*Journal of Linguistic Geography* (2024), **12**, 13–23
doi:10.1017/jlg.2024.7

CAMBRIDGE UNIVERSITY PRESS

Article

# Dialectometry-based classification of the Central–Southern Italian dialects

Antonio Sciarretta
Independent scholar

**Abstract**

This paper provides a new classification of Central–Southern Italian dialects using dialectometric methods. All varieties considered are analyzed and cast in a data set where homogeneous areas are evaluated according to a selected list of phonetic features. Using numerical evaluation of these features and the Manhattan distance, a linguistic distance rule is defined. On this basis, the classification problem is formulated as a clustering problem, and a k-means algorithm is used. Additionally, an ad-hoc rule is set to identify transitional areas, and silhouette analysis is used to select the most appropriate number of clusters. While meaningful results are obtained for each number of clusters, a nine-group classification emerges as the most appropriate. As the results suggest, this classification is less subjective, more precise, and more comprehensive than traditional ones based on selected isoglosses.

**Keywords:** Dialectometry; clustering; Central-Southern italian dialects; classification

(Received 28 February 2024; accepted 28 February 2024; First Published online 29 April 2024)

## 1. Introduction

The standard classification of peninsular Italian dialects is that proposed by G. B. Pellegrini (1977). Within the Italo-Romance branch of the Romance languages, the dialectal areas (systems) identified are: i) Tuscan, ii) Central (*Mediano*), iii) Intermediate Southern, and iv) Extreme Southern. Area i) largely corresponds to Tuscany. Area ii) comprises four subareas (Central Marchigiano in Central Marche, Umbrian, Latian in Central-Northern Latium, and Cicolano-Sabino-Aquilano between Latium and the Abruzzi). Area iii) is further subdivided into five subareas (Southern Marchigiano-Abruzzese, Molisano, Apulian, Southern Latian-Campanian, and Lucanian-Northern Calabrian). Area iv) is comprised of three subareas (Salentino, Central–Southern Calabrese, and Sicilian). These subareas, largely inspired by the administrative regions (*Regioni*) of Italy, are further subdivided into sub-subareas Ia, Ib, etc., often corresponding to a provincial (*Provincia*) level.

In SIL International's Ethnologue database (Eberhard, Simons, & Fennig, 2022), upon which ISO 693-3 is based, Italian (ita) is based on Pellegrini's Tuscan and Central, Napoletano-Calabrese (nap) is based on Pellegrini's Intermediate Southern, and Sicilian (scn) on Pellegrini's Extreme Southern. UNESCO's endangered languages list (Moseley, 2010) and the Glottolog database (Hammarström et al., 2022) adopt a virtually identical classification, albeit with slight differences in naming, even including some of Pellegrini's subareas.

While Pellegrini's primary classification is largely based on phonetic and morphological isoglosses (up to 33 for the whole of Italy), the subarea classification in Central–Southern Italy, particularly in the Intermediate Southern area, does not follow this approach—only three isoglosses are completely included within the boundaries of the area in question and have virtually no effect on the definition of subareas—but is rather grounded on administrative subdivisions. For example, the boundaries between Molisano and Southern Latian-Campanian, or between the latter and Apulian and Lucanian-Northern Calabrian, respectively, largely reflect the administrative boundaries between the corresponding regions.

The goal of this work is to investigate to what extent modern dialectometry confirms this standard classification. Dialectometry (Séguy, 1973; Goebl, 1982) aims at providing an objective view of dialect variation through the use of quantitative data analysis. In particular, dialectometric clustering has been applied to several regions, including the Netherlands (Wieling & Nerbonne, 2011), Catalonia (Valls et al. 2012), and English dialects (Wieling, Shackleton, & Nerbonne, 2013). In Italy, relevant examples mostly concern Tuscany (Montemagni & Wieling, 2016; Calamai, Piccardi, & Nodari, 2022).

In these works, various clustering techniques have been applied mainly on the basis of distance matrices, although other examples exist (Syrjänen et al., 2016). Distance matrices collect the linguistic distances between any pair of $N$ sites or areas. Linguistic distance has been defined in several different ways.

One common procedure consists in considering categorical lexical data, that is, $M$ entries in a linguistic atlas, which may have up to $P$ variants each. A distance between two sites is then defined by counting the number of pairwise variant mismatches for all features. An example is the Relative Difference Value (RDV), initially used as a difference function for unequivocal outcomes of features (Goebl, 2010) and later adapted to cover features with multiple possible outcomes (Pickl et al., 2014). A slightly modified

**Corresponding author:** Antonio Sciarretta; Email: antonio.sciarretta@ifpen.fr

**Cite this article:** Sciarretta A. (2024) Dialectometry-based classification of the Central–Southern Italian dialects. *Journal of Linguistic Geography* **12**: 13–23, https://doi.org/10.1017/jlg.2024.7

metric, the Weighted Identity Value (WIV), can use weights to emphasize some particular features (Goebl, 1982). This approach has been extended to variables/features other than lexical, that is, phonological, rules (Valls et al., 2012).

Another approach considers individual word pronunciations, which are converted in edit-distances between strings of characters, typically using one particular location as a reference. The most common edit distance used is Levenshtein distance (Levenshtein, 1966), which describes the cost (number of elementary operations) of changing one string into another or, equivalently, the character mismatches when the strings are opportunely aligned. More refined methods with variable costs of substitutions (weights) also exist, such as the PMI-based Levenshtein distance (Wieling et al, 2014). Once normalized by the length of alignment, the edit distances between $m$ word pairs can be then aggregated by taking their average (Heeringa, 2004), leading to the distance between two varieties.

Once a distance matrix is obtained, several analyses can be performed, the basic ones being beam maps, honeycomb maps, and cluster analysis. Among clustering techniques, hierarchical clustering, such as complete-linkage, UPGMA, or Ward's, has been more often used (Goebl, 2008). Partitional clustering has been somehow less used in dialectometry, although both k-means and k-medoid clustering have been applied on different kinds of linguistic data (Hyvönen, Leino, & Salmenkivi, 2007; Burridge et al., 2019; Cheshire, Mateos, & Longley, 2011; Syrjänen et al., 2016).

Hierarchical clustering or k-medoid can be used directly once the distance matrix is defined since these methods only need the distance metric between the sites. On the contrary, k-means requires one to evaluate the distance between actual sites and iteratively updated centroids, which do not correspond to any site, therefore preventing the use of pre-calculated distance matrices.

Dimensionality Reduction (DR) techniques try to reduce the number of variables while preserving the variation as much as possible. For instance, Bipartite Spectral Graph Partitioning (BSGP) using Singular Value Decomposition (SVD) has been used (Wieling & Nerbonne, 2011; Montemagni & Wieling, 2016; Wieling et al., 2013). This technique uses a binary segment substitution matrix ($N \times M$) with value $A_{ij} = 1$ when segment substitution $j$ occurs in variety $i$. SVD is applied to produce a synthetic vector of size $N + M$, which is then processed by k-means, in an attempt to simultaneously cluster sites and linguistic features that give rise to the geographical clustering.

Other dimensionality reduction techniques such as Multidimensional Scaling, Principal Component Analysis (PCA), or Factor Analysis (Pröll, Pickl, & Spettl, 2014) are usually used to discover indirectly latent clusters and dialect continua in the data, for example by converting the distance matrix into a $N \times 3$ matrix, then attributing RGB values to rows and visualizing them on maps. However, these DR techniques usually do not provide explicit clustering capability.

