Measuring dialect differences*

John Nerbonne and Wilbert Heeringa

Abstract

We measure varietal differences in general, and differences with respect to standard languages in particular ("dialectality", in Herrgen/Schmidt's sense) in order to systematize observations about dialect differences, to make sense of exceptions, and to enable measurements based on randomly selected material, thus obviating issues of potential bias. Finally, measurements allow the characterization of abstract relations among language varieties.

We illustrate some issues with simple techniques for categorical data introduced by Séguy and refined by Goebl, viz., issues concerning frequency, irrelevant variation, and competing forms. We proceed to measuring pronunciation differences, focusing on differences in the pronunciation of the same words in different varieties. Caution is needed to isolate pronunciation differences from differences in inflectional morphology, sandhi, and intonation. We characterize the difference between sound segments and develop a measure of the difference between the sequences of those segments in words, including insertions, deletions, and swaps (epenthesis, elision and metathesis).

Automating measurement techniques exposes the issue of validation, which lay largely unexamined in earlier dialectology. We propose to validate measurements based on the degree to which they correlate with dialect speakers' judgments of difference, justified by the presumed function of linguistic variation, that of signaling provenance.

1 Why Measure?

Dialectology blessedly inherits large data reserves from earlier practitioners, especially the compilers of dialect atlases designed to display variation in comparable linguistic items. These compilers collected linguistic variants such as the lexical realization of the word for 'house fly', the common order of pronominal objects, or the pronunciation of (the first vowel in) the word 'marry' throughout a large selection of sites in a language area. We refer to such items of variation as features or variables. Large collections may include several hundred sites at which hundreds of features are documented. One of the primary tasks of dialectology is to characterize this variety of speech forms. Naturally there are many others tasks, e.g. determining the extralinguistic correlates of variation, modeling the cognitive processes needed to deal with variation in form, and

* To appear in: Jürgen Erich Schmidt & Peter Auer (eds.) *Theories and Methods*, vol. within series *Language and Space*. Mouton De Gruyter: Berlin. ca. 2009.

understanding the tension between variation and effective comprehension, but many of these require good primary characterizations of the linguistic variation.

The data reserves are so large and complex that simple attempts to characterize the differences always encounter problems of different sorts, many of which are well documented. Bloomfield (1933, Chap.19) discusses how linguistic variants often do not map neatly to geography, noting exceptions of two types, linguistic exceptions, in which linguistically similar material does not project to geography in the same way (the same etymological vowel is realized differently in two different words) and geographic exceptions, where linguistic variants project fairly simply but where pockets of exception remain. A larger problem is that even very thorough studies (Kurath and McDavid 1961, König 1994) inevitably examine only a very small fraction of the features documented in substantial atlases. This leaves dialectology open to the charge statisticians have dubbed "cherry picking", i.e. picking variables that confirm the analysis one wishes to settle on. Given the tens of thousands of features which may differ between languages (in lexical items alone!), we would do well to take measures to avoid the danger of fortuitous selection.

Finally, we are motivated to develop measurement techniques in order develop hypotheses about variation that require abstract characterizations. These indeed seem within reach given the right level of abstraction.

We preview briefly how the use of measurement schemes attacks these issues. By *measuring* differences we map them to numbers. This numerical characterization enables us to integrate information from large samples of linguistic variables — just by adding the differences. We postulate that the pronunciation difference between 'night' as pronounced diphthongally in standard American, [naɪt], and the same word pronounced monophthongally in the American south, [nat], is a number that may be added to other measures of difference from a large sample. Given this, we may collect many measurements involving many pairs of pronunciations, enabling an aggregate and therefore more abstract characterization of the data. Once we have an aggregate, then we need not be distracted by linguistic exceptions, which simply contribute differently — perhaps more and perhaps less — to the sample. Geographic exceptions may of course remain, but only if the weight of a great deal of linguistic evidence bears this out. The single exceptional feature no longer spoils a characterization, which is based on the tendencies of many features. We avoid the dangers of relying on fortuitously chosen data by including a great deal of data, and abstract characterizations inherit the reliability of the mass of data on which they are based.

2 Categorical Data

The strategy of characterizing dialects on the basis of large aggregates of data samples was pioneered in dialectometry by Jean Séguy and Hans Goebl (Séguy 1973, Goebl 1984), who analyze large samples at a NOMINAL, or CATEGORICAL

level. Categorical data analysis views data as belonging either to the same or to different categories. One may then measure the SIMILARITY between the lexicalizations for the concept DRAGONFLY, i.e., variants such as 'darning needle' vs. 'dragonfly' as zero (0) in case the variants are not the same, and one (1) if they are. Alternatively, one may measure the dissimilarity or DISTANCE between the two variants, in which case the values switch. Henceforth we focus on distances rather than similarities, but we maintain that the two perspectives are interchangeable.

It is also possible to weight some items as more important than other. Goebl (1984), for example, advocates the use of a similarity measure weighted to favor infrequent coincidence, which Nerbonne and Kleiweg (2007:160–163) evaluate positively. We shall ignore such weightings in the remainder of this article, but we note that it is straightforward to apply weightings such as Goebl's to the all the aggregating procedures below.

We also need to deal with competing forms. In the simple case where a single form at one site is compared with two at another, the mean of the two distances is used. Nerbonne and Kleiweg (2003) generalize this idea.

Both Séguy and Goebl include pronunciation differences among the variables they quantify. If the (first vowel of the) word 'marry' is among the pronunciation variables, then one may review the variants to determine whether the pronunciation is [æ], [ɛ], [e], or even [æ^{w}], [æ̲], [e̬] or [ẽ]. We then check straightforwardly whether two pronunciations are the same or not. A large number of such comparisons provides a reliable basis on which to measure varietal similarity. Exact identity may be too demanding a criterion, in which case one needs a procedure testing whether two sounds belong to the same class or not, effectively dealing with the problem of individual or irrelevant variation. In the running example, one might distinguish [æ] and its variants (including [æ^{w}] and [æ̲]) from raised variants such as [ɛ] and [e] (including [e̬] and [ẽ]). Naturally, it is important to defend the classification chosen. But if the classification is sound, then it is good methodology to include this 0/1 measurement as one of a large number on which an aggregate relation is assayed.

Even though their methodology is sound, we sought to go beyond the Séguy/Goebl categorical level of analysis for pronunciation differences, first, because it requires manual intervention in isolating specific aspects of pronunciation, where an automated process is possible. Second, we wished to make fuller use of the rich dialect atlas transcripts, rather than be limited to just selected variables.

