Gabmap Tutorial

by

Therese Leinonen

Preface

Gabmap is a web application for dialectometric research developed at the University of Groningen. The link to the web application is http://www.gabmap.nl/. This tutorial gives an introduction to and overview of the basic functions in Gabmap.

The tutorial includes a number of exercises marked with this symbol:  .
Solutions to the exercises are collected in the last chapter of the tutorial.

Contents

1 Introduction

Gabmap is a web application for dialectometrics and cartography. It allows you to make mappings and statistical analyses of your dialect data. For example, you collected the local variants of the pronunciation of a large number of words in several locations in some area. Using Gabmap you can compare the pronunciation of these words in the different dialects and make maps of the results.

In this tutorial, we focus on analyzing phonetic transcriptions. However, Gabmap can also be used for analyzing numeric dialect data (for example, formant frequencies) or categorical data (for example, syntactic or morphological variables).

Gabmap offers the following tools:

Summary of the data

When you have uploaded your data in the web application you will get an overview of the number of places, number of linguistic variables, number of different symbols used in the transcriptions etc. An index map of the data locations is created and there is a possibility to create distribution maps of any specific variant(s) of a word or of a character.

Measurement of linguistic distances

Several of the analyses offered in Gabmap are based on aggregate linguistic distances between language varieties. String edit distance is used for measuring linguistic distances of transcription data. The linguistic distances are displayed in different ways on a number of maps. All maps and figures in Gabmap can be downloaded as image files.

Statistical analyses and mappings

With the statistical tools in Gabmap you can explore the structure of your dialect data. With multidimensional scaling you can explore to what extent the dialects form a continuum, while cluster analysis classifies the dialects and detects dialect areas. The results of the statistical analyses are displayed in maps and figures.

2 Getting started

The dialect data used in this tutorial comes from the Linguistic Atlas of the Middle and South Atlantic States (LAMSAS, http://us.english.uga.edu/lamsas/). We use a subset from this atlas comprising 67 counties in the state of Pennsylvania. The dialect data (Pennsylvania.u16) and the map file (Pennsylvania.kml) can be downloaded at http://www.gabmap.nl/~app/doc/tutorial/data/. Save both files to your local hard disk (right-click on the file name and choose *Save Target As...*).

2.1 Check the data files

If you want to, you can inspect the files you just downloaded. The data file is a tab separated text file which can be opened in a text editor, or, in Microsoft Excel by using the Data Import wizard. The data have the format of a table. The sites where the data was collected are listed in the first column and the linguistic variables are in the first row. The cells in the table contain transcriptions of the dialectal pronunciations of the variables at each site.

The map file can be opened and edited using Google Earth. Google Earth is a free software which has to be downloaded and installed on your computer before you can use it.

2.2 Upload the data in Gabmap

Open Gabmap at http://www.gabmap.nl/. First you have to create an account by choosing a username and password. A message will be sent to you by email when you have created your account. Follow the instructions in the email to confirm the account. If you do not use an account for more than two months it will be deleted automatically.

Once you have logged in, you can create a new project of the Pennsylvania data by following the following steps:

1. Give the project a describing name (for example, Pennsylvania) at *Description*.
2. Upload the map file (Pennsylvania.kml).
3. Upload the data file (Pennsylvania.u16).
4. Choose *string data* at *Type of data*, since the data file comprises phonetic transcriptions. As type of processing we choose *string edit distance - tokenized*.
5. Click *Create project*.

3 Overviews of the data

When a project has been created, the menus in the *project view* show all the different types of results that have been created of the data. You can click on the topics to see the results. To return to the project view, you can click on the project in the upper left corner of the screen at any time.

3.1 Indexes

Indexes are created of the data sites and of the items in the data set. By clicking on *places* in the project view, you will see index maps of all the sites.

When you click on *items* in the project view, you get a list of the linguistic variables and the number of occurrences of each variable. Note that since there might be multiple transcriptions of one item from a single site and/or missing data, the number of instances of a variable might be smaller or greater than the number of places in the data set

3.2 Data overview

By clicking on *data overview* in the project view you get all kinds of summaries of the data. For example, you get a list with the total number of places in the data, the number of linguistic variables (items), the number of different characters used in the transcriptions etc. There is also a list of all the characters and a list of tokens. A token is a combination of a phonetic symbol and diacritical marks. You can try clicking on one of the blue numbers (which indicate number of occurrences) in the character or token list to see a map of the geographic distribution of the character/token. The darker the color on the map, the more frequent the character/token. Examples of the character/token are shown below the map.

Click on the blue number next to ‘t’ in the character list. What does the map show? Can you think of a linguistic explanation for the map?

Any of the maps in Gabmap can be downloaded by right-clicking on one of the file formats to the right of the map and choosing *Save Target As...*.

The character list with the number of occurrences of each symbol can be a very useful way of detecting errors in your own data sets, since very infrequent symbols might be typos.

4 Distribution maps

Go to the project view and click on *distribution maps*. The distribution maps show the geographic distribution of a chosen variant of a linguistic variable.

We could, for example, ask ourselves: What are the main variants of the pronunciation of the word ‘Georgia’? In the drop down menu of items we choose ‘Georgia’ and click *Select item*. All variants of this variable (with the number of occurrences in brackets) will be listed. To show a map you can click on one of the variants and then on the button *Show distribution map*. The map shows in which of the sites the specific variant is used; the darker the color on the map the more frequent use.

Some of the variants of the word ‘Georgia’ occur only once, while others are frequent. Look at some of the most frequent variants of this variable. Can you identify any geographic areas? Which linguistic features are distinguishing for the different areas?

A number of variants, all of them very infrequent, start with a ‘t’ instead of a ‘d’. By pressing the *Ctrl* button on your keyboard while you click on the variants, you can choose all of them before you click *Show distribution map*. In which area are the variants starting with a ‘t’ used?

Another way of selecting multiple variants simultaneously is to use regular expressions. For example, writing ‘t’ in the *Regular expression* box will automatically select all the variants that have a ‘t’ anywhere in the transcription string.

To make distribution maps of another item, choose the item in the drop down menu and click *Select item*.

Look at the distribution maps of the word ‘thousand’. What kind of variation do you find for this item?

5 Alignments

String edit distance (also called Levenshtein distance) is used in Gabmap for measuring linguistic distances based on phonetic transcriptions. String edit distance determines the distance between two different pronunciations of a lexical item by finding the smallest cost for changing one pronunciation into the other. Changing one pronunciation into the other is done by inserting, deleting or substituting characters. This process results in an alignment of the two pronunciations.

You can click on *alignments* under *Measuring technique* in the project view. This will bring you to a view where you can choose any of the lexical items and a site in order to see the alignments.

What is the linguistic distance between the pronunciations of the word ‘rose’ in the counties Lebanon and Monroe?

When using string edit distance to measure the linguistic distances in the data set, the distance between two dialects is first calculated separately for all lexical items. After that, the aggregate distance between the two dialects is calculated as the average of all the item distances.

(Alignments are created in Gabmap only for string data. For other types of data than transcriptions other distance measures are used which will not result in alignments.)

6 Inspecting the aggregate distances

The aggregate linguistic distance between two places is the average of all the item distances computed with the string edit distance (see previous topic). Under *Differences – statistics and difference maps* you can download the aggregate distances as a table. In this table you can look up the linguistic distance between any two sites (exactly like you would look up the distance in kilometers between two places in the distance chart in a road map!).

The aggregate distances are displayed in a number of mappings.

6.1 Difference maps

When you click on *statistics and difference maps* in the project view you will see some statistics of the aggregate linguistic distances and two maps. The maps are created by drawing lines between sites indicating how similar or different two sites are linguistically. The darker the color of the line on the map, the more similar are the sites linguistically.