Recently, spatial Bayesian Clustering (BC) has been applied to linguistic data by Romano et al. (2022). While hard clustering generates clear boundaries between clusters and thus may fail to represent gradual variations in continuous dialect data, clustering is fuzzy in BC: Each point belongs to every cluster with a certain probability. Bayesian clustering yields core regions where points predominantly belong to a single cluster and gradual boundaries where points belong to multiple clusters with almost equal probabilities.

The data used for clustering are generally the entries of linguistic atlases. For the region under consideration, the web page of the Salzburg dialectometry team (Goebl et al., 2019) provides a classification based on the AIS (Jaberg & Jud, 1987) data and two hierarchical clustering algorithms. However, Central–Southern Italian dialects are classified alongside other Italian dialects: Even setting the number of clusters to the maximum value available (20), only four or five groups emerge in the region considered. Moreover, the results change dramatically depending on the corpus considered, which is probably due to the relatively low number of sites ($N$ less than 100) in the corpus.

In this work, we try to consider all Central–Southern Italian varieties, that is, more than a thousand communes in nine regions: Marche (south of the river Esino), Umbria, Latium, Abruzzi, Molise, Campania, Apulia, Basilicata, and Calabria. To obtain access to useful and homogeneous data, we select $L$ phonetic features (selected according to three guiding principles) instead of trying to gather a vocabulary of word entries. Then we apply k-means clustering to points in an abstract $L$-dimensional space. Each point represents a group of varieties that are homogeneous according to the selected phonetic features and can be represented as strings of numerical values that describe the outcomes of those features. Thanks to the relatively low dimension of the dataset ($N \times L$), clustering can be performed directly with the k-means algorithm, without the need of dimensionality reduction techniques. Distance can be calculated between any strings, also not representative of any variety, such as the k-means centroids. We adopt the silhouette analysis to choose the most appropriate number of clusters. Based on that, we propose a heuristic method to define fuzzy or transitional areas across groups.

## 2. Method

Varieties are classified according to $L = 18$ phonetic traits, which are listed in Table 1. These traits certainly represent a subset of the diatopic variation in the area considered. Their choice has been made according to three guiding principles:

- Being sufficiently compact in their areal distribution, thus avoiding the use of possibly widespread but "darting" phenomena occurring here and there, for example due to diachronic variation and the influence of Standard Italian. This criterion discarded, for example, the propagation of /u/ in pre-tonic position (Savoia & Baldi, 2016; Schirru, 2016) and the semivocalization of initial and intervocalic /v/.
- Being sufficiently widespread, concerning at least two or three provinces. For this reason, for example, the palatalization of pretonic /a/, which concerns a possibly compact but limited area in Molise (Iannacito, 2002), was discarded.
- Being sufficiently identifiable, that is, occurring in at least half a dozen words that can be retrieved in common speech, written texts, or the scientific literature. For this reason, for example, the different outcomes of -TJ- or -BJ- (Carosella, 2016), occurring in a very few common words, have been discarded.

Traditionally, most of these features are associated with "isoglosses" that have been used to define dialect groups or subgroups. For instance, phonetic trait 4 from Table 1 is the definitory isogloss that separates the Central dialects from Intermediate-Southern dialects in the classification of Pellegrini.

All varieties in the geographical space considered have been inspected and attributed a numerical value for each trait. Traits that have just two outcomes can generate either a digit 0 (in general, absence of that trait) or 1 (presence). Traits with multiple ($P$) outcomes can generate digits ranging from 0 to $P - 1$ where 0 is generally attributed to the "most standard" outcome, and the digit

**Table 1.** Set of phonetic features considered and their possible outcomes

| $\ell$ | Phonetic trait—Outcomes | Examples | $x_\ell$ | $w_\ell$ |
|---|---|---|---|---|
| 1 | Metaphony, given /-U/ | “bed” | | 0.5 |
| | Absent | ['lϵt:o] | 0 | |
| | Raising-type | ['let:u] | 1 | |
| | Diphthonigization-type | ['ljϵt:ə] | 2 | |
| | Monophthongization-type | ['lit:ə] | 3 | |
| 2 | Metaphony, given /-I/ | “good” (pl.) | | 0.5 |
| | Absent | ['bɔno] | 0 | |
| | Raising-type | ['bonu] | 1 | |
| | Diphthonigization-type | ['bwɔnə] | 2 | |
| | Monophthongization-type | ['bunə] | 3 | |
| 3 | Vocalic differentiation by position | “thing,” “mouth” | | 1 |
| | Absent | ['kɔsa], ['vok:a] | 0 | |
| | Present (central–southern origin) | ['kosa], ['vɔk:a] | 1 | |
| | Present (northern origin) | | -1 | |
| 4 | Word-final vowels | “house,” “heart,” “eight,” “wolf” | | 1 |
| | Reduction of all (/ə/) | ['kasə], ['kɔrə], ['ɔt:ə], ['lupə] | 0 | |
| | Conservation of -a, reduction of others (/a/, /ə/) | ['kasa], ['kɔrə], ['ɔt:ə], ['lupə] | 1 | |
| | Conservation of three (/a/, /e/-/ə/-/i/, /o/-/u/) or four vowels (with /i/ distinct from /e/-/ ə/) | ['kasa], ['kɔre]–['kori], ['ɔt:u], ['lupu] | 2 | |
| | Conservation of all five vowels (/a/, /e/, /i/, /o/, /u/) | ['kasa], ['kɔrə], ['ɔt:o], ['lupu] | 3 | |
| 5 | Alteration of -LL- | “horse” | | 1 |
| | Absent (/ll/) | [ka'val:u] | 0 | |
| | Palatal (/j/, /ʎ/, /ʝ/) | [ka'vaj:u] | 1 | |
| | Occlusive (/dd/) and retroflex | [ka'vad:u] | 2 | |
| 6 | Metaphony of -A- | “hands” | | 1 |
| | Absent | ['manə] | 0 | |
| | Present | ['minə] | 1 | |
| 7 | Some groups of consonants + L | “(it) rains,” “white,” “flower” | | 1 |
| | Standard (/pj/, /bj/, /fj/) | ['pjovə], ['b:jangə], ['fjorə] | 0 | |
| | Alteration of /PL/ > /kj/ | ['covə], ['b:jangə], ['fjorə] | 1 | |
| | Further alteration of /BL/ > /j/ | ['covə], ['jangə], ['fjorə] | 2 | |
| | Further alteration of /FL/ > /ʃ/, /x/ etc. | ['covə], ['jangə], ['ʃorə] | 3 | |
| 8 | Apocope of -no, -ne | “bread,” “wine” | | 0.5 |
| | Absent | ['pane], ['vino] | 0 | |
| | Only -ne | ['pa], ['vino] | 1 | |
| | Both | ['pa], ['vi] | 2 | |
| 9 | Outcomes of -LJ- | “son” | | 1 |
| | Palatal (/ʎ/) | ['fiʎə] | 0 | |
| | Approximant (/j/) | ['fijə] | 1 | |
| | Occlusive (/ɟ/) | ['fiɟə] | 2 | |
| 10 | Aspiration of -F- | “coffee” | | 1 |
| | Absent | [ka'fe] | 0 | |
| | Present | [ka'he] | 1 | |
| 11 | Rhotacization of -D- | “tooth” | | 1 |
| | Absent | ['dϵndə] | 0 | |
| | Present | ['rϵndə] | 1 | |

(*Continued*)

