core extends along the upper reaches of the Wu River, the Yuan River and the Three Gorges water system of the Yangtze, stretching from the Miao Ling range in the south, across the Dalou and Wuling mountains in the middle, to the southern foothills of the Daba Mountains in the north.

3.4.3 Transitional nodes

Betweenness centrality reveals the transitional zones that cross major dialect boundaries. In contrast to the highly concentrated core areas of node degree, high betweenness centrality does not depend on a high local degree, but rather on the node's spatial intermediary position in the global network; such nodes tend to lie at the forefront of convergence zones among multiple macro-clusters. The cities with the highest betweenness centrality, playing key bridging roles, include: Liling (betweenness score 0.2081), Changsha City (0.1979), Fenyi (0.1883), Ezhou (0.1375), Taihe (0.1342), Wan'an (0.1288), Pengze (0.1216), Tongdao (0.0848), Jingdezhen (0.0723) and Tianzhu (0.0708).

Among these two centrality measures, the only city that ranks in the top ten for both is Tianzhu County, Guizhou Province.

4. Discussion

4.1 Factors underlying the formation of the dialectal opposition pattern in the southwestern–central China region

Visualizations using MDS, RGB mapping, and Louvain community detection highlight a stark spatial dichotomy. The Southwestern Mandarin area functions as a vast, homogeneous hinterland, whereas the eastern region constitutes a highly varied highland. MRM analysis confirms that Euclidean distance outweighs topographic resistance in explaining these structural differences, a pattern fundamentally driven by divergent historical migrations.In the southwest, Ming and Qing state-sponsored campaigns directed massive influxes of Han Chinese into military and agricultural settlements (Zhang, 2018). To facilitate cross-regional communication, these diverse immigrant languages underwent intense koineisation, ultimately consolidating Southwestern Mandarin as the regional lingua franca.

Conversely, central and eastern China absorbed fragmented, cumulative migrations since the Tang and Song dynasties. Prolonged settlement within isolated river basins amplified internal differentiation. This historical depth established regions like Jiangxi, Hubei and Hunan as prominent mixed dialectal centers. The Gan dialect, for instance, emerged not from a single ancestor, but as a composite product of repeated contact between northern migrants and indigenous languages such as Wu and Old Xiang (Sagart, 2002). Local administrative and geographic boundaries further reinforced this complexity, fostering numerous transitional zones.

Collectively, these dynamics support a cultural diffusion model for the southwestern–central region, establishing that the scale and intensity of “socio-political migrations” (Yang et al., 2024) successfully leveled the insulating effects of natural topography.

4.2 Characteristics of dialectal evolution in central China

Moving beyond qualitative phonological comparisons, our FCM evolutionary tree (Tables 3 and 4) and Louvain community partitions demonstrate that Hakka and Gan initially embed within the same macro-scale transitional complex, corroborating their shared northern migrant substratum. As clustering granularity increases, however, the Hakka nodes separate into an independent community. This bifurcation reflects distinct historical trajectories: Hakka ancestors were isolated in the rugged mountainous junctions of Fujian, Guangdong, and Jiangxi, shielding them from subsequent Northern Chinese linguistic erosion. Conversely, Gan speakers in the middle-lower Gan River plains maintained continuous wave-like contact with surrounding Jiang-Huai Mandarin and Xiang dialects. Therefore, “Hakka-Gan unity” and “Hakka-Gan independence” are not mutually exclusive, but rather represent sequential historical phases of dialectal evolution.

Furthermore, the FCM analysis yields three principal conclusions regarding the transitional nature of central China’s dialects. First, transitional zones persist regardless of clustering granularity, indicating that regional dialect contact forms broad, continuous belts of lexical and phonetic convergence rather than discrete boundaries. Second, dialects that undergo continuous fission nevertheless retain certain components that are stably transmitted. The theory of orderly heterogeneity, advanced by William Labov and others, demonstrates that a mixed state can itself constitute a stable dialectal structure(Guy & Hinskens, 2016). Also, contact-zone dialects undergo long-term internal adjustments to form self-consistent, structurally robust new systems (e.g., the Suantang speech of Tianzhu County). Third, linguistic variation is not exclusively gradual; it also encompasses abrupt shifts. Concepts such as “competitive restructuring” (Li, 1997) and “systematic re-differentiation” confirm that dialects periodically undergo rapid structural reorganization under contact pressure.