3 String Distance Measure

Fortunately, there are effective algorithms available for comparing strings, i.e. sequences of symbols. Gusfield (1999: Ch.11) is an excellent summary of the current state of the art. Rather than try to summarize all the techniques,

we present EDIT DISTANCE here, also known as LEVENSHTEIN DISTANCE and STRING (EDIT) DISTANCE. We can approach edit distance from two perspectives, and it will be instructive to use both in this overview.

One the one hand, one may ask how many operations of a simple sort are required to transform one string into another. We illustrate how one dialectal pronunciation of German *Durst* 'thirst', namely [tʊəʃ] (Aachen) is transformed into another, [tɔʃt] (Vielbrunn). By writing each derivation step, we can see the operations at work:

t	ʊ	ə	ʃ		
t	ɔ	ə	ʃ		substitute [ɔ] for [ʊ]
t	ɔ		ʃ		delete [ə]
t	ɔ		ʃ	t	insert [t]

If each operation is associated with a cost, e.g. one, then the Levenshtein distance is the sum of the least costly set of operations mapping one string to another. Since the three operations above are indeed minimal, the distance between the two strings is the sum of the cost of the operations, three. This is naturally rough; we examine more sensitive costs in Sec. 3.3 below. Gusfield (1999:11.1–11.3) presents a dynamic programming algorithm for calculating Levenshtein distance efficiently.

The alternative perspective is that of alignment. Proceeding from the least costly derivation, we align identical segments that remain constant during the derivation and all pairs of segments where one was substituted for the other. Finally we align symbols with an empty segment in case they are involved in deletions or insertions. The result in this case is as follows.

	t	ʊ	ə	ʃ	
	t	ɔ		ʃ	t
Costs		1	1		1

Given a derivation, the alignment may be recovered. Alignments are an important check on the quality of measurements (see Sec. 5), and we may search the aligned segment pairs (above [ʊ/ɔ],[ə/] and [/t]), for regular correspondences (see Kondrak 2002, Prokic 2007).

The procedure is normally applied to the entire dialect material available. We measure the differences not only between all pairs of transcriptions for Durst 'thirst', but also those of dozens to hundreds more. It is sensible to use material which represents true pronunciation differences as purely as possible, and which therefore contains as few differences as possible due to inflectional morphology, sandhi, and intonation. After determining all of the pronunciation distances for all of the words, the pronunciation distance between sites is simply the mean pronunciation distance of all the words in the atlas's sample. These site × site differences may be further analyzed using hierarchical clustering or multidimensional scaling (see Nerbonne, Heeringa and Kleiweg 1999, Heeringa 2004, or below for examples). Peter Kleiweg maintains an interactive demo at www.let.rug.nl/kleiweg/lev/ which includes various operation weights as well as the opportunity to use segment distances derived from features.

We turn now to the linguistic interpretation of this sort of measurement. We note first that the measurement may be automated so as to require no manual intervention (see R*u*G/L[04], www.let.rug.nl/kleiweg/L04/), and second that it is sensitive to a large number of pronunciation differences, not merely to a limited set chosen by the investigator.

Third, however tempting it might be to interpret the insertions, deletions and substitutions as historical sound changes, we warn that the operations work strictly on the surface. Alignments provide documentation of sound correspondences, but the procedure is not designed to discover or simulate historical changes.

Fourth, even though there is a unique least cost of operations mapping from one string to another, there may be different sets of operations that corresponds to different alignments. This is one reason we are motivated to explore more sensitive measures of phonetic overlap (see Sec. 3.3 below).

Fifth and finally, there are many potential refinements of the procedure which might be expected to yield more precise measurements. In 3.2 we discuss how one might discount the effects of fast speech rules, and in 3.3 how to build more phonetic sensitivity into the measure, e.g., by weighting the substitutions phonetically. In 3.5 we consider enforcing a syllabicity constraint, making the calculation sensitive to context, and including metathesis as a basic operation.

3.1 Formal properties

We note some formal properties of edit distance, reviewing material in Kruskal (1999) and Gusfield (1999). First, edit distance is a generalization of Hamming distance, which makes no provision for insertions and deletions. We illustrate Hamming distance below:

	t	ʊ	ə	ʃ
	t	ɔ	ʃ	t
Costs		1	1	1

Even though the distance happens to be the same in this case, it is clear that the Hamming procedure misses the coincidence of [ʃ] in the two strings, so that it is less suitable for dialectological application. The psycholinguistics of spoken word recognition is extremely sophisticated in some aspects, but it uses Hamming distance — on carefully controlled material in which insertions and deletions do not occur — as a measure of word similarity (Luce and Pisoni 1998). See also Sec. 6.

There are even less sophisticated measures, such as Jaccard distance and Dice, which ignore the order of segments (Manning, Prabhakar and Schütze, 2008).

Second, we note that Levenshtein measures are distances in the mathematical sense, i.e. numbers greater than or equal to zero; zero just in case two strings are the same; symmetric, so that the distance from string s1 to s2 is always the same as the distance from s2 to s1; and in conformance with the so-called triangle inequality: the distance from s1 to s2 is less than or equal to the sum of the

distances from s1 to s3 and s3 to s2, for any string s3. The fact that the distances are symmetric makes them unsuitable for modeling some problems, for example, the CONFUSABILITY MATRICES phoneticians compile from how frequently one segment may be mistaken for another (Johnson, 2004, Chap.4). If one modeled confusability by similarity, implemented by edit distance on samples of the segments' spectrograms, the attempt would be limited by the inherent symmetry of edit distances.

3.2 Herrgen-Schmidt Dialecticality Measures

Independent of the work on Levenshtein distance, Herrgen and Schmidt (1989) sketched a procedure for comparing pronunciations of entire words which has been applied in several projects and publications. Our presentation follows Lameli (2004: Chap.5).

The Herrgen-Schmitt procedure shares the motivation to evaluate a large sample of comparable pronunciations, and to include all the material in each word transcription rather than only selected differences. The procedure assumes feature-based descriptions of each segment, which are interpreted numerically. For example if there are four vowel heights, corresponding to [i], [e], [ɛ] and [a], then the height difference between [i] and [e] is one, and that between [i] and [a] three. It is challenging to specify a segment distance table exhaustively (see Sec. 3.3 below), but Lameli completes the table successfully. The segment distance table, thus derived, is then used as the basis for calculating word differences, which are simply the sum of differences in aligned segments.