In the first map, only neighboring sites are connected by lines. In the second map, lines are drawn across a larger geographic area.

Compare the two difference maps. In which areas are there large dialect differences? In which areas are the dialects very similar to each other? Are there abrupt dialect borders?

6.2 Reference point maps

Another way of visualizing linguistic distances are the *reference point maps*. These maps, made popular in Goebl and Haimerl's Visual Dialectometry (VDM, http://www.dialectometry.com/dmdocs/englisch_fr.html), show the linguistic distances from one chosen site to all other sites in the data set.

Go to *Differences – reference point maps*. Choose a place in the drop down list and click *Show map*. The chosen reference site is displayed by a star. All other sites are colored so that the sites that are linguistically most similar to the reference site are the darkest. The lighter the color the more different the dialect.

Compare the reference point maps of Pittsburgh and Philadelphia. What do these maps tell about the dialects?

7 Multidimensional scaling

Multidimensional scaling (MDS) is used for exploring dialect continua. In contrast to the reference point maps, which show the linguistic distances only from one chosen site to all other sites, MDS visualizes the linguistic relationships between all the sites in the data set simultaneously.

MDS is a statistical method which reduces complex distance data into interpretable dimensions. Generally, MDS is used to scale data into two or three dimensions, since more dimensions are very difficult to visualize. When applying MDS to dialect data, three dimensions are generally enough since many studies have shown that three dimensions explain at least around 90% of the total variance in dialect data.

If you go to *Multidimensional scaling - mds plots* in Gabmap, you see the result of MDS to two dimensions. All places have got a position in a two-dimensional coordinate system. Hovering the mouse cursor over the dots in the plot displays the site names. Linguistically similar sites are found close to each other in the plot, while linguistically different ones have a large distance in the plot. The plot can help to identify outliers in the data set or to recognize dialect groups (that is, clouds in the plot).

Which county in Pennsylvania has the most divergent dialect?

The MDS map displays the result of MDS to three dimensions. In MDS to three dimensions, positions in a three-dimensional space are assigned to all varieties included in the analysis. The three-dimensional space can be thought of as this color cube:

All dialects in the data set get a position in this cube and at the same time a specific color. Dialects that are linguistically similar to each other are placed close to each other in the MDS space and hence get similar colors, while dialects that are very different

from each other are assigned positions and colors far away from each other in the cube (for example, in opposite corners of the cube).

The three-dimensional MDS map (*Multidimensional scaling – mds maps*) is created by coloring the area of each site on the map with the color assigned by MDS. Separate maps of each of the first dimensions of the MDS are displayed under the three-dimensional MDS map in Gabmap.

Look at the three-dimensional MDS map of Pennsylvania. Are there any abrupt dialect borders? Are there transitional borders? Can clear dialect groups be identified?

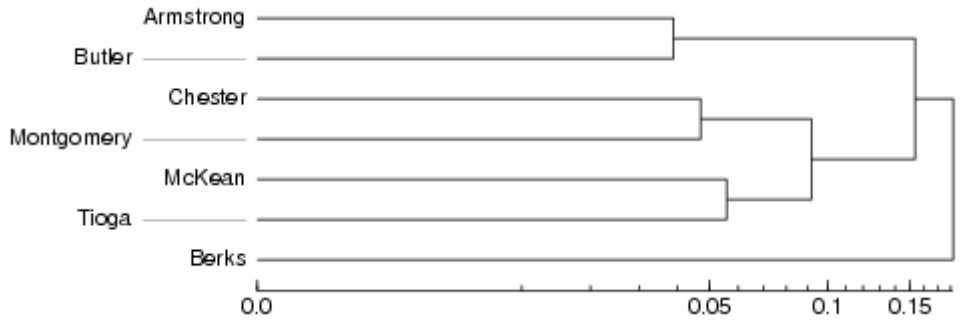
Compare the MDS map to the reference point map (see previous topic) of the place Huntingdon. What do the two different methods of visualizing dialect differences tell about dialects?

8 Cluster analysis

While MDS displays dialect continua, cluster analysis can be used for classifying dialects and identifying dialect areas. Clustering is the process of dividing a set of objects into groups (clusters). In our case, the objects are geographic places and we want to divide them into groups based on their linguistic similarity.

Cluster analysis is applied to the distance matrix with the pair-wise aggregate linguistic distances between places. In clustering, groups are merged based on similarity. To start with, each place is a cluster of its own, a cluster with only one element. The two places that have the smallest linguistic distance in the distance table are merged into a cluster. Then the difference is calculated between that new cluster, and all remaining places. Based on the new distances, again, the objects with the smallest difference are merged. And so on, until all places are merged into one big cluster.

The history of the clustering procedure is displayed in a dendrogram:

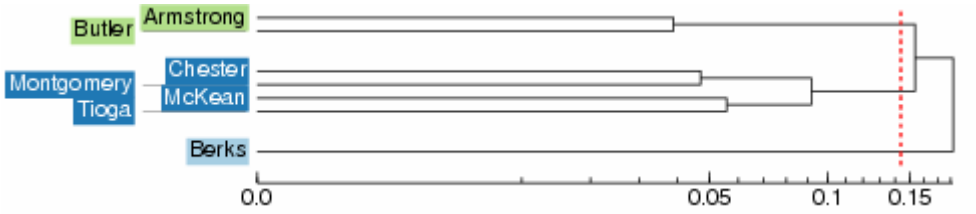


In this example dendrogram, Armstrong and Butler are the two places with the smallest linguistic distance in the distance matrix and they have been grouped together first. After that Chester and Montgomery have been joined, and in the third step McKean and Tioga. After these pairs have been formed, the clusters containing Chester and Montgomery and McKean and Tioga are joined to form a cluster with four items. Berks has such a large distance to all other sites that it has been merged with the rest only in the very last step of clustering.

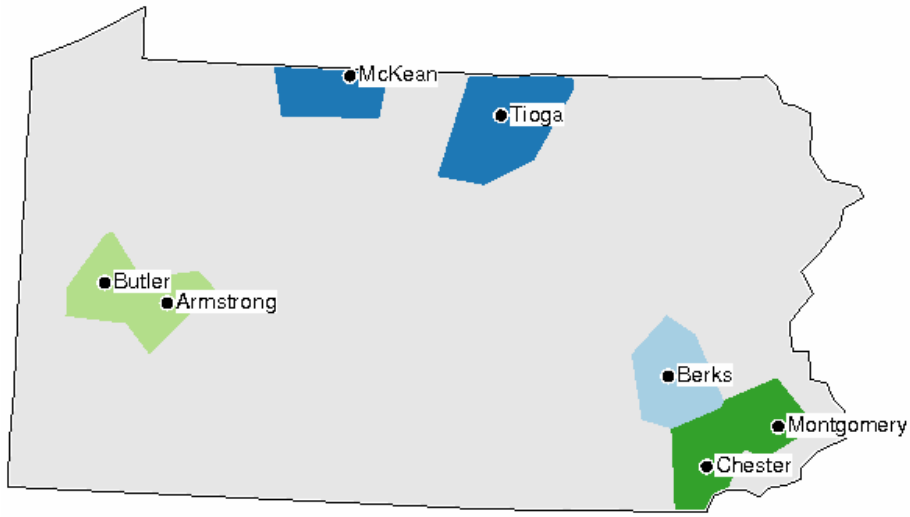
When making a dialect classification, you proceed from the right to the left in the dendrogram until you find the break point with as many branches as you want groups in your classification. The break point is marked with a read line in the dendrograms below.