4.3 Tianzhu County as a key marker of significant variation

In contrast to traditional expert-dependent site selection, topological network analysis precisely identifies dialect contact foci. In our spatial network, Tianzhu County (Guizhou Province) ranks among the highest in both degree and betweenness centrality, marking it as a critical hub for investigating dialect contact and variation in the southwestern–central region.

Geographically situated at the tri-provincial junction of Guizhou, Hunan, and Guangxi, Tianzhu serves as a mountainous corridor bordered by the Xiang dialect area to the east and the Southwestern Mandarin area to the west and south. While Southwestern Mandarin (locally known as Kehua) functions as the regional lingua franca, the county's linguistic ecology is highly diverse, encompassing indigenous Suantang speech alongside minority languages such as Dong and Miao.

The taxonomic classification of Suantang speech exemplifies this complex contact dynamic. Although fundamentally distinct from Southwestern Mandarin, Suantang has undergone significant structural convergence with it, evidenced by shifting aspiration patterns in ancient fully voiced initials, matching high-register falling tones, and identical entering tone mergers (Liu, 2014). Consequently, its classification remains highly contested: some assign it to the Chang–Yi subgroup of Xiang based on tonal similarities, others link it genealogically to

Sichuan's Jingzhou accent, while some propose an independent "Southwestern Xiang–Southeastern Guizhou" classification. Furthermore, its phonology, lexicon, and grammar exhibit profound substratum influences from Miao and Dong. Ultimately, despite being shaped by these four converging linguistic forces, the Suantang system has not collapsed. Instead, through long-term internal readjustment, it has stabilized into a self-consistent, exclusive dialect capable of independent evolution(Chen, 2005). This micro-level linguistic complexity empirically corroborates the macro-scale dialect contact phenomena identified in our computational modelling.

4.4 Limitations and prospects

While this study advances the quantitative analysis of dialectal geography in the southwestern–central region, two primary limitations remain. First, our computational model exhibits reduced cluster coherence at finer resolutions, indicating that feature dimensionality reduction and unsupervised clustering alone cannot losslessly resolve complex micro-level dialect structures. Second, the current spatial modeling framework excludes critical socio-geographic variables, such as hydrography, transportation networks, historical population dynamics, and ethnic settlement patterns. Future research must integrate these dimensions into comprehensive spatial regression models to quantify their independent contributions.

Despite these limitations, this study provides two distinct methodological contributions. First, we propose a geospatial statistical paradigm for fieldwork site selection. By utilizing topological network computations to pinpoint dialect contact foci, this approach replaces empirical, point-to-point selection, significantly optimizing survey efficiency and systematically identifying high-value localities. Second, we empirically verify the geographic mechanisms driving dialect contact. By applying multiple regression on distance matrices (MRM) to quantify topographic resistance, our method moves beyond qualitative visual map interpretation, providing robust statistical evidence for the cultural diffusion model in this region.

5. Conclusion

In this study, we conducted a large-scale dialectometric and spatial topological analysis of dialect survey data from the southwestern–central China region, revealing that its dialectal landscape exhibits a pronounced bilayer spatial structure. Moving beyond traditional models of unidirectional migration, our multidimensional computational approach demonstrates that a macroscale gradual continuum is shaped by successive systemic migrations, while fragmented residual dialect islands are preserved by complex micro-scale topography. Furthermore, multiple regression on distance matrices (MRM) confirms that the intensity of historical population replacement was sufficient to level the static resistance of natural geography, indicating a topological imprint woven jointly by horizontal transmission and independent evolution. By systematically integrating spatial data science methodologies, such as multidimensional scaling, fuzzy c-means clustering, and Louvain community detection, this study not only provides objective geospatial resolution to long-standing historical controversies, such as the unity versus independence of the Hakka and Gan dialects, but also reshapes the traditional research paradigm of dialectology. Ultimately, by utilizing topological indices to pinpoint key variation hubs like Tianzhu County, we successfully couple macro-scale spatial dialectometric exploration with micro-scale experimental phonetic evidence, offering a robust, interdisciplinary framework that is highly extendable to the study of regional dialects worldwide.