There are several unique aspects of this research line. First, the work focuses less on measuring dialect differences, more on measuring differences between a dialect on one hand and the standard language on the other. This reflects the interests of the researchers, but it also allows them to incorporate a second unique feature, a normalization with respect to fast-speech rules. The Herrgen-Schmidt procedure does not regard differences as genuine if they might have arisen through the application of a fast speech rule. For example German *Lippen* 'lips' may have a canonical standard pronunciation as ['lɪp.ən], but it is pronounced in fast speech as ['lɪp.m̩]. Dialect pronunciations which elide the schwa and assimilate the final [n] should not be measured as differing from standard German. In unpublished work at Groningen we have experimented with implementing several fast-speech rules in standard German, measuring the distance of a dialectal form to each "allegro" variant, and then using the least value as the distance.

Third, Herrgen-Schmidt's rules are sensitive to small differences, but they also set maximal segment difference values so that word measurements are not overwhelmed by single-segment differences. Fourth, there is still no automatic procedure for applying Herrgen-Schmidt's difference metric. This probably derives from detailed and complex rules for handling some special cases such as the German /a/ and foreign borrowings. This overview is too brief to review all of the details.

3.3 Segment Distances

It is linguistically natural to wish to incorporate more phonetic sensitivity into the string distance measure. After all, an English dialect with 'path] as [paθ] is closer to one with [pæθ] than to a third with [pɛθ]. There is also a technical reason for preferring more sensitive measures of segment difference, namely the wish to avoid multiple alignments. The more sensitive the phonetic measure, the less likely these are.

Accordingly, not only Herrgen-Schmidt (above), but also Kessler (1995), Nerbonne and Heeringa (1997), Kondrak (2002, Chap. 4.5) Heeringa (2004, Chap.3–4), McMahon, Heggarty, McMahon and Maguire (2007) and others have proceeded from segment distance tables, and, for the most part, have used the Levenshtein procedure sketched above. Heeringa (2004) devotes ninety pages to comparing segment distances derived from the feature system in *The Sound Pattern of English* (Chomsky and Halle, 1968), two different systems developed to score transcription quality, and, a fourth system Heeringa develops based on curve distance in canonical spectrograms. All of the systems allow for the representation of diphthongs, affricates, stress, length, syllabicity, and a wide range of secondary articulations. In the same spirit as the Herrgen-Schmidt maximal values for segment distances, Heeringa uses a logarithmic correction to limit the impact of differing segments, following Stevens's (1975) idea that psychophysical reactions scale logarithmically.

Heeringa found modestly superior analyses using spectrogram-based segment distance tables. See Sec. 5 and 6 below on comparing putative improvements, and see the appendix for one comparison of analyses based on phonetic features vs. the binary same/different distinction.

We note further that, given a table of segment distances, however derived (called ALPHABETIC WEIGHTS in Gusfield, 1999), the dynamic programming algorithm computing Levenshtein distance always returns the optimal alignment with respect to that table (Kruskal 1999: Sec. 5, Gusfield 1999: 11.5), i.e. the alignment that minimizes the sum of the aligned segments' differences. We therefore say that the Levenshtein procedures "lifts" the segment distance table to a sequence distance measure. Heggarty et al. (2007) aim to align words with respect to etymological frames, which would seem to require moving from two-string alignments to the simultaneous alignment of three strings, but their procedure is not explained in detail (see Gusfield, 1999, Ch.14 on (difficult!) multiple string alignment).

3.4 Other Related Work

Kessler (1995) first applied edit distance to dialect material. Like Herrgen-Schmidt he proceeded from a segment distance table, which he also compared to a binary scheme in which segments were either alike or different, concluding that the latter was superior. Kessler applied clustering to check whether the edit distances could delineate dialect areas. Nerbonne, Heeringa and Klei-

weg (1999) followed with an analysis of Dutch and introduce multi-dimensional scaling (MDS) as a means of further analyzing the average pronunciation differences. Both clustering and MDS are illustrated in Sec. 4 below. Heeringa (2004) analyzed Norwegian and Dutch, comparing many options for phonetic representation (see Sec. 3.3 above), clustering and MDS. It is the most thorough treatment of the subject to-date. There are also analyses of German, Bulgarian, American English, Italian, Sardinian, Bantu, and some Indo-Iranian and Turkic varieties.

Kondrak (2002) modified edit distance to discount mismatches near the beginnings and ends of words and applied his algorithm to diachronic phonology, experimenting on Indo-European and Algonquian languages. By keeping track of frequent operations, he could detect regular sound correspondences and cognates. McMahon, Heggarty, McMahon and Maguire (2007) apply an alignment algorithm to English dialects and analyze results using phylogenetic algorithms designed to infer genealogical trees (or networks).

3.5 Other refinements

The success of the techniques raises questions about its linguistic underpinnings, some of which we explore here. Early on, inspection of the alignments induced by the dynamic programming algorithm revealed alignments such as the following two:

t	ʊ	ə	ʃ		t	ʊ	ə	ʃ
t	ɔ	ʃ	t		t	ɔ	ʃ	t

Each of these requires three operations, yielding distances of three, but intuitively the first violates a SYLLABICITY constraint by allowing the consonant [ʃ] to replace the vowel [ə]. It is straightforward to enforce this constraint, and Heeringa, Kleiweg, Gooskens and Nerbonne (2006) claim that this yields superior analyses.

While almost all applications have ignored the context of sound correspondences, Heeringa et al. (2006) operationalize context by applying the algorithm not to single segments, but rather to bigrams, a standard technique for incorporating context in computational systems. Results were slightly, but consistently better.

There are also extensions of the Levenshtein procedure allowing metatheses (also known as 'swaps' or 'transpositions'). See Sankoff & Kruskal (1999:9ff,212). Metatheses are rare in languages analyzed to-date, with Bulgarian a notable exception. Adding swaps to the dynamic programming algorithm can be quite difficult when segment distance tables are used (Wagner, 1999).