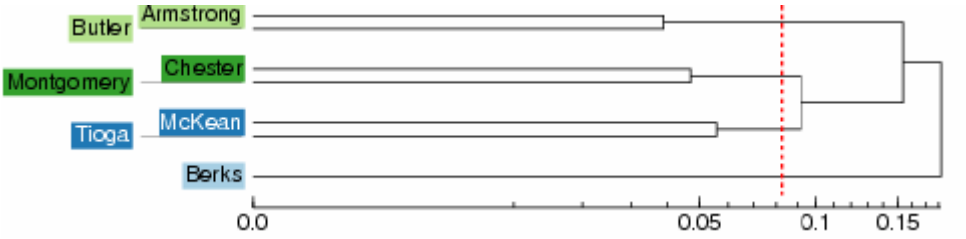
So if one would like to make a division into two clusters of this small data set, Berks would be in one group by itself and all the other places would form the second group together:

In a division into three groups, there would be one group with one member (Berks), one group with four members, and one group with two members:



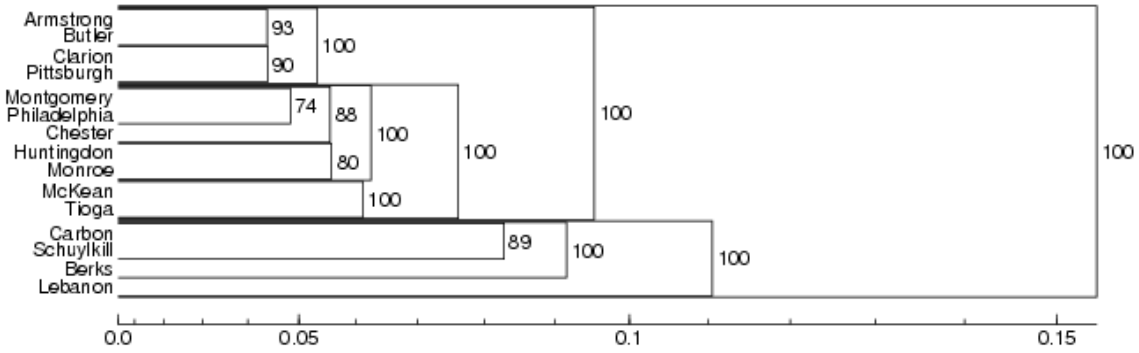
A map can be created by coloring the area of each cluster with a distinct color. The colors in these maps are arbitrary. Similarity of colors does not imply linguistic similarity, but each distinct color simply denotes one cluster. This is how a cluster map with four clusters would look:





8.1 Fuzzy clustering

A problem with cluster analysis is that it is a relatively unstable method. Small changes in the distance matrix can lead to large changes in the clustering results, which makes the results of clustering unreliable. Different methods have been developed to solve this problem. One such method is *fuzzy clustering*. Fuzzy clustering means that the original distance matrix is contaminated with (varying) small amounts of random noise. This is done several times and each time clustering is performed on the contaminated matrix. After that we count how many times each cluster has appeared. Clusters that appear in many runs of the analysis with added noise are particularly stable ones. The results are displayed in a probabilistic dendrogram:



The percentages in the probabilistic dendrogram indicate how many times each cluster was encountered in the repeated clustering with noise. In the example above, the largest clusters have been encountered in all the iterations and have a probability of 100%. We can therefore be pretty sure that these are real clusters. On the lower levels the percentages are somewhat smaller, so we cannot be completely sure about these clusters.

Go to *Fuzzy clustering - probabilistic dendrogram* in your Gabmap project.

Which dialect groups in Pennsylvania can be identified with high confidence?

Note that the colors in the probabilistic dendrogram are from a different analysis and are there to help you identify the places in the map under the dendrogram. This map is also displayed at *Fuzzy clustering - fuzzy cluster maps*. The map visualizes something between MDS and cluster analysis: main dialect groups are identified in the map, but continuous relationships are displayed for places which cannot be put in one group with high probability. The map is created by running MDS on the branch lengths of the dendrogram (so-called cophenetic distances) instead of on the original linguistic distances.

8.2 Discrete clustering

Results of clustering without noise are found under *Discrete clustering – cluster maps and dendrograms* in Gabmap. Remember that caution should be taken when interpreting these results, because cluster analysis is not a stable technique (see above).

You can inspect the results of four different clustering algorithms in this view in Gabmap. The four methods are *Complete Link*, *Group Average*, *Weighted Average* and *Ward's Method*. The different algorithms have different ways of determining how distances between newly formed clusters are calculated in the clustering process. The different methods have different biases, for example, Ward's Method favors equal size clusters, while the other methods are more true to the original linguistic distances.

The default cluster map represents eight clusters obtained by using Weighted Average for clustering.

Look at the map of eight clusters with Weighted Average as clustering method. Compare this map to the difference maps (*Differences – statistics and difference maps*). Is there an agreement between the line maps on the one hand and the cluster map on the other? What is similar? What is different?

Now, let's change clustering method. You can change the parameters in the box under the map in this view. Choose *Ward's Method* instead of *Weighted Average*, and then click *Change settings*. The default number of clusters displayed is eight, but you can change the number of clusters displayed yourself. Choose six instead of eight, and then click *Change settings* again.

Compare the results of six clusters using Ward's Method to the results of fuzzy clustering. Are the results similar?

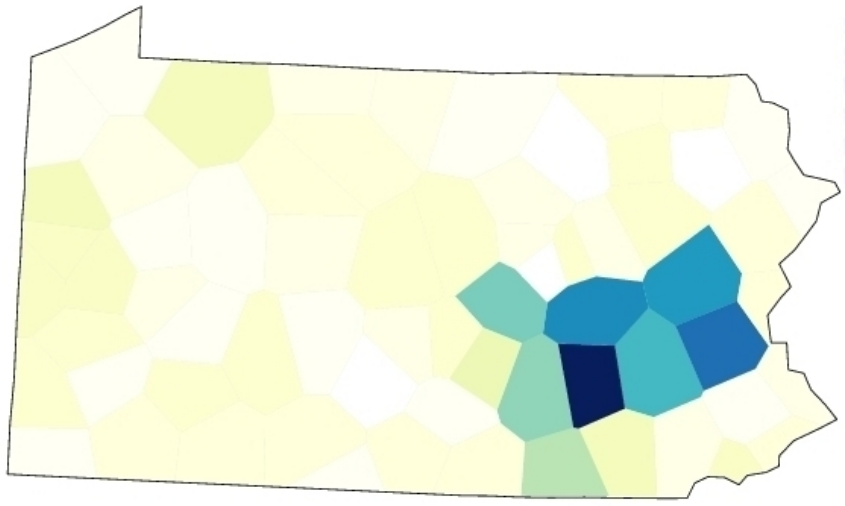
Compare the map of six clusters using Ward's Method to the one of eight clusters with Weighted Average as clustering method. Which one is better?

An additional method for validating the results of cluster analysis is available under *Discrete clustering – cluster validation*. A cluster maps is displayed, and under the map, there is a plot which is actually the same MDS plot which you find at *Multidimensional scaling – mds plots*. The plot is colored according to the chosen cluster analysis, so that you can compare the clustering results with the results of MDS. The plot helps you to see how well separated the clusters are. If no clear, separated clouds of dots can be identified in the plot, the data set is probably truly continuous and it does not make sense to use cluster analysis.

9 Solutions

9.1 Data overview

Click on the blue number next to ‘t’ in the character list. What does the map show? Can you think of a linguistic explanation for the map?



found 7315 items, showing 200

show codes

nu tʃərsɪ

Carbon — New Jersey

tɛnəsii

Clarion — Tennessee

The map above is the one you should see after clicking on the number of occurrences of ‘t’ in the character list. As you can see, there is an area in the south-east of Pennsylvania where ‘t’ is used much more frequently than in the rest of Pennsylvania (the darker the color on the map, the more frequent the symbol).