6. Bibliographical References

Bao, H., & Yan, S. (1986). Hunan dialect regions. *Dialect*, (4), 273–276.

Barthelemy, M. (2011). Spatial networks. *Physics Reports*, *499*(1–3), 1–101. https://doi.org/10.1016/j.physrep.2010.11.002

Cao, Z. (2011). Geographical distribution types of Chinese dialects. *Language Teaching and Research*, (5), 11–19.

Chen B. (2005). Two modes of language contact leading to the differentiation of Chinese dialects. *Journal of Peking University (Philosophy and Social Sciences Edition)*, (2), 43–50.

Chen, H., & Bao, H. (2007). The Chinese dialects of Hunan Province (draft). *Dialect*, (3), 250–259.

Curtin, K. M. (2007). Network analysis in geographic information science: Review, assessment, and projections. *Cartography and Geographic Information Science*, *34*(2), 103–111. https://doi.org/10.1559/152304007781002163

Fang, H. (2021). Experimental phonetics methods for dialect phonetics research. *Oriental Linguistics*, (1), 123–148.

Goebl, H. (2006). Recent advances in salzburg dialectometry. *Literary and Linguistic Computing*, *21*(4), 411–435. https://doi.org/10.1093/llc/fql042

Guo, L. (2009). *Phonological Study of Southwestern Mandarin in Hubei* [Doctoral dissertation, Fudan University]. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CDFD&dbname=CDFD0911&filename=2009208871.nh

Guy, G. R., & Hinskens, F. (2016). Linguistic coherence: Systems, repertoires and speech communities. *Lingua*, *172–173*, 1–9. https://doi.org/10.1016/j.lingua.2016.01.001

Heeringa, W., & Nerbonne, J. (2001). Dialect areas and dialect continua. *Language Variation and Change*, *13*(3), 375–400. https://doi.org/10.1017/S0954394501133041

Huang, H., Grieve, J., Jiao, L., & Cai, Z. (2024). Geographic structure of chinese dialects: A computational dialectometric approach. *Linguistics*, *62*(4), 937–976. https://doi.org/10.1515/ling-2021-0138

Huisman, J. L. A., Franco, K., & Van Hout, R. (2021). Linking Linguistic and Geographic Distance in Four Semantic Domains: Computational Geo-Analyses of Internal and External Factors in a Dialect Continuum. *Frontiers in Artificial Intelligence*, *4*, 668035. https://doi.org/10.3389/frai.2021.668035

Li, D. (2007). Hunan Gan dialect segmentation. *Dialect*, (3), 260–264.

Li, L. (1997). Sixty years of investigation and research on Southwestern Mandarin. *Dialect*, (4), 249–257.

Liu, Z. (2004). A Century of Equal Discussion on the Division of Chinese Dialects. *Academic Research*, (4), 125–130.

Liu, Z. (2010). The historical layers of the pronunciation of fricatives in the Xi initial in various dialects of Hakka, Jiangxi, Guangdong and Pingzhu. *Nankai Linguistics Journal*, (1), 31–36, 185–186.

Liu, Z. (2014). *Suantang Dialect Research* [Doctoral dissertation]. Hunan Normal University.

Qiao, Q. (2025). The role of historical documents on Chinese dialects in the study of the history of Chinese dialects. *Journal of Chinese Linguistics*, (3), 88–96.

Quilodrán, C. S., Currat, M., & Montoya-Burgos, J. I. (2025). Benchmarking the Mantel test and derived methods for testing association between distance matrices. *Molecular Ecology Resources*, *25*(2), e13898. https://doi.org/10.1111/1755-0998.13898

Sagart, L. (2002). Gan, hakka and the formation of Chinese dialects. *Dialect Variations in Chinese*, 129–154.

Xie, L. (2006). Regional Divisions of Gan Dialect (Draft). *Dialect*, (3), 264–271.

Xie, L. (2008). Chinese dialects in Jiangxi Province. *Dialect*, (2), 117–122.