It has been straightforward to interpret STRESS as a vowel feature. Under this scheme the verb 'contract' [kən.ˈtrækt] and the noun [ˈkɔn.trækt] differ at two positions, the vowels. Without special treatment, the words would have differed in the first vowel, but not in the second. Because TONE is at times not realized on single segments, but rather in complicated ways that depend on the sequence of syllables, it is less straightforward to incorporate tone into the

distance measure in a general way. Gooskens and Heeringa (2006) contrast the three tone patterns of Norwegian and compare the degree to which perceptual distances can be predicted by prosodic differences as opposed to by segmental phonological differences measured using edit distance.

Finally, we should like to add that, although it is good scientific practice to prepare one's data carefully to eliminate potential confounds and "noise", it is not always practical. Fortunately, the procedures we sketch are robust enough to function well even on noisy data. The Dutch dialect source used by Heeringa (2004) was analyzed in two ways, once restricted to cognate words (allowing morphological differences, cf. 'dog' and 'doggy'), and once comparing all semantically similar words, including unrelated words (cf. 'friend' and 'buddy'). Given a 125 word sample, results correlate nearly perfectly ($r = 0.98$). The replication on a Norwegian sample was slightly lower ($r = 0.95$, $n = 58$).

We consider other potential refinements in Section 6 below, "Emerging Issues".

4 Example

The *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) comprises material collected on the Eastern seaboard of the U.S. from 1933 through 1974. The area extends from Northern Florida through New York state and includes all the intermediate states on the Atlantic, plus West Virginia. Our focus here is on the data collected by Guy Lowman in 1933–1936, roughly 70% of the total (see Fig.1 for a map). We focus only on Lowman's work to avoid confounds in field workers' techniques and/or transcriptions (Nerbonne and Kleiweg, 2003).

The data were obtained using a questionnaire in which respondents were asked how they expressed everyday things and events, e.g., "If the sun comes out after a rain, you say the weather is doing what?" (used to elicit 'clearing up', 'fairing off' and forty other dialectal variants. The LAMSAS material is publicly accessible for reanalysis (see http://us.english.uga.edu/lamsas/; Kretzschmar, 1994) and contains the responses of 1162 informants interviewed in 483 communities. The responses to 151 items are included in the web distribution, the basis for the work here. We analyze Lowman's data in what follows.

In Lowman's section of the database there are 92,537 transcriptions involving 1.3 million phonetic tokens collected from 363 locations, and 797 informants. There are, on average, 14.0 characters per string, which are parsed into 7.9 sound tokens (IPA segments, often with diacritic) per string. There are 1,677 unique sounds (combinations of base segments and various diacritics) and 1,132 unique vowel sounds alone. LAMSAS usefully contains both phonetic and orthographic transcriptions, allowing us to focus measurements on comparable material, ignoring lexical and morphological differences.

We implemented the LAMSAS feature system as a segmental basis (but see the Appendix for a comparison of the results based on a binary same/different

distinction). We present the features for vowel system in the table below, and suppress discussion of the consonantal features in the interest of saving space. If one chooses to adopt an existing feature system, as we did, there arises the question of how to interpret the feature values numerically. Thus Kurath and Lowman distinguished ten different vowel advancement positions, and fifteen vowel heights, but we need to assign numbers to the positions if we are to use them to measure segment differences. The table below shows how we did this, allowing six as a maximum advancement difference and only three and a half as a maximum height difference. Wherever the values of one feature may differ more than those of another, the scheme in the table effectively weights the first feature more heavily. As just noted, advancement may differ by as much as 6, while rounding can not differ by more than one, so differences in advancement can count much more. Diacritics representing stress, rhotism, pharyngealization and devoicing were each capable of adding maximally one unit of difference, and intermediate differences, including those indicated by diacritics, were interpolated. The differential weightings of the features are given implicitly by the difference in feature's extreme values. The decision to make advancement count more heavily than height is based on the idea that a change in vowel advancement marks a dialect speaker more saliently than a difference in vowel height, but we concede that this decision is moot.

One vowel feature, 'direction' was not instantiated anywhere in the LAMSAS database, so it is omitted from the table below, and Lowman and Kurath allow for six degrees of rounding, which we simplified to five when we found only five levels distinguished in the data.

Vowel Feature	Possible Values
v-advanced	-3, -2, -1, 0, 0.4, 1, 1.4, 2, 2.4, 3
v-high	-1.75, -1.5, -1.25, -1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75
v-rounded	-1, -0.5, 0, 0.5, 1
v-long	-0.5, 0, 0.5, 1
v-stress	0, 0.35, 0.7
v-nasal	0, 1
v-rhotic	0, 1
v-pharyng.	0, 1
v-voice	0, 1

The feature names reflect their normal phonetic (articulatory) interpretation. The stress which is marked on a syllable is interpreted as a property of the vowel, which is why it appears in the Table above. Vowels receive either stress, secondary stress, or no stress. Vowels were interpreted as voiced except when explicitly marked as voiceless, in which case they bore the feature [-voice]. Lowman rarely added a diacritic indicating the pharyngealization of a vowel, which is interpreted by [v-pharyng.] feature where it occurs. Vowels written as superscripts (e.g., the second parts of laxing diphthongs) are not interpreted by

a feature [±super] — but rather through a weighting. Comparisons involving superscripted vowels count only 50% of what they would cost if the segments were not superscripted. The idea is naturally that such minor articulations contribute less to pronunciation difference.

We calculate the distance between two vowels first by simply summing the differences of all the feature values, $\sum_f |f_v - f_{v'}|$. In order to emphasize the importance of slight differences as opposed to larger ones, we work with the logarithm of that sum, as introduced above. Finally, we wish to work with a scale with a genuine zero:

$$d(v, v') = \log(1 + \sum_f |f_v - f_{v'}|)$$

We applied the unigram Levenshtein model using the segment distances just sketched, without any word-length normalization. We compared the pronunciations of each pair of sites in Lowman's data, using all of the words common to each pair. The result is a large pronunciation distance table, or MATRIX, providing a mean pronunciation distance for each pair of sites. As distances are symmetric, we ignore half of the cells in the table.

We turn now to the results, beginning with the dialect areas which emerge from the pronunciation comparison. They are detected via hierarchical clustering (Nerbonne, Kleiweg, Heeringa and Manni, 2008), a technique for recognizing groups in data, and are shown in Fig. 1.