**Table 1.** (*Continued*)

| $\ell$ | Phonetic trait—Outcomes | Examples | $x_\ell$ | $w_\ell$ |
|---|---|---|---|---|
| 12 | Degemination of -RR- and other geminates | “ground” | | 1 |
| | Absent | ['tɛr:a] | 0 | |
| | Present (of -rr-) | ['tɛra] | 1 | |
| | Present (of -rr- and others) | ['tɛra] | 2 | |
| 13 | Postnasal sonorization of stops and progressive assimilation in groups of /n/ + stops | “spring,” “when” | | 1 |
| | Both present | ['fonde], ['kwan:o] | 0 | |
| | Only assimilation | ['fonte], ['kwan:o] | 1 | |
| | Both absent | ['fonte], ['kwando] | 2 | |
| 14 | “Florentine” Anaphonesis | “tongue” | | 1 |
| | Absent | ['leŋgwa] | 0 | |
| | Present | ['liŋgwa] | 1 | |
| 15 | Some groups of consonants + J | “arm,” “to eat,” “to go out” | | 1 |
| | Standard (/ʧ/, /ɲ/, /j/) | ['vraʧ:ə], [ma'ɲ:a], ['ji] | 0 | |
| | Alteration of /kj/ > /ʦ/ | ['vraʦ:ə], [ma'ɲ:a], ['ji] | 1 | |
| | Further alteration of /ngj/ > /nʤ/ | ['vraʦ:ə], [ma'nʤ:a], ['ji] | 2 | |
| | Further alteration of /j/ > /ʃ/ | ['vraʦ:ə], [ma'nʤ:a], ['ʃi] | 3 | |
| 16 | Group R + J | “baker” | | 1 |
| | Central–southern /r/ | [for'naro] | 0 | |
| | Tuscan /j/ | [for'najo] | 1 | |
| 17 | Group S + J | “kiss” | | 1 |
| | Postalveolar (/ʃ/) | ['vaʃə] | 0 | |
| | Alveolar (/s/) | ['vasə] | 1 | |
| 18 | Tonic vowel system | “snow,” “month,” “cross” | | 1 |
| | Common Romance | ['nevə], ['mesə], ['kroʧə] | 0 | |
| | “Romanian” | ['nevə], ['mesə], ['kruʧə] | 1 | |
| | “Sardinian” | ['nivə], ['mɛsə], ['kruʧə] | 2 | |
| | “Sicilian” | ['nivə], ['misə], ['kruʧə] | 3 | |

increases with the degree of deviation from this standard. The numerical values of each outcome are also listed in Table 1. In case of intermediate, simultaneous, or uncertain outcomes, sometimes fractional values have been used.

Resulting from this encoding, each dialect corresponds to a string of $L$ digits, $\{x_\ell\}_1^L$. Varieties that are geographically adjacent and share the same string are considered as equal and form a “homogeneous area” (HA) for the purposes of this study. In the whole space, no less than $N = 647$ homogeneous areas have been identified in this way: 111 in Latium, 101 in Calabria, 89 in the Abruzzi, 83 in Campania, 79 in Basilicata, 76 in Apulia, 40 in Molise, 44 in Marche, 24 in Umbria. The localization of these areas is shown schematically in Figure 1. Their actual extension and the varieties included in each of them are detailed in the companion web site.[1]

Each homogeneous area represents one point in the data set used for the classification. The metrics used is the Manhattan distance

$$D_{ij} = \sum_{\ell} w_\ell \left| x_{i\ell} - x_{j\ell} \right| , \tag{1}$$

where $|\cdot|$ denotes the absolute value and $w$ is a vector of weights. In this study, $w_\ell$ is always 1 except for $\ell = \{1, 2, 8\}$ where $w = 0.5$ has been used; see Table 1. We note that this procedure is roughly equivalent to “count the isoglosses” between two different locations.

Based on this metric, a k-means algorithm has been used to classify the $N$ $L$-dimensional points into $K$ groups. This well-known algorithm tries to attribute each point to one of the clusters by minimizing the within-cluster sum of Manhattan distances, that is,

$$\min \sum_{k=1}^{K} \sum_{x_i \in C_k} \sum_{\ell=1}^{L} w_\ell |x_{i\ell} - m_{k\ell}| , \tag{2}$$

where the centroid $m_k$ is defined as the mean of points belonging to cluster $k$ $(C_k)$,

$$m_k = \frac{1}{card(C_k)} \sum_{x_j \in C_k} x_j . \tag{3}$$

In practice, the algorithm proceeds iteratively. First, a set of $K$ means is randomly generated. Then, each point is attributed to the cluster with the “nearest” mean. Further, means are recalculated based on the points attributed to each cluster. This process is repeated for $T$ iterations. However, the algorithm is not guaranteed

https://doi.org/10.1017/jlg.2024.7 Published online by Cambridge University Press

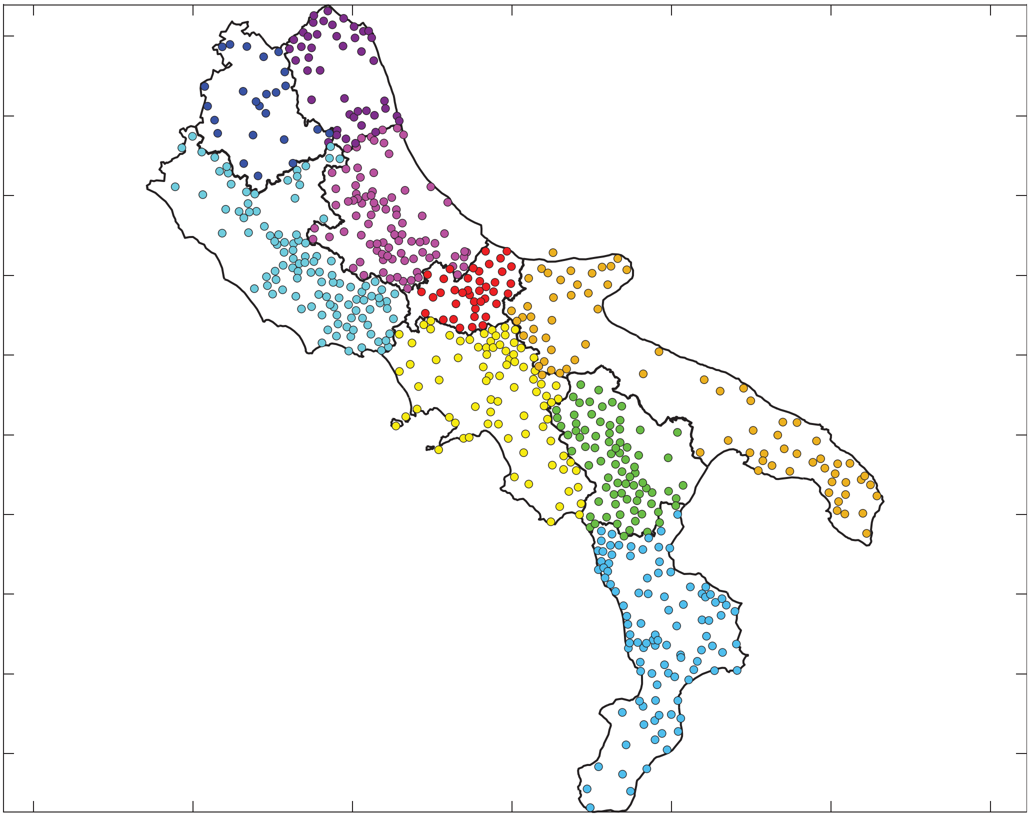

**Figure 1.** Localization of the homogeneous areas (circles). Each color corresponds to one of the administrative regions. Boundaries between regions are drawn

to find the optimum, that is, the clustering that minimizes the objective in (2) (Russel & Norvig, 2020). For this reason, the algorithm is run for $R$ times, each time with a different (random) initialization of the means. For each run, the objective is calculated, and finally the run with the minimal objective is chosen as the result. For this study, the algorithm is parametrized with $T = 20$ and $R = 200K$.