Yang, C., Zhang, X., Yan, S., Yang, S., Wu, B., You, F., Cui, Y., Xie, N., Wang, Z., Jin, L., Xu, S., & Zhang, M. (2024). Large-scale lexical and genetic alignment supports a hybrid model of han Chinese demic and cultural diffusions. *Nature Human Behaviour*, *8*(6), 1163–1176. https://doi.org/10.1038/s41562-024-01886-9

Zhang, M., Jian, J., & Wuyun, P. (2016). A quantitative analysis of the transmission patterns of Minnan dialect. *Language Science*, *15*(5), 533–542.

Zhang, X. (2018). *A Study on the Phonetics of the Dong Dialect in Guizhou, Hunan, Guangxi and Yunnan* [Doctoral dissertation]. Shanghai Normal University.

Zhou, Z., & Lo, K. (1991). Migrations in Chinese history and their legacy on chinese dialects. *Journal of Chinese Linguistics Monograph Series*, (3), 29–49.

7. Language Resource References

Cao, Z. (2008). *Linguistics atlas of Chinese dialects* (Vols. 1–3). Beijing: The Commercial Press.

Supplemental (not included in the main text pages):

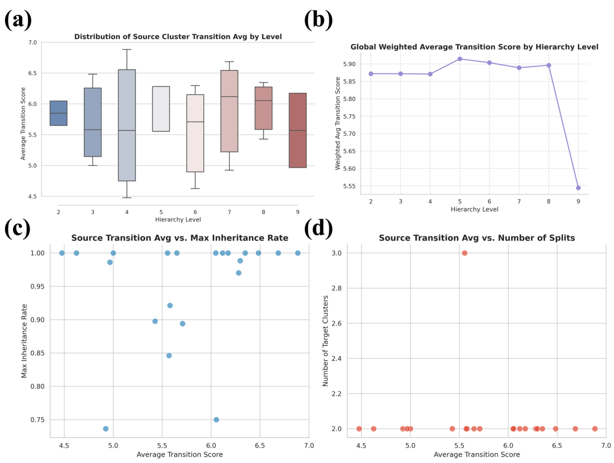
PROVINCE	MAJOR DIALECT CATEGORY	REPRESENTATIVE SUBGROUP / SUB-SUBGROUP	DIALECT LOCALITIES INCLUDED（*IDS*=269）
YUNNAN 云南	Southwestern Mandarin	Dianzhong subgroup 滇中片	Kunming (昆明), Chuxiong (楚雄), Dali (大理), Jianshui (建水), Simao (思茅), Wenshan (文山), Huaning (华宁), Malong (马龙)
	Southwestern Mandarin	Dianxi subgroup 滇西片	Lincang (临沧), Yongsheng (永胜), Baoshan (保山)
	Southwestern Mandarin	Diannan/Dianbei subgroup 滇南片/滇北片	Zhaotong (昭通), Yanjin (盐津), Huize (会泽), Fuyuan (富源)
GUIZHOU 贵州	Southwestern Mandarin	Kungui/Qianbei subgroup 昆贵片/黔北片	Anshun (安顺), Dafang (大方), Weining (威宁), Guiyang (贵阳), Duyun (都匀), Qinglong (晴隆), Dejiang (德江), Tongren (铜仁), Xishui (习水), Yuqing (余庆), Zhengan (正安), Zunyi County (遵义县)
	Southwestern Mandarin	Cenjiang subgroup 岑江片	Zhenyuan (镇远), Liping (黎平)
	Southwestern Mandarin	Qian (West) South subgroup/ Mixed 黔（西）南片/混合区	Libo (荔波), Anlong (安龙)