Fig. 1 shows a map of the entire LAMSAS area, where the areas covered by field workers other than Lowman are shaded gray. Even though the clustering procedure worked only on a matrix of pronunciation distances, note that it detects geographically coherent areas, a first indication that the analysis is working. The dendrogram on the right shows that the major division in the data is between the two most southern subareas and the four northern ones, where the split runs to the south of the (northern) Virginia border; in the northern cluster, there is a second north-south split running through the north of Pennsylvania. This is a defensible subdivision of the LAMSAS speech area, even if it differs from Kurath and Lowman's opinion in failing to confirm their "midland" area, extending from western Pennsylvania southward along the Appalachian mountains.

Because the pronunciation distance is numerical, we can also apply multidimensional scaling (MDS) to our results to attempt to view it in a more simplified form. The result is found in Fig. 2. Heeringa (2004: Ch.6.2) explains MDS in more detail. We note only that MDS tries to place each site in a coordinate space of few dimensions. We use a three-dimensional solution which accounts for over 90% of the variation in the original distance table. If we map each of the dimensions in the three-dimensional solution to intensities of red, green and respectively blue, we obtain the map in Fig. 2.

The MDS analysis is striking for the prominently distinguished position of southeast Pennsylvania. The explanation for the area's unusual status is sug-

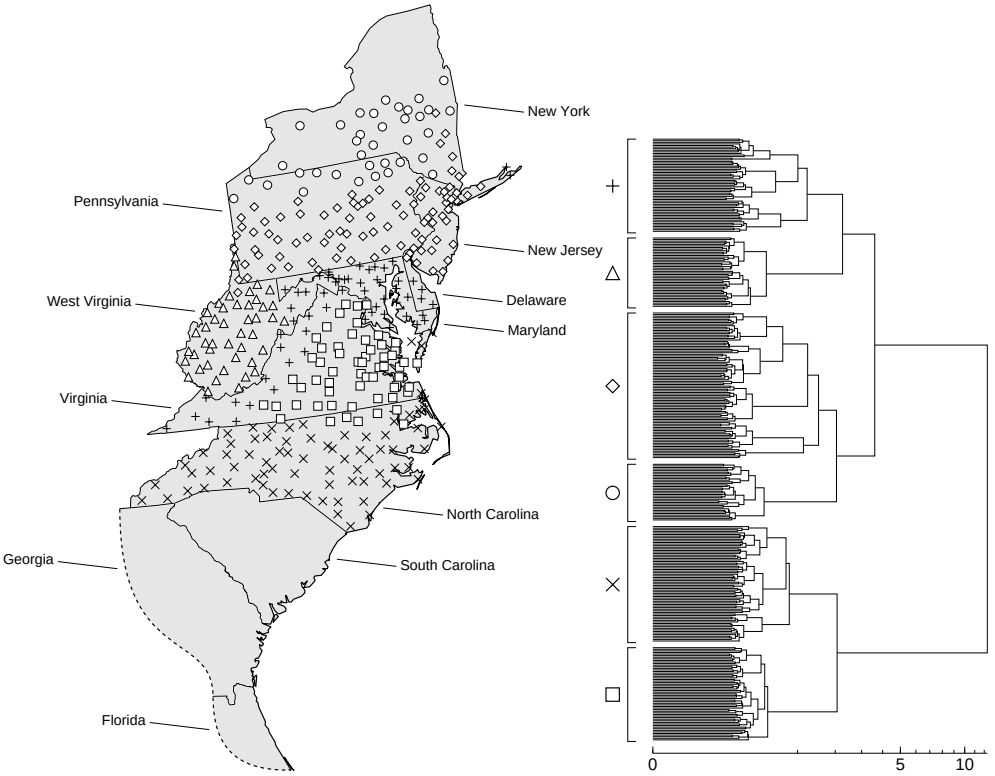

Figure 1: Left, the dialect areas emerging from the aggregated pronunciation difference measurements and right, the dendrogram, providing a key. See text for discussion.

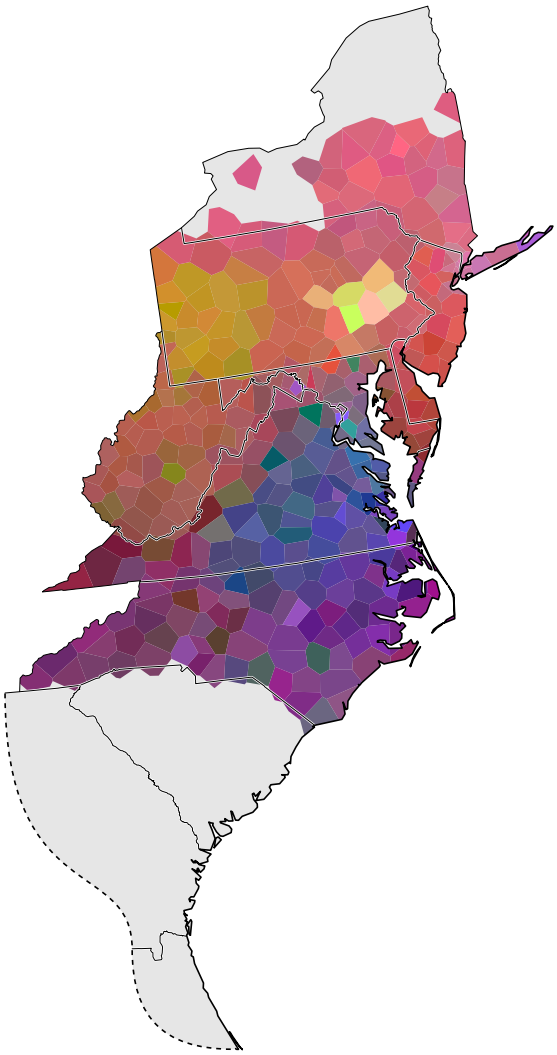

Figure 2: The three-dimensional MDS analysis of the pronunciation distances, rather better suited for portraying dialect continua. See text for discussion.

gested in pronunciations such as ['tʃɔr.tʃə] 'Georgia', where the devoiced syllable onsets (expected [ʤ]) betray the German influence on the speakers interviewed (speaker PA7C in Lancaster in fact spoke German as his first language).