To choose the optimal number of clusters $K$, the *silhouette analysis* is used. According to this method, a silhouette metric is defined as a function of the number of clusters as

$$\sigma(K) = \left\langle \frac{b_i - a_i}{\max(a_i, b_i)} \right\rangle, \tag{4}$$

where $\langle\cdot\rangle$ denotes the average over all points $i$, and

$$a_i = \frac{1}{\mathrm{card}(C_I) - 1} \sum_{j\in C_I, j\neq i} D_{ij}\ , \quad b_i = \min_{J\neq I} \frac{1}{\mathrm{card}(C_J)} \sum_{j\in C_J} D_{ij} \tag{5}$$

are the mean distance between point $i$ and all other points in the same cluster $C_I$ and the smallest mean distance of $i$ to all points in any other cluster, respectively. The optimal number of clusters is chosen so as to maximize the silhouette. The silhouette coefficient $SC = \max_K \sigma(K)$ summarizes the final result.

It is a common opinion that belonging to one particular dialectal group is not a rigid attribute; instead, transition bands exist. To confirm this intuition quantitatively, we have used the following method. We compute the distance between each HA and the centroids of all clusters,

$$D_i^k = \sum_{\ell} w_\ell |x_{i\ell} - m_{k\ell}|\ . \tag{6}$$

The lowest distance corresponds by definition to the cluster $k$ to which the HA is member. If the difference between the second lowest distance (say, with cluster $h$) and the lowest distance is less than a specified fraction of the lowest distance, then that HA is marked as a transitional area between cluster $k$ and cluster $h$,

$$i \in C_{KH} \ \text{ if } \ D_i^h < D_i^l \ \ \forall l \neq \{k,h\} \ \cap D_i^h < (1+\xi) D_i^k. \tag{7}$$

## 3. Data

Data for all varieties considered (see Table 2 for provinces and province codes) have been collected from multiple and diverse sources, including material covering the phonetics of specific varieties (see Selected Sources: Specific Varieties on the companion website), larger areas or entire regions (see Selected Sources: Larger Areas on the companion website), comprehensive monographies (see Selected Sources: Comprehensive Monographies on the companion website), and linguistic atlases (see Selected Sources: Linguistic Atlases on the companion website), including acoustic atlases. Other speech material available on the web, both ethnographic and spontaneous, has also provided data for certain dialectal traits. Good dialectal dictionaries, although often written by non-professional researchers, have been found for many varieties. Dialectal literature (mostly poetry) in specific varieties and collections covering broader areas has been also perused, particularly for those traits that are unambiguous when written. Many of these non-scholarly sources are listed in the companion website. Less canonically, many data have been obtained by

https://doi.org/10.1017/jlg.2024.7 Published online by Cambridge University Press

**Table 2.** List of province codes. Regional capital cities in bold

| MARCHE | | CAMPANIA | |
|---|---|---|---|
| **Ancona** | AN | Avellino | AV |
| Ascoli Piceno | AP | Benevento | BN |
| Fermo | FM | Caserta | CE |
| Macerata | MC | **Napoli** | NA |
| UMBRIA | | Salerno | SA |
| **Perugia** | PG | APULIA (*Puglia*) | |
| Terni | TN | **Bari** | BA |
| LATIUM (*Lazio*) | | Barletta-Andria-Trani | BT |
| Frosinone | FR | Brindisi | BR |
| Latina | LT | Foggia | FG |
| Rieti | RI | Lecce | LE |
| **Roma** | RM | Taranto | TA |
| Viterbo | VT | BASILICATA | |
| ABRUZZI (*Abruzzo*) | | Matera | MT |
| **L'Aquila** | AQ | **Potenza** | PZ |
| Chieti | CH | CALABRIA | |
| Pescara | PE | Cosenza | CS |
| Teramo | TE | **Catanzaro** | CZ |
| MOLISE | | Crotone | KR |
| **Campobasso** | CB | Reggio di Calabria | RC |
| Isernia | IS | Vibo Valentia | VV |

**Table 3.** Clustering results as a function of the number of clusters *K*

| $K$ | $\sigma$ | Main new divide (w.r.t. $K-1$) | Groups identified |
|---|---|---|---|
| **2** | **(0.397±0.000)** | Salerno–Lucera–Vieste (SLV) | Northern space vs. Southern space |
| 3 | (0.364±0.001) | Gaeta–Sora–Termoli (GST), Alento–Agri–Taranto–Brindisi | Northern, **Central**, Southern subspaces |
| **4** | **(0.366±0.00)** | GST, SLV, Alento–Crati–Nardò–Brindisi | Northern, **Campanian-Molisan, Apulian-Lucanian**, Southern subspaces |
| 5 | (0.336±0.003) | Sora–L'Aquila–S. Benedetto | Northern subspace, **Abruzzese**, Campanian-Molisan, Apulian-Lucanian, Southern subspaces |
| 6 | (0.323±0.003) | Foggia–Potenza–Cassano | Northern subspace, Abruzzese, Campanian-Molisan, **Apulian**, **Irpino-Lucanian**, Southern subspace |
| 7 | (0.322±0.000) | Pollino–Sila–Lamezia | Northern subspace, Abruzzese, Campanian-Molisan, Apulian, Irpino-Lucanian, **Cosentino**, **Salentino-Calabrian** |
| **8** | **(0.322±0.001)** | Latina–Ancona | **Perimedian, Median**, Abruzzese, Campanian-Molisan, Apulian, Irpino-Lucanian, Cosentino, Salentino-Calabrian |
| 9 | (0.315±0.002) | Irregular | Perimedian, Median, Abruzzese, **Samnite**, **Neapolitan-Molisan**, Apulian, Irpino-Lucanian, Cosentino, Salentino-Calabrian |
| 10 | (0.315±0.005) | Irregular | As above, but Salentino split from Calabrian |
| 11 | (0.313±0.001) | Irregular | As above, but Irpino-Lucanian split in two groups |

inspecting, searching, and sometimes querying dialect-oriented groups on social networks such as Facebook. Older scholarly data have been systematically verified or discarded by perusing the (written) conversations found in these groups.

As a result, a database containing thousands of observations has been prepared and is available to the readers upon request to the author. Based on the database, the strings for each variety have been constructed and the homogeneous areas identified.

Inspection of unclassified results already provides some useful insight. For instance, it is possible to graphically represent on a map the distances from a given HA, creating similarity maps as defined by Goebl et al. (2019). Moreover, "isogloss maps" and "beam maps" have been also created. Examples of the latter for all regions considered are shown in the companion website, where only "beams" corresponding to distances $D \leq 1$ are plotted, depicting the emergence of dialectal continua. However, this analysis yields many small continua and a large number of isolated areas (whose distance with all conterminous areas is larger than 1), making a significant classification impossible. For this purpose, the most useful analysis is that of clustering, presented in the next section.

## 4. Results

Clustering with several values of $K$ ranging from 2 to 11 have been run and inspected. For higher values of $K$ the overall results become very sensitive to the random initialization, are unstable, and thus are not shown. Table 3 summarizes the main divides (traditionally, the "isoglosses") that characterize each new partition, as well as the new groups that emerge from it.

The silhouette factor as a function of $K$ is shown in Figure 2. Values are given as the mean of four series of runs plus/minus the standard deviation. When the latter is small, it means that the results are stable when different series of runs are executed. As can be observed, the factor $\sigma$ generally decreases with the number of clusters, with the coincidence intervals of two consecutive $K$ that are generally not overlapping. However, three values of $K$ emerge as local maxima, namely, $K = 2,\ 4$, and 8. These partitions are all very stable, as evidenced by a variance of the silhouette coefficient low or null. A fourth cluster number, $K = 10$, has a mean silhouette factor that is close to that with $K = 9$ and presents an overlap of the respective confidence intervals, meaning that for some series of runs the silhouette could be higher with $K = 10$ than with $K = 9$. However, partition $K = 10$ is also the least stable, with a large variance due to concurrent clustering results.