SICHUAN 四川	Southwestern Mandarin	Chengyu subgroup 成渝片	Chengdu (成都), Pingchang (平昌), Huaying (华蓥), Qingchuan (青川), Wangcang (旺苍), Xichang (西昌), Gulin (古蔺), Beichuan (北川), Yanting (盐亭), Zizhong (资中), Suining (遂宁), Changning (长宁)
	Southwestern Mandarin	Guanchi subgroup (including entering tone) 灌赤片（含入声）	Luding (泸定), Leshan (乐山), Ya'an (Baoxing / Hanyuan) (雅安(宝兴/汉源)), Pingshan (屏山), Fushun (富顺), Miyi (米易)
CHONGQING 重庆	Southwestern Mandarin	Chengyu subgroup 成渝片	Chongqing (重庆), Dazu (大足), Wulong (武隆), Yunyang (云阳), Zhongxian (忠县), Qijiang (綦江)
	Southwestern Mandarin	Qianbei subgroup 黔北片	Xiushan (秀山)
HUBEI 湖北	Southwestern Mandarin	Wutian/Chengyu subgroup 武天片/成渝片	Enshi (恩施), Hefeng (鹤峰), Zhongxiang (钟祥), Yidu (宜都), Yuan'an (远安), Zigui (秭归), Qianjiang (潜江), Wuhan (武汉), Yingcheng (应城), Laohekou (老河口)
	Southwestern Mandarin	Ebei subgroup 鄂北片	Zaoyang (枣阳), Yunxian (郧县), Fangxian (房县), Guangshui (广水)
	Gan	Datong/Huaiyue subgroup 大通片/怀岳片	Ezhou (鄂州), Hong'an (红安), Huangmei (黄梅), Qichun (蕲春), Wuxue (武穴), Yingshan (英山), Daye (大冶), Huangshi (黄石), Yangxin (阳新)
	Jianghuai Mandarin	Huangxiao subgroup 黄孝片	Honghu (洪湖), Jianli (监利), Shishou (partly belongs to Southwestern Mandarin) (石首)
	Gan/Mandarin	Mixed area 混合区	Chibi (赤壁), Chongyang (崇阳), Jiayu (嘉鱼), Tongcheng (通城), Tongshan (通山), Xianning (咸宁)
HUNAN 湖南	Xiang	Chang-Yi subgroup (New Xiang) 长益片（新湘语）	Changsha County (长沙县), Liuyang (浏阳), Ningxiang (宁乡), Wangcheng (望城), Changsha City (长沙市), Xiangtan County (湘潭县), Xiangxiang (湘乡), Yiyang (益阳), Yuanjiang (沅江), Nanxian (南县), Taojiang (桃江), Anhua (安化), Tianzhu (天柱)
	Xiang	Lou-Shao subgroup (Old Xiang) 娄邵片（老湘语）	Lengshuijiang (冷水江), Lianyuan (涟源), Loudi (娄底), Shuangfeng (双峰), Xinhua (新化), Shaoyang City (邵阳市), Dongkou (洞口), Longhui (隆回), Shaodong (邵东), Shaoyang County (邵阳县), Wugang (武冈), Xinning (新宁), Xinshao (新邵)
	Xiang	Chen-Xu subgroup 辰溆片	Chenxi (辰溪), Xupu (溆浦), Yuanling (沅陵)
	Southwestern Mandarin	Changhe/Jiyong subgroup 常鹤片/吉永片	Anxiang (安乡), Changde (常德), Hanshou (汉寿), Linli (临澧), Taoyuan (桃源), Yongzhou (永州), Qiyang (祁阳), Dayong (Zhangjiajie) (大庸/张家界)