Nerbonne (2006) traces which linguistic features contribute the most to the aggregate dialectal differences shown in the map, but focused on the four southernmost areas shown in Fig. 1. We focus on the most important distinctions here. Using factor analysis it could be shown that the different reductions of unstressed vowels, [ə] vs. [ɪ] (the latter including [ɨ]) in the unstressed closed syllables of words such as 'closet', 'kitchen', and 'Baltimore' (second syllable) are most strongly coherent among the collection sites. Closely aligned with this shift we noted that the same varieties which used the higher version of the reduced vowel ([ɪ]) also fronted the [u] to [ʉ] in words such as 'St. Louis' and 'Tuesday', likewise fronting the onset of the diphthong in the second syllable in 'Missouri'. Finally the [ɔ/ɑ] distinction in 'hog[pen]' and 'Florida' aligned well with [ə/ɪ] distinction. Note that only the [ɔ/ɑ] distinction is phonemic in some varieties of American English, while the other two are subphonemic. The paper also examines several less prominent alternations, include (i) rhotic vs. nonrhotic pronunciations, where the latter varieties likewise demonstrate a lowering of [ɔ] in words such as 'forty' or 'storm'; (ii) a contrast between raised and unraised [ɨ] as the last syllable in 'Tuseday', 'foggy' and 'thirty'; (iii) another [ə/ɪ] distinction, but this time in open syllables such as the final syllable in 'sofa', 'Georgia' and 'Russia'; (iv) the raising of [ɛ], most extremely to [ɪ], e.g. in the first syllable in 'Tennessee' (in general before [n]); and (v) fronted vs. nonfronted versions of the lax [ʊ] in words such as 'wood' and 'good'. It is striking that the [aɪ/a] shibboleth was not among the most prominent distinctions.

Kurath & McDavid (1961) also discuss each of the features. The aggregate, dialectometric perspective adds the opportunity to quantify the importance of features, which Kurath and McDavid viewed cartographically.

5 Evaluation

Above we criticized the traditional method for having little to say about which features and which features' distributions are important. This arises in traditional methodology because there are simply too many features, and thus too many distributions and isoglosses to choose from, leaving the method quite underdetermined. This problem arises in the evaluation of older methodology.

But it would be naïve to think that we make ourselves immune to methodological cares simply by aggregating. The problem of which features to choose we indeed avoid by obtaining a large, representative sample, preferably via the (standard) random selection (Bolognesi and Heeringa, 2005). But there remain many options in measuring pronunciation differences (see 3.3 and 3.5 above). Defining segment distances alone involves distinguishing about twenty features and five or so values per feature. In addition we may process unigrams or bi-

grams, to treat diphthongs and/or affricates as one segment or two, to normalize based on word length or based on segment or word frequency, to aggregate feature differences via addition or via a Euclidean distance, ... The list is substantial, leading to the question: If we experiment extensively, and finally hit upon a pleasing analysis, have we attained substantial insights or have we been lucky due to the plethora of analytical options? This is our evaluation problem.

As we suggest by introducing the issue this way, we believe that the evaluation problem has always been present in dialectology, but it is more obviously present in complex systems involving measurements where many options are available to analysts. In the present discussion we consider only the problem of evaluating the pronunciation measurements, and ignore issues concerning further analysis, in particular those involving clustering.

Psychometrics divides these issues are into two sub-issues, CONSISTENCY and VALIDITY. Consistent measurements tend to provide the same information. We are encouraged when measurements continue to function well when used on new data, a simple sort of consistency check. Cronbach's alpha is a more formal check on consistency (Nunnally, 1978). We first measure the inter-item correlation, e.g., between the word 'dog' and the word 'cat' by first obtaining two site $\times$ site tables of pronunciation distances, one for 'dog' and another for 'cat'. We then calculate the correlation between the two tables, measuring, informally, the degree to which the two words provide the same indication of linguistic distance. We repeat this process for every pair of words in the sample (a time-consuming task), thereby obtaining the mean inter-item correlation. Fig. 3 below graphs Cronbach's alpha as a function of r, the mean inter-item correlation coefficient, and n, the number of items. Given a mean inter-item correlation, Cronbach's alpha indicates whether enough material is present in the sample. Nunnally (1978) suggests that 0.7 or 0.8 is a satisfactory level, and Cronbach's alpha = 0.97 for the LAMSAS sample, indicating that the signal is very consistent, given the amount of material.

We emphasize that Cronbach's alpha depends on the data set being analyzed. In our experience 30-word sets normally show levels above 0.8, but new data, especially with low inter-item correlation, could differ.

We say that a measure is VALID if it measures what it purports to, and this leads immediately to more reflective questions about the goals of dialectology. We can claim to measure similarity in pronunciation, but similarity with respect to what? Our thinking has evolved on this. Assuming that expert opinion is well-founded, we once calculated the degree to which our analyses coincided (Heeringa, Nerbonne and Kleiweg, 2002). While this is worthwhile, still, if we are ambitious, we should like to improve on earlier expert opinion if possible. We now prefer to begin from the premise that one of the goals of dialectology is to *characterize the signals of provenance* normally present in speech. Naturally this begins with detecting those signals and then proceeds to investigate their structure. We speak of "signals" to emphasize that people should be able to receive them, and this leads immediately to the idea that we should validate

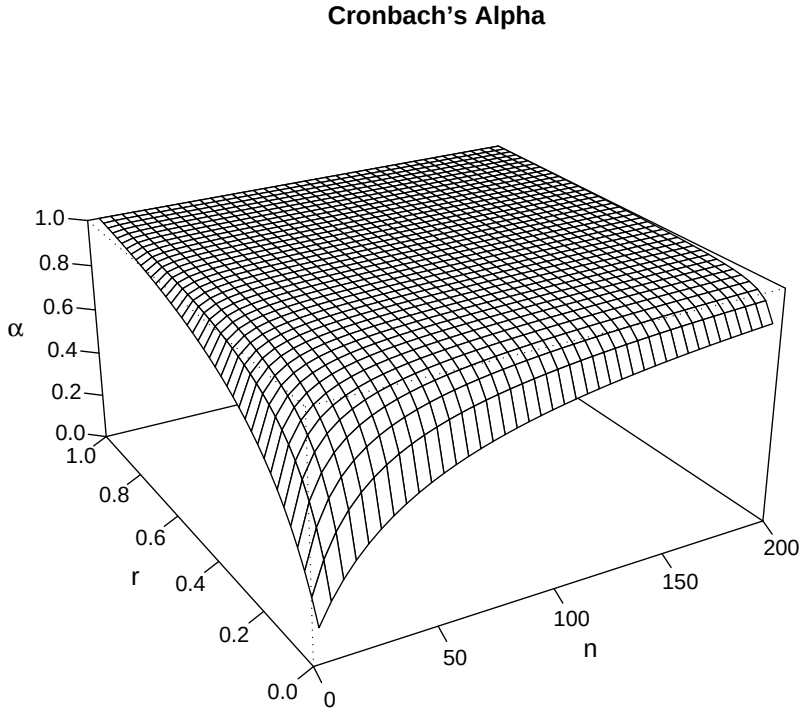

Figure 3: Cronbach's alpha is a function of the inter-item correlation r (left axis), which measures whether different items are giving the same indication, and n (right axis), the sample size. It shows whether samples are large enough to provide consistent signals.

our measurements by comparing them to dialect speakers' judgments.