The first optimal classification with $K = 2$ divides the overall space considered into a Northern and a Southern space, separated by a line that resembles the traditional Salerno–Lucera (actually, Salerno–Lucera (FG)–Vieste (FG), SLV) isogloss bundle. For instance, around this line lays the northern limit of KJ > /tʦ/ (see trait 15 in Table 1).

https://doi.org/10.1017/jlg.2024.7 Published online by Cambridge University Press

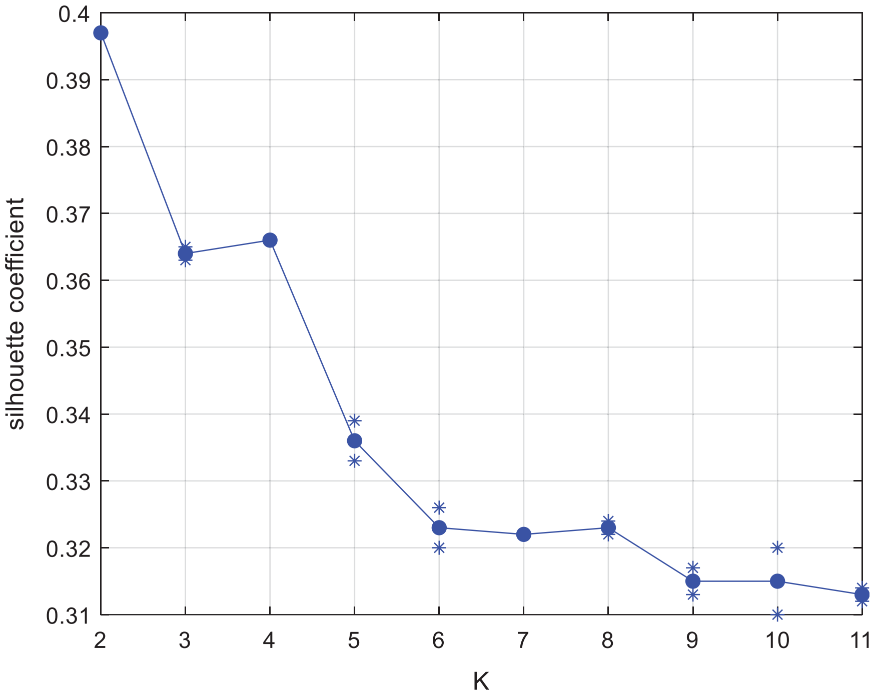

**Figure 2.** Silhouette coefficient as a function of the number of groups

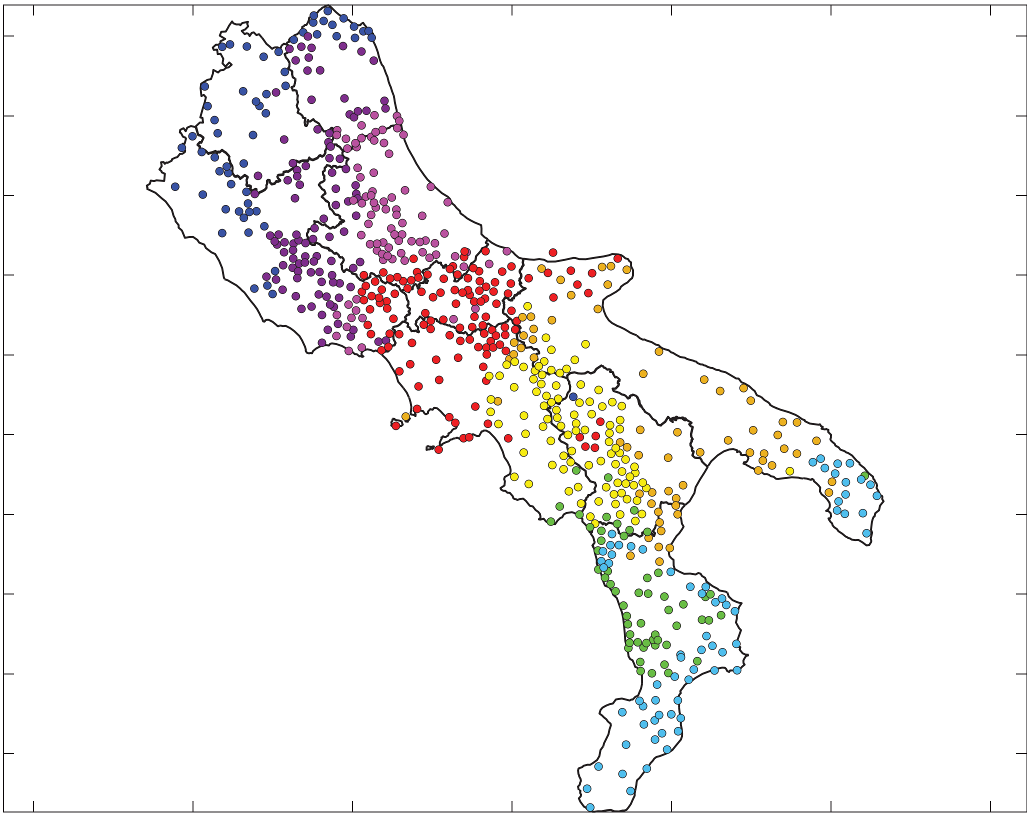

**Figure 3.** HA clustered in $K = 8$ groups: schematic representation. Each color corresponds to one group: blue (Perimedian), purple (Median), pink (Abruzzese), red (Campanian-Molisan), orange (Apulian), yellow (Irpino-Lucanian), green (Cosentino), light blue (Salentino-Calabrian)

In the partition with $K = 4$ each of these subspaces splits in two. Thus, a Northern subspace is separated from a Central-Northern subspace by a line running from around Gaeta (LT) on the Tyrrhenian coast to around Termoli (CB) on the Adriatic coast, with an elbow around Sora (FR). The Central-Northern subspace is separated from a Central–Southern subspace by an SLV line, although not exactly coincident with the previous one. Finally, the Central–Southern subspace is separated from a Southern subspace by two lines, one running from around the mouth of the Alento river (SA) on the Tyrrhenian coast to the mouth of the Crati river (CS) on the Ionian coast, the other running from around Nardò (LE) on the Ionian coast to around Brindisi on the Adriatic coast.

https://doi.org/10.1017/jlg.2024.7 Published online by Cambridge University Press