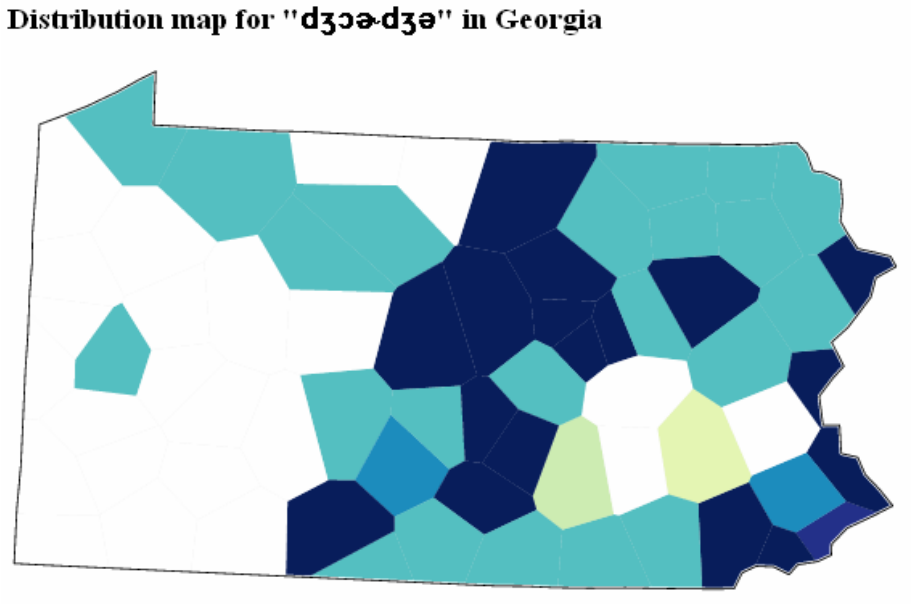
Under the map a number of samples of transcriptions including a ‘t’ are shown. Because ‘t’ is a very frequent symbol, all of the occurrences of ‘t’ cannot be shown in this list, but a random sample is made. If you were lucky, some of the transcriptions in the random sample that was generated for you showed words that are pronounced with ‘t’ in the blue area on the map, but without ‘t’ in other parts of Pennsylvania. The first sample in the sample list above, the pronunciation of the item ‘New Jersey’ in Carbon, is one example of this.

The following exercises in this tutorial will show some more elaborate ways of detecting geographic distribution patterns of dialectal features.

9.2 Distribution maps

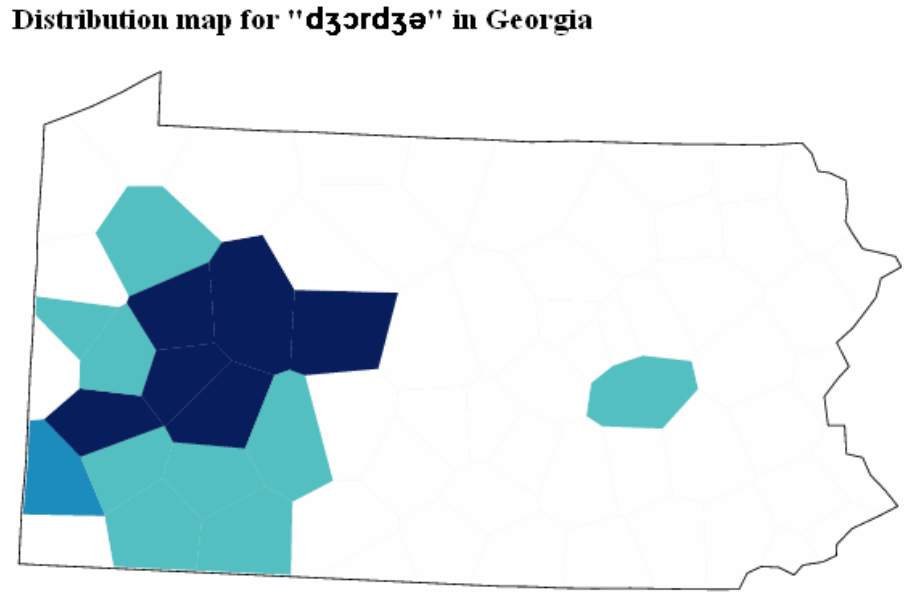
Some of the variants of the word ‘Georgia’ occur only once, while others are frequent. Look at some of the most frequent variants of this variable. Can you identify any geographic areas? Which linguistic features are distinguishing for the different areas?

The most frequent variant of ‘Georgia’ is [dʒɔə-dʒə] with 75 occurrences in the data set. If you choose this variant you should see the following map:



The map shows that the variant [dʒɔə-dʒə] is used mostly in eastern and northern parts of Pennsylvania.

The second most frequent variant of ‘Georgia’ is [dʒɔrdʒə] with 23 occurrences. This variant is used mostly in western parts of Pennsylvania:



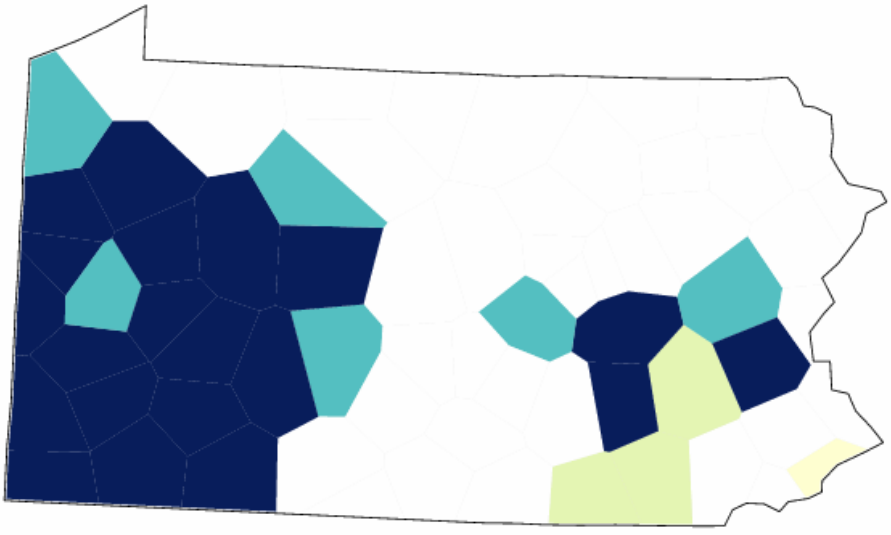
Using regular expressions

Using the *Regular expressions* box comes in handy when you want to find out which features in the word 'Georgia' that are the most decisive ones for the observed geographic distributions. Of the two main variants that we just made the maps of, the western one has an [r], while the eastern and northern variant has an r-colored schwa [ɚ] instead of [r].