Gooskens and Heeringa (2004) have developed this idea and made it operational. They analyzed Norwegian pronunciations along the lines sketched in Sec. 3 above, and they played recordings of the pronunciations to dialect speakers, who judged how similar the pronunciations were to their own. Gooskens and Heeringa then measured the correlation between the perceptual judgments and pronunciation distances ($r = 0.7$). We have also applied this in comparing different versions of the pronunciation difference measure (Heeringa et al. 2006), and it seems the best way to validate the work independently. We should add that most analyses indicate that the various different versions of the pronunciation measure do not differ significantly when evaluated strictly. It is surprisingly difficult to demonstrate the superiority of the more discriminating systems.

6 Emerging Issues

If we are satisfied that pronunciation distance measures contribute valuably to the dialectologist's toolbox, then several opportunities arise, and several further tasks suggest themselves. Perhaps the most exciting opportunity is the chance to attempt to characterize the general relation between geographic and linguistic distance. Séguy (1971) shows that dialectometric distances grow sublinearly as a function of geography, a result which most studies have confirmed to-date. Nerbonne and Heeringa (2008) replicate Séguy's finding using Dutch data and argue that this contradicts Trudgill's well-known "gravity model". The present overview is too brief to go into further detail, but our argument made essential use of the measurement techniques presented here.

The success of the technique suggests that it should be possible to detect recent borrowings as those words that show an unexpectedly low distance given overall varietal distances, and it is exciting to consider what other cultural markers might be studied quantitatively to understand the degree to which linguistic variation behaves like other cultural traits. Finally, we wish to know the correlation of different linguistic levels — lexical, phonological and syntactic — as signals of provenance.

There are also several points at which we feel the techniques presented here should be improved. It should be clear that we are not satisfied with the minor benefits that have been adduced for systems with segment distances. Linguistically we are quite certain that finer distinctions may be reliably drawn, and it is a puzzle as to why their benefit should be so hard to demonstration. Our current hypothesis is that the high level of aggregation may be hiding the benefits. It would also be sensible to explore the relation between dialect perception and other aspects of phonetics, asking whether and which traits they share (Wieling and Nerbonne, 2007). Finally, we look forward to the development of better techniques for identifying the major linguistic factors in aggregate comparison.

References

Alewijnse, Bart, John Nerbonne, Lolke van der Veen, and Franz Manni 2007 A Computational Analysis of Gabon Varieties. In: Petya Osenova et al. (ed.) *Proceedings of the RANLP Workshop on Computational Phonology Workshop at the conference Recent Advances in Natural Language Processing.* Borovetz. 3–12.

Bloomfield, Leonard 1933 *Language.* New York: Holt, Winehart and Winston.

Bolognesi, Roberto and Wilbert Heeringa 2005 *Sardegna fra tante linge. Il contatto linguistico in Sardgna dal Medioevo a oggi.* Cagliari: Condaghes.

Chomsky, Noam and Morris Halle 1968 *The Sound Pattern of English.* New York: Harper and Row.

Goebl, Hans 1984 *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF.* 3 Vol. Tübingen: Max Niemeyer.

Gooskens, Charlotte and Wilbert Heeringa 2004 Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16(3), 189–207.

Gooskens, Charlotte and Wilbert Heeringa 2006 The Relative Contribution of Pronunciation, Lexical and Prosodic Differences to the Perceived Distances between Norwegian dialects. *Literary and Linguistic Computing* 21(4): 477–492.

Gusfield, Dan 1999 Algorithms on Strings, Trees and Sequences. Computational Science and Computational Biology. Cambridge: Cambridge University Press. (11997)

Heeringa, Wilbert 2004 Measuring Dialect Pronunciation Difference Using Levenshtein Distance. Ph.D. Diss., University of Groningen.

Heeringa, Wilbert, John Nerbonne and Peter Kleiweg 2002 Validating Dialect Comparison Methods. In: Wolfgang Gaul and Gerd Ritter (eds.) *Classification, Automation, and New Media. Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation.* Heidelberg: Springer. 445–452

Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens and John Nerbonne 2006 Evaluation of String Distance Algorithms for Dialectology. In: John Nerbonne and Erhard Hinrichs (eds.), *Linguistic Distances,* ACL Workshop held at ACL/COLING, Sydney, 51–62. Shroudsburg, PA: ACL.

Herrgen, Joachim and Hans-Jürgen Schmidt 1989 *Dialektalitätsareale und Dialektabbau. In Wolfgang Putschke, Werner Veith & Peter Wiesinger (eds.) Dialektgeographie und Dialektologie. Günter Bellman zum 60en Geburtstag von seinen Schülern und Freunden.* Marburg: Elwert. 304–346. (Deutsche Dialektgeographie, Vol.90)

Johnson, Keith 2004 *Acoustic and Auditory Phonetics.* London: Blackwell (1st ed., 1997).

Kondrak, Grzegorz 2002 Algorithms for Language Reconstruction. Ph.D. Diss., University of Toronto.

König, Werner 1994 *dtv-Atlas zur deutschen Sprache.* München: Deutscher Taschenbuchverlag. (1st ed. 1978).

Kretzschmar, William, Jr. 1994 *Handbook of the Linguistic Atlas of the Middle and South Atlantic States.* Chicago: University of Chicago Press.

Kruskal, Joseph 1999 An Overview of Sequence Comparison. In: David Sankoff and Joseph Kruskal (eds.) 1–44.

Kurath, Hans and Raven McDavid 1961 *The Pronunciation of English in the Atlantic States: Based upon the Collections of the Linguistic Atlas of the Eastern United States.* Ann Arbor: University of Michigan Press.