	Southwestern Mandarin	Qianbei/Xiangxi subgroup 黔北片/湘西	Baojing (保靖), Fenghuang (凤凰), Guzhang (古丈), Huayuan (花垣), Jishou (吉首), Longshan (龙山), Yongshun (永顺), Luxi (泸溪), Hongjiang (洪江), Zhijiang (芷江), Zhongfang (中方)
	Gan	Yiliu subgroup 宜浏片	Liling (醴陵), Youxian (攸县), Chaling (茶陵), Yanling (炎陵)
	Gan	Leizi subgroup 耒资片	Hengyang County (衡阳县), Hengnan (衡南), Hengshan (衡山), Hengdong (衡东), Changning (常宁), Leiyang (耒阳), Qidong (祁东), Anren (安仁)
	Hakka/Others	Local dialects/Mixed 土话/混合	Chenzhou (郴州), Guidong (桂东), Guiyang (桂阳), Jiahe (嘉禾), Linwu (临武), Rucheng (汝城), Yizhang (宜章), Yongxing (永兴), Zixing (资兴), Daoxian (道县), Dong'an (东安), Jianghua (江华), Jiangyong (江永), Lanshan (蓝山), Xintian (新田), Ningyuan (宁远), Huitong (会同), Jingzhou (靖州), Xinhuang (新晃), Mayang (麻阳)
	Gan/Xiang	Mixed area 混合区	Yueyang (city / county) (岳阳(市/县)), Huarong (华容), Linxiang (临湘), Miluo (汨罗), Pingjiang (平江), Xiangyin (湘阴), Zhuzhou (株洲), Tongdao (通道)
JIANGXI 江西	Gan	Chang-Jing subgroup 昌靖片	Nanchang (city / county) (南昌(市/县)), Xinjian (新建), Anyi (安义), Jinxian (进贤), Fengxin (奉新), Gao'an (高安), Jing'an (靖安)
	Gan	Yiliu subgroup 宜浏片	Yichun (宜春), Pingxiang (萍乡), Yifeng (宜丰), Shanggao (上高), Wanzai (万载), Tonggu (铜鼓), Lianhua (莲花), Shangli (上栗), Luxi (芦溪)
	Gan	Jifu subgroup 吉抚片	Ji'an County (吉安县), Jishui (吉水), Xiajiang (峡江), Xingan (新干), Taihe (泰和), Wan'an (万安), Anfu (安福), Yongfeng (永丰), Yongxin (永新), Jinggangshan (井冈山), Suichuan (遂川), Fuzhou (抚州), Yihuang (宜黄), Le'an (乐安), Chongren (崇仁)
	Gan	Fuguang subgroup 抚广片	Dongxiang (东乡), Jinxi (金溪), Nancheng (南城), Nanfeng (南丰), Lichuan (黎川), Guangchang (广昌), Zixi (资溪)
	Gan	Yingyi subgroup 鹰弋片	Yingtan (鹰潭), Guixi (贵溪), Yujiang (余江), Yugan (余干), Wannian (万年), Poyang (鄱阳), Pengze (彭泽), Leping (乐平)
	Gan	Huaiyue/Mandarin 怀岳片/官话	De'an (德安), Duchang (都昌), Hukou (湖口), Jiujiang (九江), Ruichang (瑞昌), Xingzi (星子), Wuning (武宁), Xiushui (修水), Yongxiu (永修)
	Hui	Wuyuan subgroup	Wuyuan (婺源), Dexing (德兴)