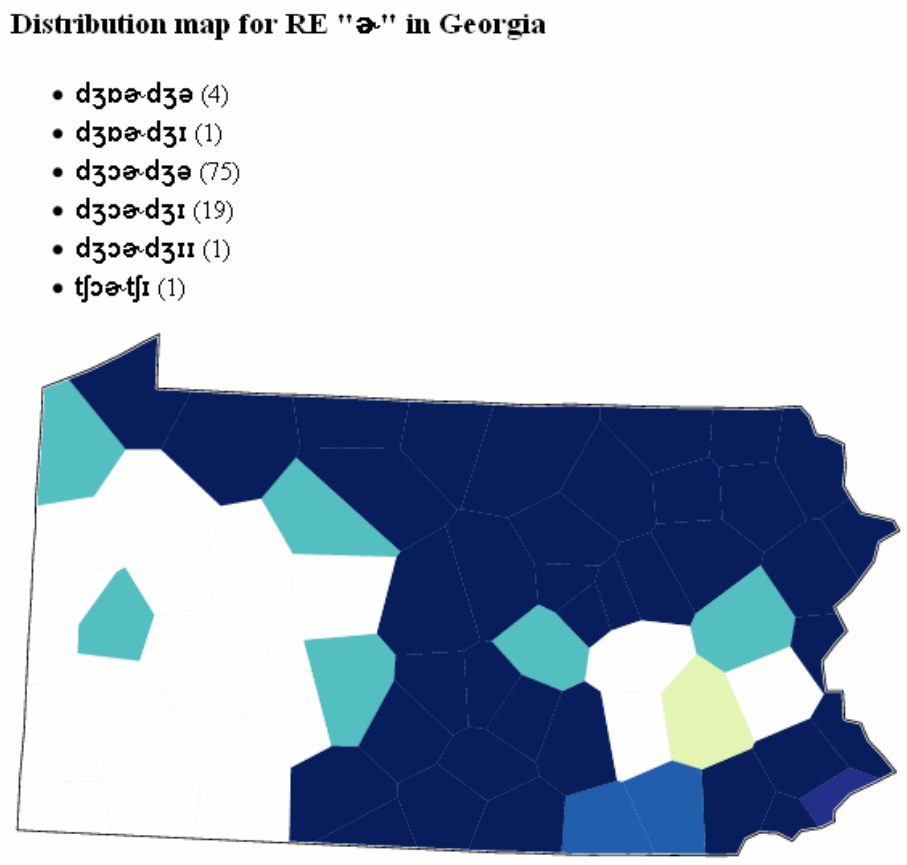
Typing 'r' in the box for regular expressions and clicking *Show distribution map* will automatically select all variants including an 'r' anywhere in the transcription and show the distribution map for these variants:

Distribution map for RE "r" in Georgia

- **dʒordʒə** (1)
- **dʒordʒɪ** (1)
- **dʒɒrdʒə** (1)
- **dʒɔrdʒə** (23)
- **dʒɔrdʒɪ** (11)
- **dʒɔrdʒɪɪ** (1)
- **dʒɔərdʒə** (4)
- **tʃɔrtʃə** (4)
- **tʃɔɾtʃə** (1)
- **tʃɔərtʃə** (2)
- **tʃɔəɾtʃə** (1)

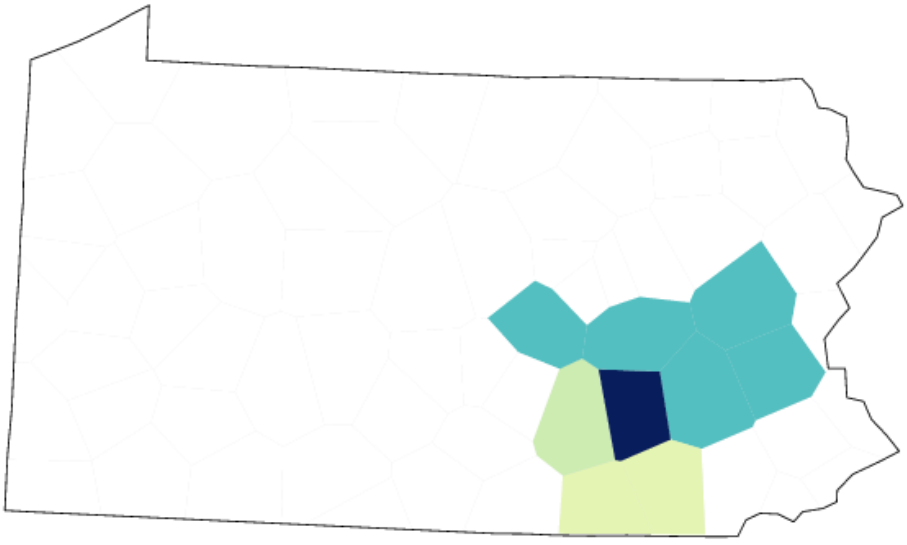


If you type 'ɚ' in the regular expressions box, you get the opposite map in this case (since IPA symbols are not available on the keyboard you can, for example, copy the symbol 'ɚ' from the character list and paste it in the box):



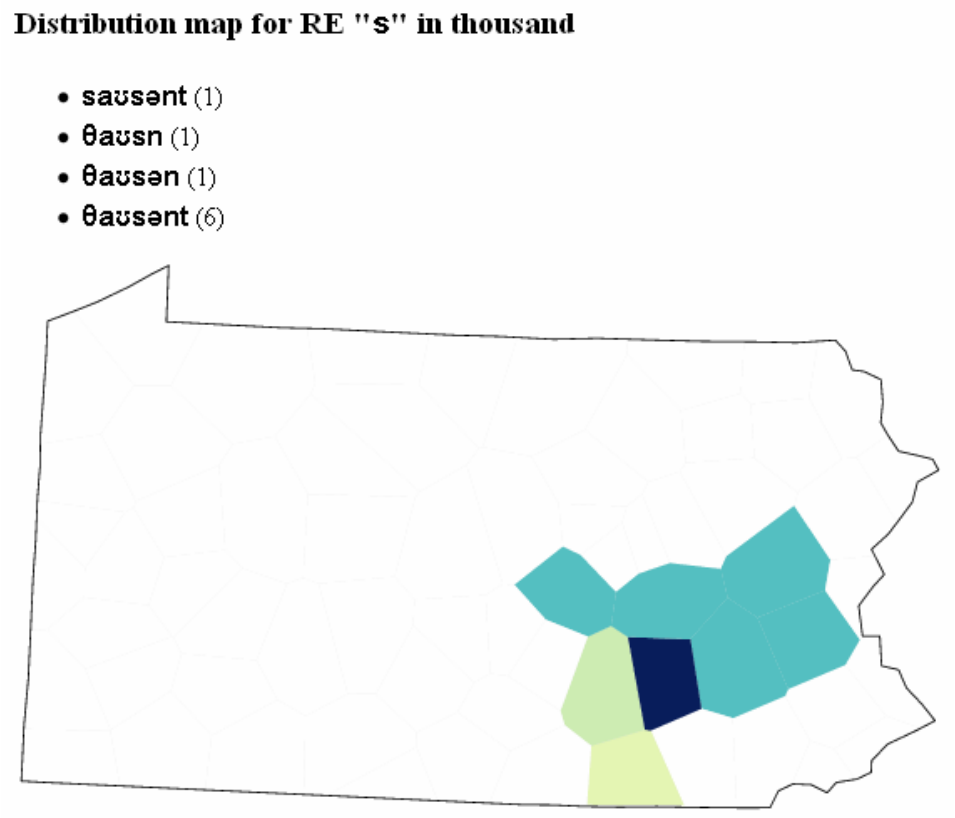
A number of variants, all of them very infrequent, start with a ‘t’ instead of a ‘d’. By pressing the *Ctrl* button on your keyboard while you click on the variants, you can choose all of them before you click *Show distribution map*. In which area are the variants starting with a ‘t’ used?

If you chose all variants starting with a ‘t’, this is the map you should see:



Look at the distribution maps of the word ‘thousand’. What kind of variation do you find for this item?