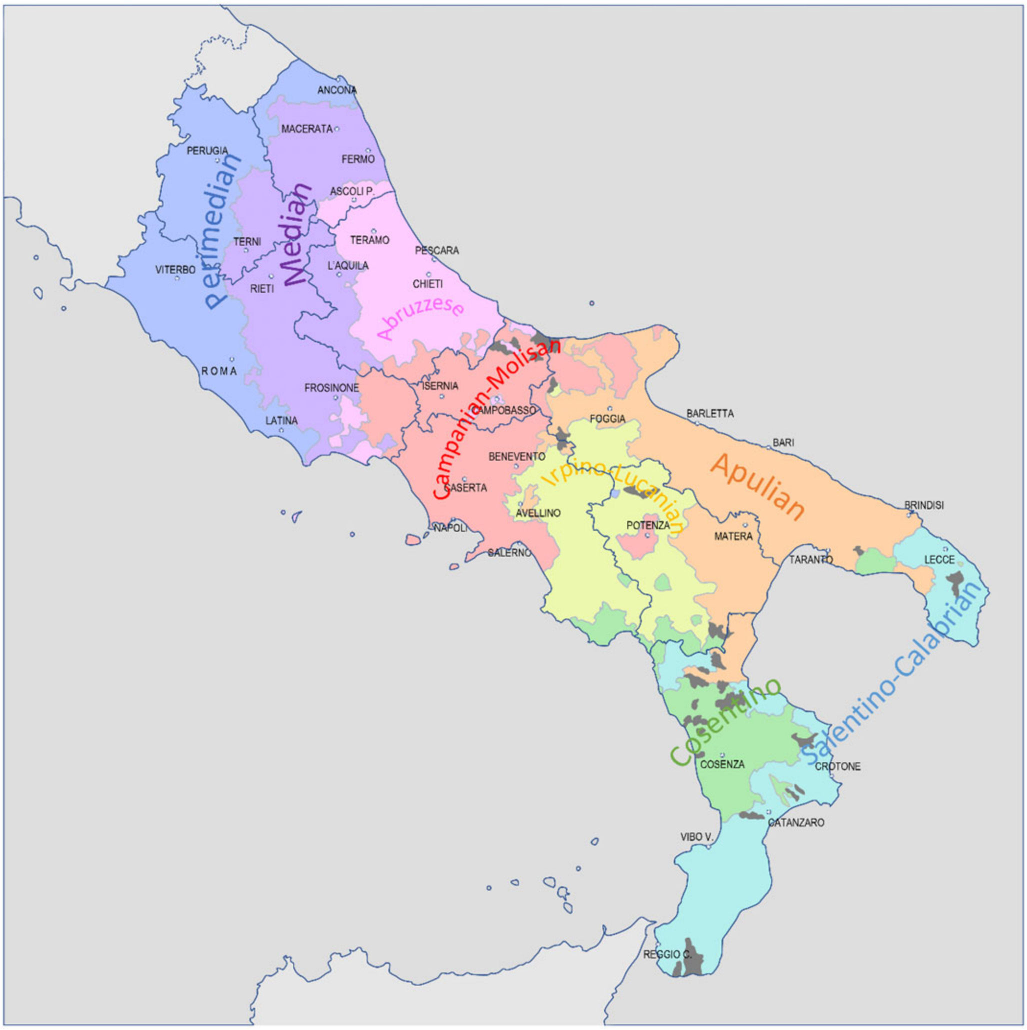

**Figure 4.** HA clustered in $K = 8$ groups: a linguistic map with actual group boundaries. Colors of groups correspond to those of Figure 3.

The next optimal classification is with $K = 8$. Incidentally, this value of $K$ almost matches the number of administrative regions (*Regioni*) in the space considered. This partition could be thus a promising basis for the definition of more accurate "regional languages" in this half of Italy. The groups identified by clustering with $K = 8$ are listed in Table 3 and detailed here from north to south as:

1. "Perimedian" group, including provincial capitals AN, PG, VT, RM, LT, and areas in Northern Marche, in Central-Western Umbria, in Western Latium, besides a hamlet (*frazione*) in Basilicata (a Marchigiano colony).
2. "Median" group, including provincial capitals MC, FM, TN, RI, FR, AQ, and areas in Central Marche, South-Eastern Umbria, Central Latium, Western Abruzzi.
3. "Abruzzese" group, including provincial capitals AP, TE, PE, CH, and areas in Southern Marche, Eastern Abruzzi, Southern Latium, besides a few smaller areas in Molise.
4. "Campanian-Molisan" group, including provincial capitals IS, CB, BN, NA, CE, SA, and areas in South-Eastern Latium, Molise, Northern-Central Campania, besides some smaller areas in Northern Apulia and Basilicata around and including the provincial capital PZ (Gallo-Italic colonies).
5. "Apulian" group, including provincial capitals FG, BAT, BA, TA, MT, BR, and areas in Northern-Central Apulia, South-Eastern Basilicata, North-Eastern Calabria, besides some smaller areas in Central Campania.
6. "Irpino-Lucanian" group, including provincial capital AV and areas in South-Eastern Campania, Western Basilicata, and North-Eastern Apulia.
7. "Cosentino" group, including provincial capital CS and areas in Southern Campania (likely having a Greek substratum), Northern Calabria, besides some smaller areas in Basilicata (most of them being or having been Gallo-Italic colonies).
8. "Salentino-Calabrian" group, including provincial capitals LE, KR, VV, CZ, RC, and areas in Southern Apulia and Central–Southern Calabria, besides some smaller areas in Northern Calabria.

Figure 3 shows the attribution of each HA to one of the nine clusters, identified by a color. It must be noted that the K-means algorithm has no knowledge about the spatial correlation between the HA, each of them representing a "point" in an 18-dimensioned space, with these points that can be geographically ordered in any arbitrary way. However, the spatial consistency of the results is striking, and the groups obtained clearly recall traditional regions and dialectal groups. The actual boundaries between the eight groups can be traced on a map, as depicted in Figure 4.

The boundary between groups 1 and 2 recalls a well-known isogloss, the Northern limit of simultaneous NT > /nd/ and ND >

https://doi.org/10.1017/jlg.2024.7 Published online by Cambridge University Press

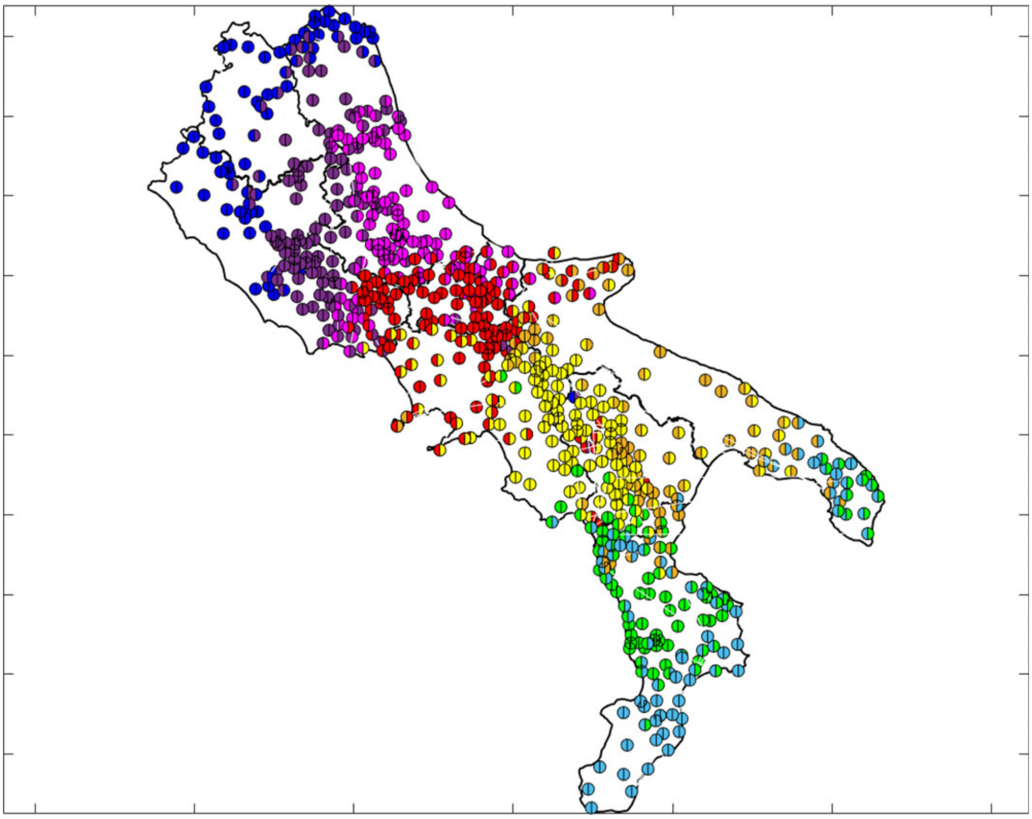

**Figure 5.** HA clustered in $K = 8$ groups with second-best clusters: schematic representation. Core clusters are identified by the left-half color of the circles; second-best clusters in transitional area are identified by the right-half color