		婺源片	
	Wu/Gan	Shangli/Mixed 上丽片/混合	Guangfeng (广丰), Shangrao (上饶), Yushan (玉山), Yanshan (铅山), Hengfeng (横峰), Yiyang (弋阳)
	Hakka	Ninglong/Yuxin subgroup 宁龙片/于信片	Ganxian (赣县), Nankang (南康), Dayu (大余), Shangyou (上犹), Chongyi (崇义), Xinfeng (信丰), Quannan (全南), Dingnan (定南), Longnan (龙南), Anyuan (安远), Xunwu (寻乌), Yudu (于都), Xingguo (兴国), Ningdu (宁都), Ruijin (瑞金), Shicheng (石城), Huichang (会昌)
	Mandarin/Others	Mixed area 混合区	Jingdezhen (景德镇), Fuliang (浮梁), Xinyu (新 余), Fenyi (分 宜), Fengcheng (丰城), Zhangshu (樟树)

Supplemental **Table** 1: 269 Dialect Points in Southwest-Central China (Dialect zoning reference: Bao Houxing & Yan Sen, 1986; Chen Hui & Bao Houxing, 2007; Li Dongxiang, 2007; Xie Liuwen, 2006, 2008)

cities	id	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
重庆	52	A	A	A1	G	A	B	A	D6	B1#A1	D1	A1	D	A	B	A1
大足	53	A	A	A1	G	A	B	A	A1	A1	A1	A1	D	A	B	A1
武隆	54	A	A	A1	G	A	B	B	D2	A1	A1	A1	D	A	B	A1
秀山	55	A	A	A1	G	A	B	A	C2	A1	A1	A1	D	A	B	A1
云阳	56	A	A	A1	G	A	B	B#A	C2	B1	A1	A1	D	A	B	A1
忠县	57	A	A	A1	G	A	B	A	C2	A1	A1	A1	D	A	B	A1
綦江	58	A	A	A1	G	A	B	C1	C2	A1	D1	A1	D	A	B	A1
安顺	301	A	A	E12#A1	G	A	B	A	A1	A1	A4	E1	D	B	B	A1
大方	302	A	A	A1	G	A	B	A	C3	A1	A4	E1	D	B	B	A1
威宁	303	A	A	A1	G	A	B	A	C2	A1	A4	A1	D	A	B	A1
贵阳	304	A	A	A1	G	A	B	A	D2	A1	A4	E1	D	A	B	A1
黎平	305	A	A	A1	G	A	B	A	A1	A1	A4	A1	D	A	B	A1
天柱	306	A	A	A1	G	A	B	A	C2	A1	A1	A1	D	B	B	A1
镇远	307	A	A	A1	G	A	B	A	C2	A1	A1	A1	D	B	B	A1

Supplemental **Table** 2 Partial Data Structure for 269 Dialect Points in Southwest-Central China - Top 15*15 (rows/columns)



Supplemental **Fig**. 1. Analysis of FCM Results: (a) Distribution of transition scores at different levels (Box plot), (b) Transition scores vs. number of splits (Scatter Plot), (c) Transition scores vs. stability/inheritance rate (Scatter Plot), (d) Global weighted average transition score trends (Line Plot)

Evolutionary stage	**Parent**	**Child**	**Main lineage inheritance rate**	**Representative cities of the parent node region**
K=2->K=3	K2_2	K3_1	1.00	Southwestern Mandarin (Chengyu subgroup etc.): Lincang, Linli, Leshan, Xishui, Yunyang, etc. (100 localities)
K=5->K=6	K5_2	K6_6	0.97	Southwestern Mandarin (Chengyu subgroup etc.): Jiayu, Dafang, Dali, Dazu, Tianzhu, etc. (101 localities)
K=6->K=7	K6_6	K7_1	0.98	Southwestern Mandarin (Chengyu subgroup etc.): Huaying, Gulin, Dafang, Dali, Dazu, etc. (100 localities)

Supplemental **Table** 3 Stable lineages of core dialects

Splitting stage	**Parent**	**Cities included in the parent node**	**Main lineage**	**Inheritance rate A**	**Split lineage**	**Inheritance rate B**
K=2->K=3	K2_1	Gan–Hakka–Xiang transitional dialects (Ninglong subgroup etc.): Wan'an, Wannian, etc. (117 localities)	K3_2	1.00	K3_3	0.65
K=2->K=3	K2_2	Southwestern Mandarin (Chengyu subgroup, Hunan Changhe subgroup, etc.): Zhongfang, Lincang, etc. (149 localities)	K3_1	1.00	K3_3	0.34
K=3->K=4	K3_3	Gan–Hakka–Xiang transitional dialects (Ninglong subgroup etc.): Wan'an, Wannian, etc. (167 localities)	K4_4	0.92	K4_3	0.71
K=4->K=5	K4_2	Southwestern Mandarin (Chengyu subgroup, Hunan Changhe subgroup, etc.): Lincang, Linli, etc. (78 localities)	K5_5	1.00	K5_2	0.57
K=4->K=5	K4_4	Xiang–Lou-Shao subgroup: Wan'an, Wannian, etc. (165 localities)	K5_1	0.84	K5_2	0.43
K=5->K=6	K5_1	Gan–Hakka–Xiang transitional dialects (Ninglong subgroup etc.): Wan'an, Wannian, etc. (130 localities)	K6_3	1.00	K6_4	0.80
K=6->K=7	K6_4	Gan–Hakka–Xiang transitional dialects (Ninglong subgroup etc.): Wan'an, Wannian, etc. (151 localities)	K7_4	0.89	K7_6	0.82
K=7->K=8	K7_1	Xiang–Lou-Shao subgroup: Lincang, Linli, etc. (86 localities)	K8_2	1.00	K8_4	0.89

Supplemental **Table** 4 Key dialect splitting events