The first vowel in the word ‘thousand’ has the realizations [a], [æ] and [ɛ]. The variation in this vowel, however, does not show any very clear geographic variation. A feature that shows more geographic coherence is realization of the ‘s’ in ‘thousand’. For example, choosing all variants with a voiceless [s] gives the following map:



The map of voiceless [s] in ‘thousand’ looks very similar to the one of [t] in ‘Georgia’ above. They both display the same south-eastern area in Pennsylvania. This part of Pennsylvania was settled mainly by Germans. When the LAMSAS data was collected a number of the speakers in this area still had “Pennsylvania Dutch” (the variety of German spoken in Pennsylvania) as their first language and English as their second language. German language has influenced the English spoken in this area. The maps above show examples of features that are the result of German influence on English.

9.3 Alignments

What is the linguistic distance between the pronunciations of the word ‘rose’ in the counties Lebanon and Monroe?

To see the alignments, you first have to choose the item ‘rose’ in the drop down item list. Then you choose one of the sites Lebanon or Monroe in the list of places and click *Show alignments*. This will show the list of all the alignments of the word ‘rose’ for the selected place. In the list, you find the pair Lebanon-Monroe:

Lebanon — Monroe

r	o		s	
r	o	ʊ	z	
		1	1	2

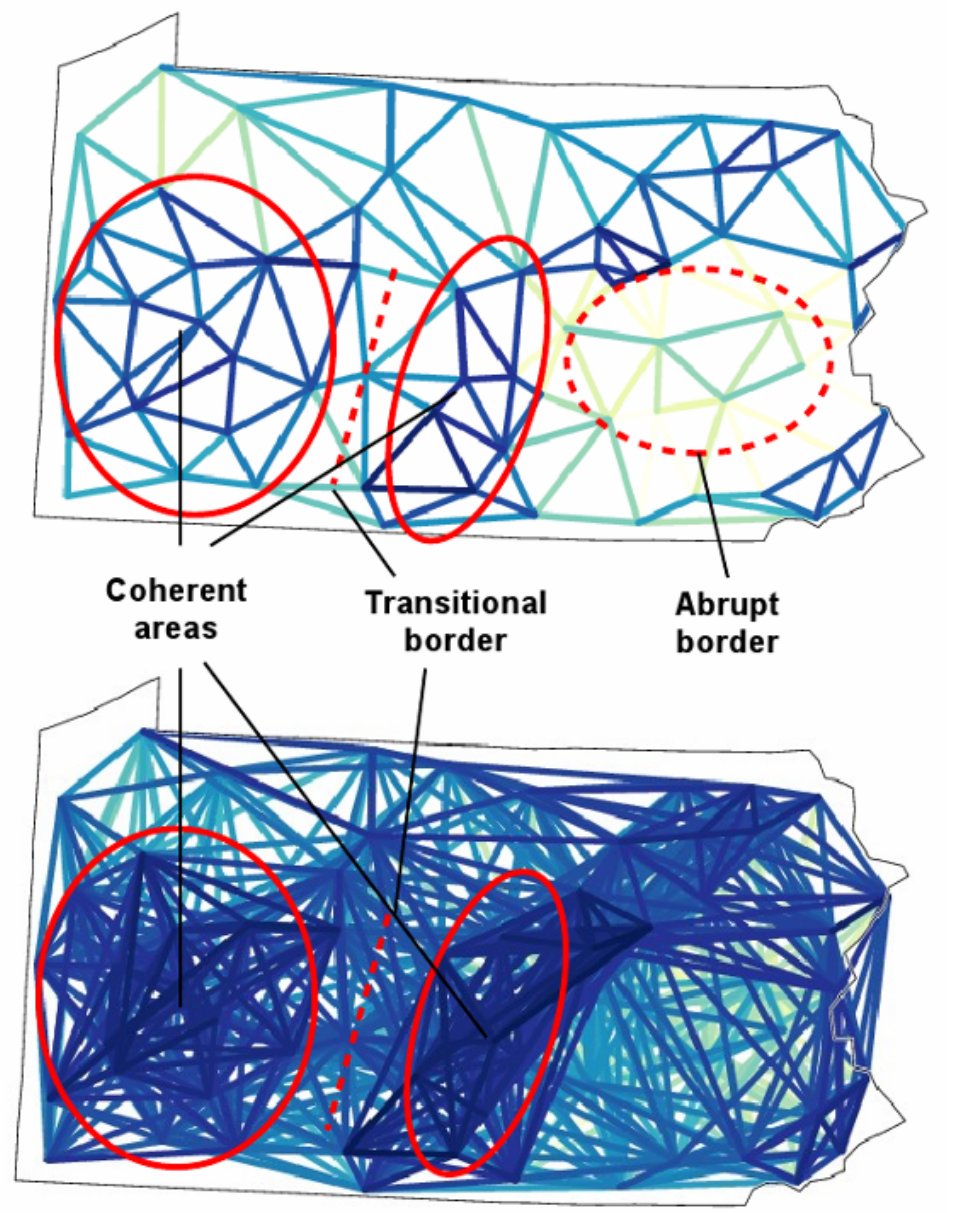
The alignment shows that in Lebanon the word ‘rose’ is pronounced [ros], while the pronunciation in Monroe is [roʊz]. Because the pronunciation in Monroe is diphthongized while Lebanon has a monophthongal vowel, the first operation is the insertion of ʊ, which gives a cost of 1. The second cost is a substitution cost: at the end of the word Lebanon has a voiceless s, while Monroe has voiced z. The sum of the two operations is 2, which is the linguistic distance between the two pronunciations.

9.4 Difference maps

Compare the two difference maps. In which areas are there large dialect differences? In which areas are the dialects very similar to each other? Are there abrupt dialect borders?

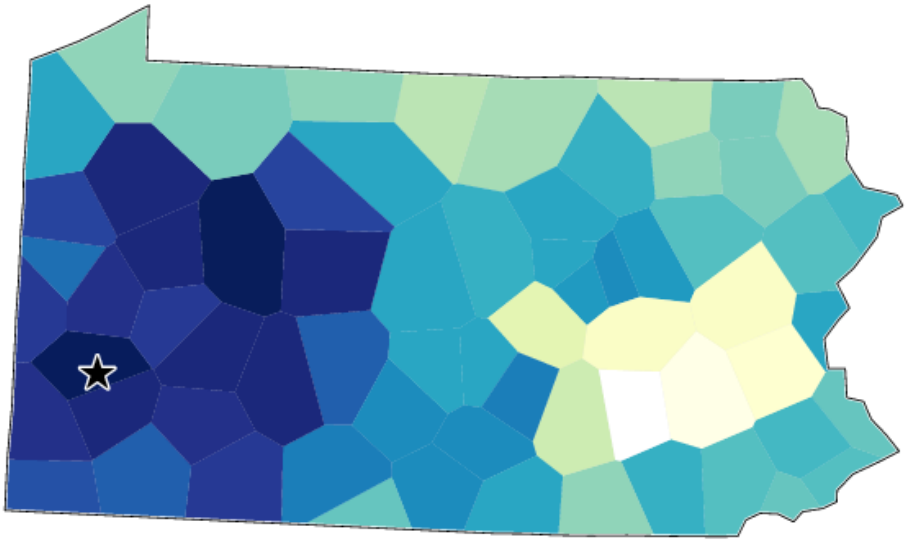
Both of the maps display dark lines in the south-west of Pennsylvania, suggesting that the varieties spoken in this area are relatively similar to each other. Another coherent area is found in central Pennsylvania. Between these two areas there are lines which are neither very dark nor very light, suggesting a transitional border.

In eastern Pennsylvania there is an area with very large dialect differences. The dialect border around this area is an abrupt one, and also the dialects within this area are quite different from each other. The lower map shows that there are dialects to the north and to the south of this area that are more similar to each other than to the ones that are geographically in between.