/nn/ (see trait 13 in Table 1), which traditionally separates the Central Italian dialects into a “Perimedian” and a “Median” section (whence the naming of groups 1 and 2 used here). Boundary 2–3 runs similarly to another definitory isogloss, the Northern limit of [ə] (see trait 4 in Table 1), which serves to separate Central from Southern (“Neapolitan language”) dialects in traditional classifications. Boundary 3–4, or the GST bundle introduced above, is similar to the Northern limit of PL > /kj/ (see trait 7 in Table 1) or isogloss 21 in Pellegrini’s map. The boundary between 4 on one side and 5 and 6 on the other is the SLV bundle discussed above. The boundary between 5 and 6 on one side and 7 and 8 on the other recalls the Northern limit of non-standard tonic vowel systems (trait 18 in Table 1), which is different from isogloss 25 (Southern limit of [ə]), which is traditionally used to separate the Intermediate Southern dialects from the Extreme Southern dialects (“Sicilian” language). Finally, the North–South boundary between groups 6 and 7 on one side and 5 and 8 on the other matches almost perfectly a less used isogloss—the Western limit of LJ > /ɟ/ (see trait 9 in Table 1)—whereas in traditional classifications the corresponding boundaries are purely administrative.

Figure 5 (schematic view) and Figure 6 (pictorial) show the transitional areas identified with the method (7) of second-best clusters (with $\xi = 0.5$). These results suggest the existence of such areas at the geographical boundary between groups 1 and 2 (in Marche, Umbria, and Latium), 2 and 3 (in Marche, Abruzzi, and Latium), 2 and 4 (in Latium), 3 and 4 (in Latium, Abruzzi, Molise, and Apulia), 4 and 5 (in Apulia), 4 and 6 (in Campania), 5 and 6 (in Apulia and Basilicata), 6 and 7 (in Basilicata), 5 and 7 (in Calabria), 5 and 8 (in Apulia and Calabria), 7 and 8 (in Calabria). Again, these results are consistent with geography in the sense that transitional areas are generally identified between clusters that actually share a geographical border.

## 5. Conclusions

In this paper, we have presented a dialectometry-based study aimed at classifying the Romance varieties of Central–Southern Italy. We have analyzed the thousands of varieties under study and operated a massive pre-treatment of data available from many sources. Instead of trying to gather a vocabulary of word entries, we opted for a reduced data set, where each variety is characterized with respect to 18 phonetic traits, including the isoglosses that have been traditionally used by linguists to define dialectal groups. On this basis we have identified 647 homogeneous areas grouping conterminous varieties that share the same traits. As a result, we have got an operating data set of 647 points in an 18-dimensional space, where we could define linguistic distances. We have then formulated the problem as a clustering problem, that is, find the $K$ clusters of those points that minimize the within-cluster linguistic distance. We have used a k-means algorithm to cluster and an ad-hoc rule to define second-best clusters and transitional areas. We have used silhouette analysis to select the most appropriate number of clusters.

The results are geographically consistent, although the algorithms used have no information about the actual geographical distance between areas or the boundaries shared by them. The groups identified for various numbers $K$ resemble but do not coincide with the regional varieties traditionally invoked. For example, when the partition with $K = 3$ is compared with the traditional high-level (Pellegrini’s areas) tripartite grouping into Central, Intermediate Southern, and Extreme Southern, the results do not match unless the Central area includes the Abruzzese.

The methods used suggest that clustering with 8 groups is the most appropriate choice. The dialectal groups identified (labelled as Perimedian, Median, Abruzzese, Campanian-Molisan, Apulian,

https://doi.org/10.1017/jlg.2024.7 Published online by Cambridge University Press

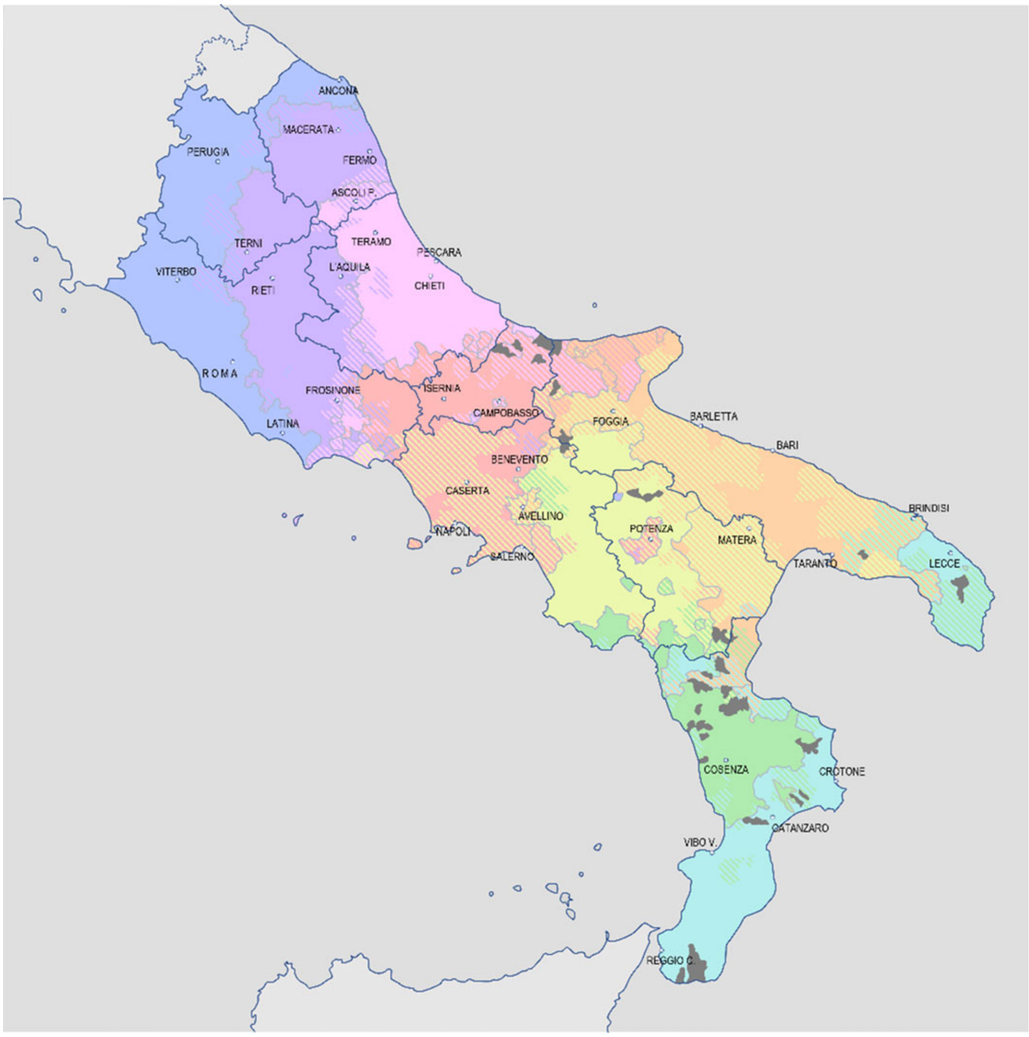

**Figure 6.** HA clustered in $K = 8$ groups with second-best clusters: a linguistic map with actual group boundaries and transitional areas (hatched)

Irpino-Lucanian, Cosentino, and Salentino-Calabrese) again do not coincide with the regional varieties (Pellegrini's subareas) traditionally invoked. The six geographic boundaries that can be roughly traced between them (considering that the geographical representations of the clusters are not perfectly connected in the topological sense) loosely run along known isoglosses, which are all among the 18 traits considered. However, in no way are these isoglosses favored a priori; it is the algorithm that "naturally" selects them in the optimization process, which in turn depends on the entire set of traits considered. This contrasts with the traditional classification that is based on a mixture of fewer definitory isoglosses and administrative or historical boundaries.