Lameli, Alfred 2004 *Standard und Substandard. Regionalismen im diachronen Längschnitt.* Wiesbaden: Franz Steiner Verlag. (ZDL Beiheft 128)

Luce, Paul A., and David B. Pisoni 1998 Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing* 19(1):1–36.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze 2008 *Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

McMahon, April, Paul Heggarty, Robert McMahon and Warren Maguire 2007 The sound patterns of Englishes: Representing phonetic similarity. *English Language and Linguistics* 11.1 113–142.

Nerbonne, John 2006 Identifying Linguistic Structure in Aggregate Comparison. *Literary and Linguistic Computing* 21(4). 463–476.

Nerbonne, John and Wilbert Heeringa 1997 Measuring Dialect Distance Phonetically In: John Coleman (ed.) *Workshop on Computational Phonology.* Special Interest Group of the Association for Computational Linguistics. Madrid, 11–18.

John Nerbonne and Wilbert Heeringa 2007 Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation. In: Sam Featherston and Wolfgang Sternefeld (eds.) Roots: Linguistics in Search of its Evidential Base Berlin: Mouton De Gruyter. 267–297

Nerbonne, John, Wilbert Heeringa and Peter Kleiweg 1999 Edit Distance and Dialect Proximity. In: David Sankoff and Joseph Kruskal (eds.), v–xv.

Nerbonne, John and Peter Kleiweg 2003 Lexical Distance in LAMSAS. *Computers and the Humanities* 37(3), 339–357.

Nerbonne, John and Peter Kleiweg 2007 Toward a Dialectological Yardstick. *Journal of Quantitative Linguistics.* 14(2–3):148–166.

John Nerbonne, Peter Kleiweg, Wilbert Heeringa and Franz Manni 2008 Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering. In: Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt & Reinhold Decker (eds.) *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society.* Berlin: Springer. 647–654. (*Studies in Classification, Data Analysis, and Knowledge Organization*)

Nunnally, Jum C. 1978 *Psychometric Theory.* New York: McGraw-Hill.

Prokić, Jelena 2007 Identifying Linguistic Structure in a Quantitative Analysis

of Dialect Pronunciation. Proceedings of the ACL 2007 Student Research Workshop, Prague. 61–66.

Sankoff, David and Joseph Kruskal (eds.) 1999 *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison.* Stanford: CSLI. (1st ed., 1983)

Séguy, Jean 1971 La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35: 335–357.

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

Stevens, S. Smith 1975 *Psychophysics: Introduction to its Perceptual, Neural and Social Prospects.* New York: John Wiley.

Vieregge, Wilhelm H., Rietveld, Antonius C. M., and Jansen, C. I. E. 1984 A distinctive feature based system for the evaluation of segmental transcription in Dutch. In Van den Broecke, M. P. R. and Cohen, A., editors, *Proceedings of the* 10^{th} *International Congress of Phonetic Sciences,* pages 654–659, Dordrecht: Foris.

Wagner, Robert A. 1999 On the complexity of the string-to-string correction problem. In David Sankoff & Joseph Kruskal (eds.), 215–235.

Wieling, Martijn and John Nerbonne 2007 Dialect Pronunciation Comparison and Spoken Word Recognition In Petya Osenova et al. (eds.) *Proceedings of the RANLP Workshop on Computational Phonology, workshop at Recent Advances in Natural Language Processing,* Borovetz, 71–78.

Appendix

In this appendix we compare the feature-based measurements in Sec. 4 to measurements based on the binary same/different distinction. In both the feature-based and the binary system we set the distance between vowels and consonants to be prohibitively high, effectively enforcing the syllabicity constraint noted in the text.

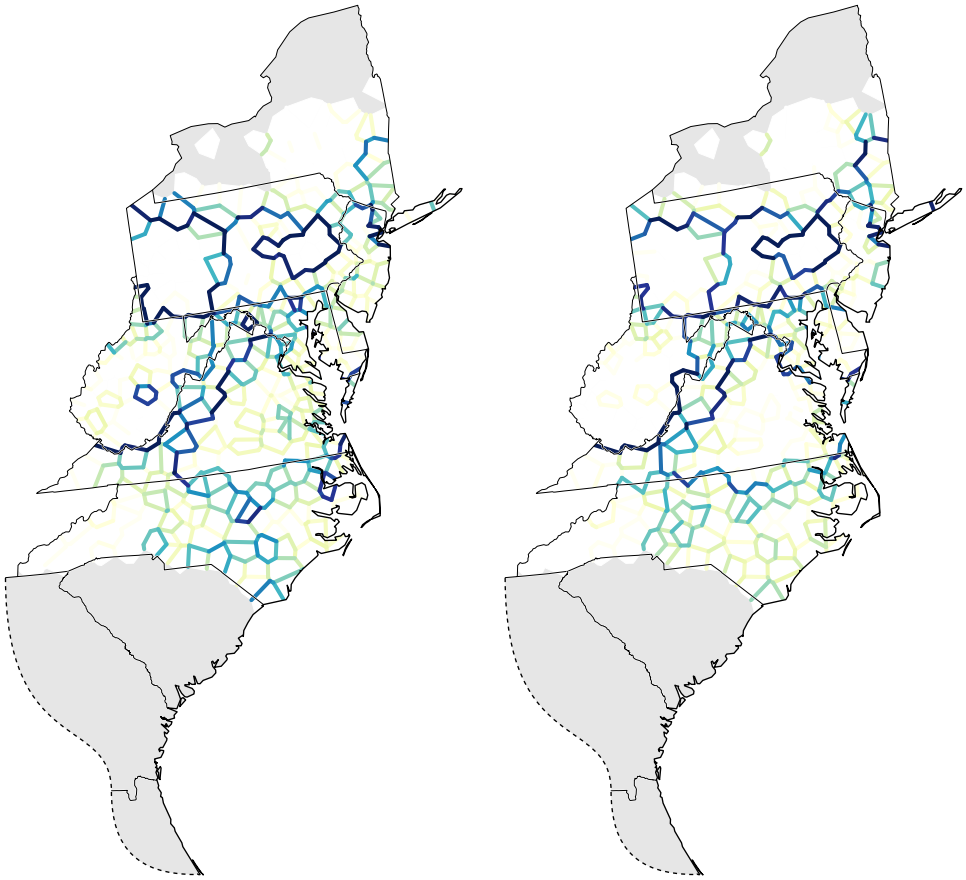

Figure 4: Maps contrasting the measures based on features (left) with the maps based on binary segment distances (but respecting V/C distinction on right.