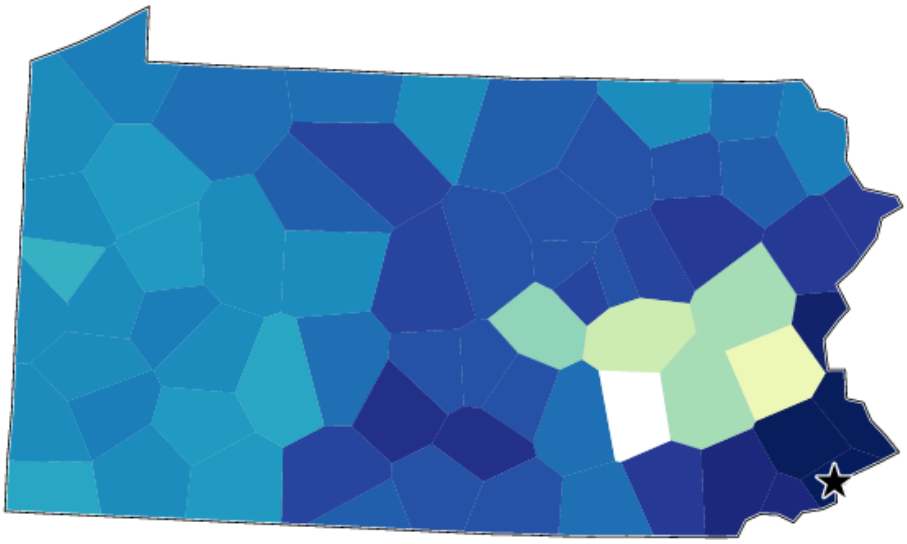

Coherent areas
Transitional border
Abrupt border

9.5 Reference point maps

Compare the reference point maps of Pittsburgh and Philadelphia. What do these maps tell about the dialects?



Pittsburgh



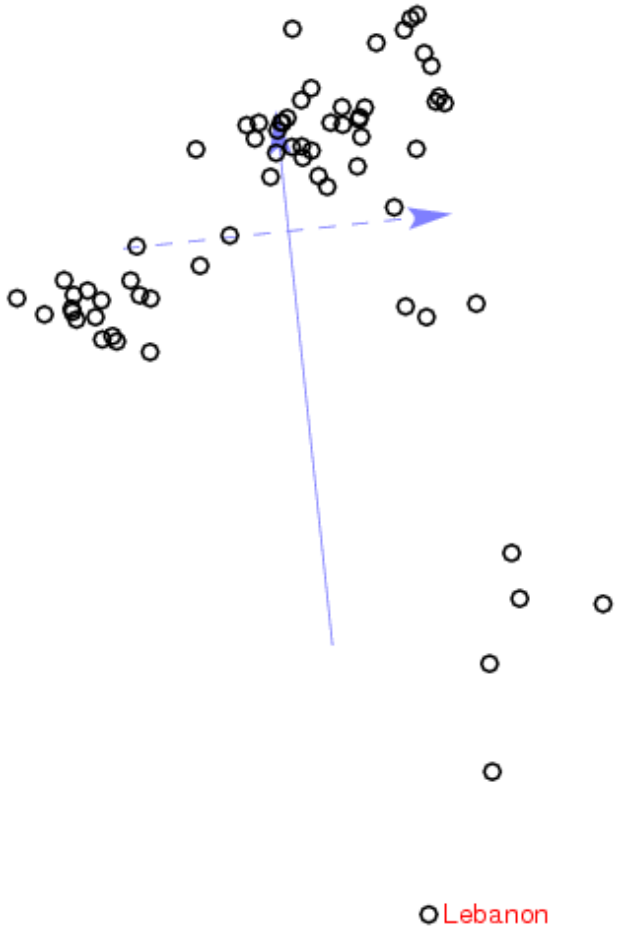
Philadelphia

You can think of the reference point maps as "How different or similar would other dialects sound to a person in Pittsburgh/Philadelphia?". For a person in Pittsburgh, the dialects in the surrounding area are all very similar to the Pittsburgh variety, but both in the north and in the east people sound quite different. For a person in Philadelphia only the geographically closest places have a pronunciation very similar to the one in Philadelphia. When you get further away dialects get gradually more and more different, but the most different sounding dialects are actually found in an area very close by, to the north-west of Philadelphia.

9.6 Multidimensional scaling

Which county in Pennsylvania has the most divergent dialect?

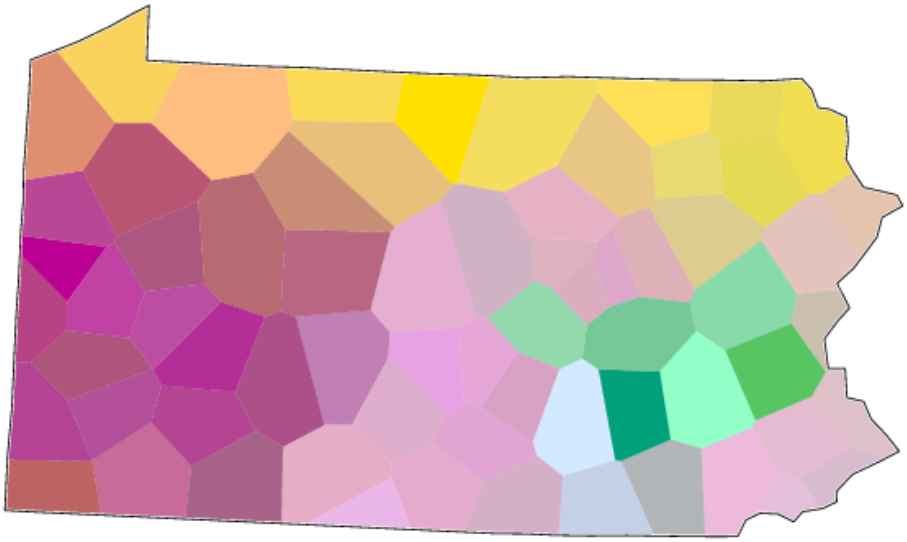
MDS plot:



The site furthest away from all other sites in the MDS plot is Lebanon.

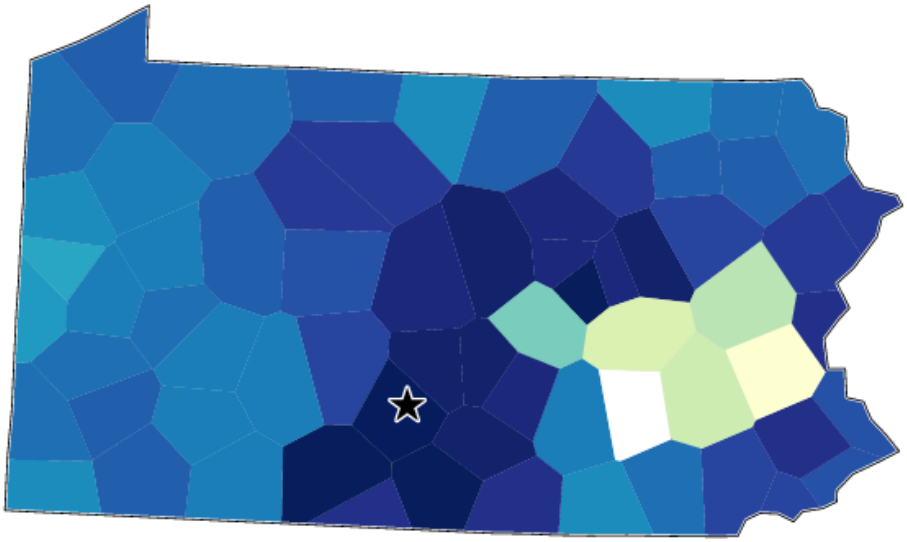
Look at the three-dimensional MDS map of Pennsylvania. Are there any abrupt dialect borders? Are there transitional borders? Can clear dialect groups be identified?