We conclude that a classification based on these grounds is less arbitrary than traditional ones as it considers multiple dialectal traits on an equal footing. It is also less subjective since the partitioning is made by an algorithm that tries to minimize a clearly defined objective function. Another strength of the method is that it can be readily adapted as long as new data are available, varieties evolve, or corrections are made to the data set.

**Supplementary material.** The supplementary material for this article can be found at https://doi.org/10.1017/jlg.2024.7.

**Competing interests.** The author declares none.

## Note

[1] The author maintains a popularization website (in Italian) at www.asciatopo.altervista.org/dialettolog.html, which contains a detailed description of the varieties and their arrangement in homogeneous areas and several linguistic maps.

## References

Burridge, James, B. Vaux, M. Gnacik, & Y. Grudeva. 2019. Statistical physics of language maps in the USA. *Physical Review E 99*. 032305. doi: 10.1103/PhysRevE.99.032305.

Calamai, Silvia, D. Piccardi, & R. Nodari. 2022. Quantifying folk perceptions of dialect boundaries. *A case study from Tuscany (Italy). Journal of Linguistic Geography 10*(2). 87–111.

Carosella, Maria. 2016. Per una ridefinizione delle sezioni orientali della Cassino-Gargano e della Salerno (o Eboli)-Lucera. *L'Italia dialettale. Rivista di dialettologia italiana 77*. 7–92.

Cheshire, J. A., P. Mateos, & P. A. Longley. 2011. Delineating Europe's cultural regions: Population structure and surname clustering. *Human Biology 83*. 573–598.

Eberhard, David M., G. F. Simons, & C. D. Fennig (eds.). 2022. *Ethnologue: Languages of the World*, 25th edn. Dallas, TX: SIL International. Online version: www.ethnologue.com.

Goebl, Hans. 1982. *Dialektometrie*, vol. 157. Vienna: Verlag der Osterreichischen Akademië der Wissenschaften.

Goebl, Hans. 2008. Le laboratoire de dialectométrie de l'Université de Salzbourg. Un bref rapport de recherche. *Zeitschrift für französische Sprache und Literatur 118*(1). 35–55.

Goebl, Hans. 2010. Dialectometry and quantitative mapping. In A. Lameli, R. Kehrein, & S. Rabanus (eds.), *Language and space. An international handbook of linguistic variation, vol. 2: Language mapping*, 433–457. Berlin & New York: de Gruyter.

Goebl, Hans, Edgar Haimerl, Pavel Smečka, Bernhard Castellazzi, & Yves Scherrer. 2019. *Dialectometry AIS*. Retrieved from http://dialektkarten.ch/dmviewer/ais/index.en.html (March 2024).

Hammarström, Harald, R. Forkel, M. Haspelmath, & S. Bank. 2022. *Glottolog 5.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from http://glottolog.org (March 2024).

Heeringa, W. J. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. dissertation. Groningen: University of Groningen.

Hyvönen, Saara, Antti Leino, & Marko Salmenkivi. 2007. Multivariate analysis of Finnish dialect data—an overview of lexical variation. *Literary and Linguistic Computing 22*(3). 271–290.

Iannacito, Roberta. 2002. L'assimilazione progressive nel dialetto molisano di Villa San Michele (IS). *Italica 79*(4). 509–524.

Jaberg, Karl & Jacob Jud. 1987. *Atlante linguistico ed etnografico dell'Italia e della Svizzera meridionale* (AIS). Milan: Unicopli. Retrieved from https://navigais-web.pd.istc.cnr.it/ (March 2024).

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10*(8). 707–710.

Montemagni, Simonetta & Martijn Wieling. 2016. Tracking linguistic features underlying lexical variation patterns: A case study on Tuscan dialects. In M.-H. Côté, R. Knooihuizen, & J. Nerbonne (eds.), *The future of dialects*, 117–135. Berlin: Language Science Press.

Moseley, Christopher (ed.). 2010. *Atlas of the world's languages in danger*, 3rd edn. UNESCO Publishing. Online version https://unesdoc.unesco.org/ark:/48223/pf0000187026.

Pellegrini, Giovan Battista. 1977. *Carta dei dialetti d'Italia*. Pisa: Pacini.

Pickl, Simon, Aaron Spettl, Simon Pröll, Stephen Elspass, Werner König, & Volker Schmidt. 2014. Linguistic distances in dialectometric intensity estimation. *Journal of Linguistic Geography 2*(1). 25–40.

Pröll, Simon, Simon Pickl, & Aaron Spettl. 2014. Latente Strukturen in geolinguistischen Korpora. In Michael Elmentaler, Markus Hundt, & Jürgen Erich Schmidt (eds.), *Deutsche Dialekte. Konzepte, Probleme, Handlungsfelder. Akten des 4. Kongresses der Internationalen Gesellschaft für Dialektologie des Deutschen (IGDD)*, vol. *4*, 247–258. Stuttgart: Steiner.

Romano, Noemi, P. Ranacher, S. Bachmann, & Stéphane Joost. 2022. Linguistic traits as heritable units? Spatial Bayesian clustering reveals Swiss German dialect regions. *Journal of Linguistic Geography 10*(1). 11–22.

Russel, Stuart & Peter Norvig. 2020. *Artificial intelligence: A modern approach*. Englewood Cliffs, NJ: Prentice Hall.

Savoia, Leonardo M. & Benedetta Baldi. 2016. Propagation and preservation of rounded back vowels in Lucanian and Apulian varieties. *Quaderni di Linguistica e Studi Orientali/Working Papers in Linguistics and Oriental Studies 2*. 11–58.

Schirru, Giancarlo. 2016. Propagginazione e flessione nominale in alcuni dialetti italiani centro-meridionali. *Atti del Sodalizio Glottologico Milanese 8–9*. 121–130.

Séguy, Jean. 1973. La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane 37*. 1–24.

Syrjänen, Kaj, Terhi Honkola, Jyri Lehtinen, Antti Lein, & Outi Vesakoski. 2016. Applying population genetic approaches within languages. *Language Dynamics and Change 6*. 235–283.

Valls, Esteve, John Nerbonne, Jelena Prokić, Martijn Wieling, Esteve Clue, & Maria Rosa Lloret. 2012. Applying the Levenshtein distance to Catalan dialects: A brief comparison of two dialectometric approaches. *Verba: anuario galego de filoloxia 39*. 35–61.

Wieling, Martjin, J. Bloem, K. Mignella, M. Timmermeister, & John Nerbonne. 2014. Validating and using the PMI-based Levenshtein distance as a measure of foreign accent strength. Poster presented at Methods in Dialectology XV, Groningen (The Netherlands), August 11–14, 2013.

Wieling, Martijn & John Nerbonne. 2011. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language 25*. 700–715.

Wieling, Martijn, R. G. Shackleton, & John Nerbonne. 2013. Analyzing phonetic variation in the traditional English dialects: Simultaneously clustering dialects and phonetic features. *Literary and Linguistic Computing 28*(1). 31–41.

https://doi.org/10.1017/jlg.2024.7 Published online by Cambridge University Press