The green area in the south-east forms a clear dialect group. The rest of Pennsylvania can roughly be divided into East, West and North, with gradual transitions between the areas.



MDS map

Compare the MDS map to the reference point map of the place Huntingdon. What do the two different methods of visualizing dialect differences tell about dialects?



ref. point map

In the reference point map of the place Huntingdon we can see that the dialects in central Pennsylvania, close to Huntingdon, are very similar to the dialect of Huntingdon. Places in the west and in the north of Pennsylvania all have a quite similar shade of blue, indicating intermediate linguistic distances to Huntingdon.

Similarly to the reference point map, the MDS map shows that central Pennsylvanian dialects are similar to each other. But in the MDS map we can also see that western and northern Pennsylvanian dialects are actually very different from each other even though they are about equally different from the central Pennsylvanian dialects.

While the reference point map shows the view of someone standing in Huntingdon listening to the other Pennsylvanian dialects, the perspective of the MDS map is more that of an “objective observer” who is trying to hear all the difference between Pennsylvanian dialects.

9.7 Fuzzy clustering

Which dialect groups in Pennsylvania can be identified with high confidence?

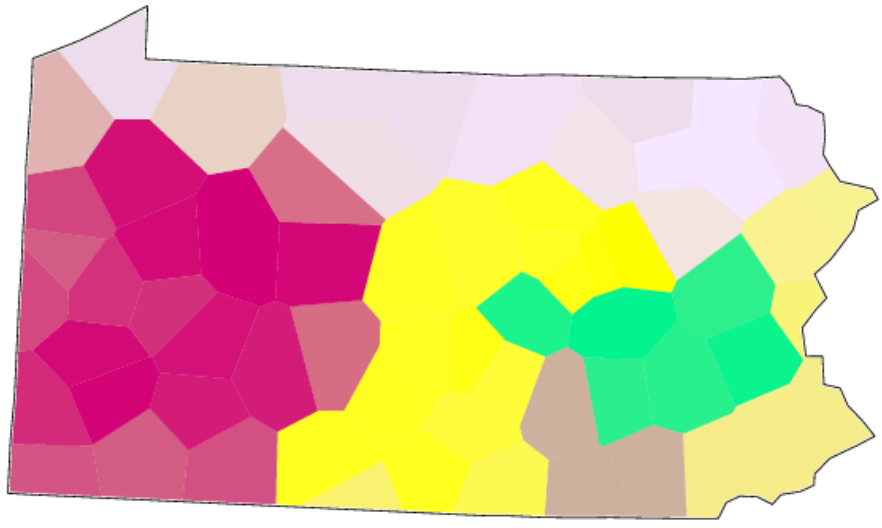
With fuzzy clustering, a division into two groups can be made of the Pennsylvanian dialects with 100% confidence. This two-way split separates south-eastern (green; German influenced, see Section 9.2) dialects from the rest of the Pennsylvanian dialects. The hierarchical structure within the green cluster cannot be detected with very high confidence.

The rest of the dialects can roughly be divided into three groups: one small cluster (brown; the counties Lancaster, York, and Dauphin), a western cluster (red), and a larger cluster with mainly northern and central dialects.

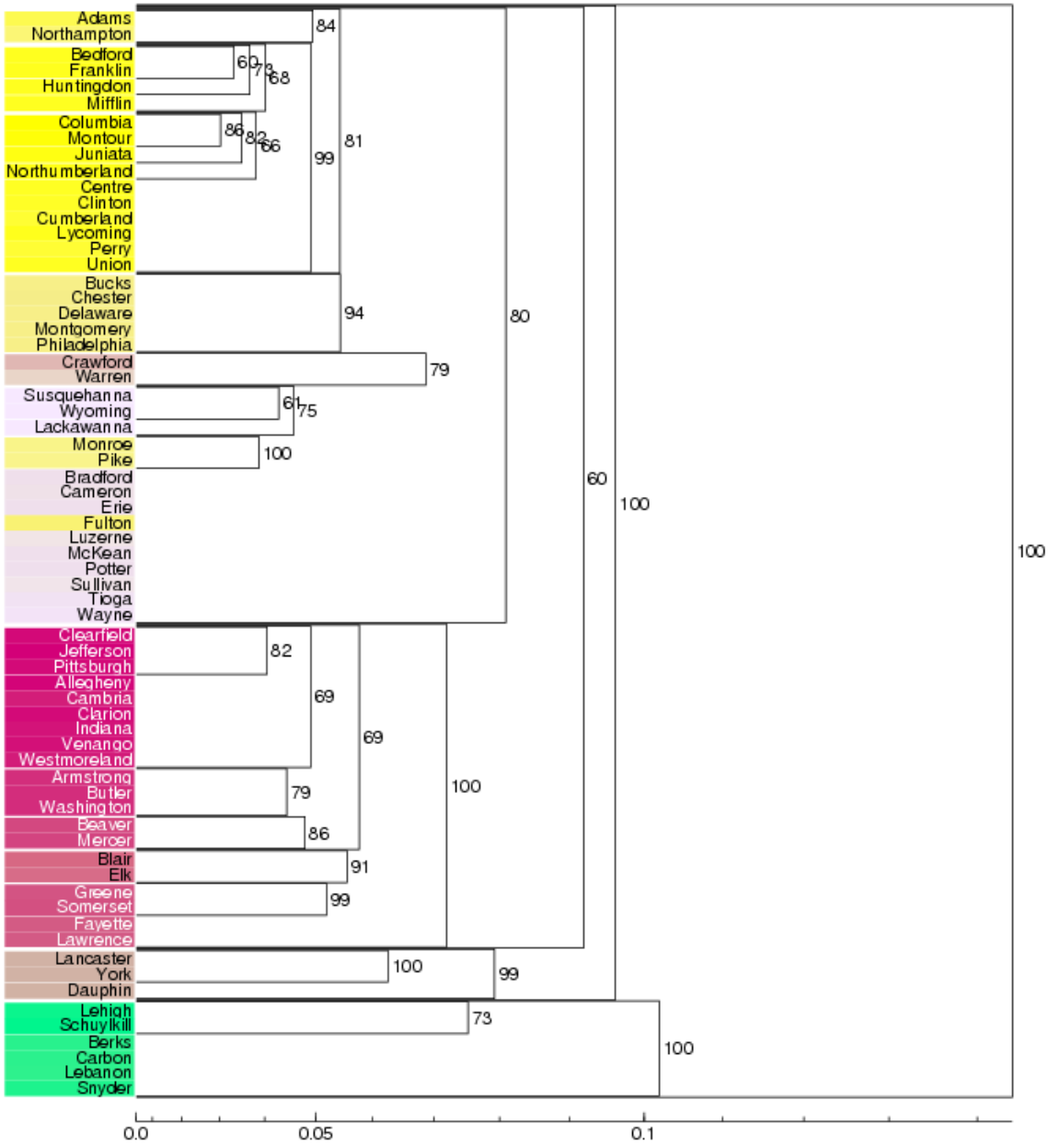
Within the northern and central cluster, the central (yellow) area is detected with 99% confidence. The counties in the south-eastern corner of Pennsylvania (Bucks, Chester, Delaware, Montgomery, Philadelphia) also form a dialect group with very high confidence (94%).

At lower hierarchical levels there are some small clusters that seem very stable. For a dialect division the large groups are generally more interesting than small subclusters.

fuzzy cluster map:



probabilistic dendrogram:

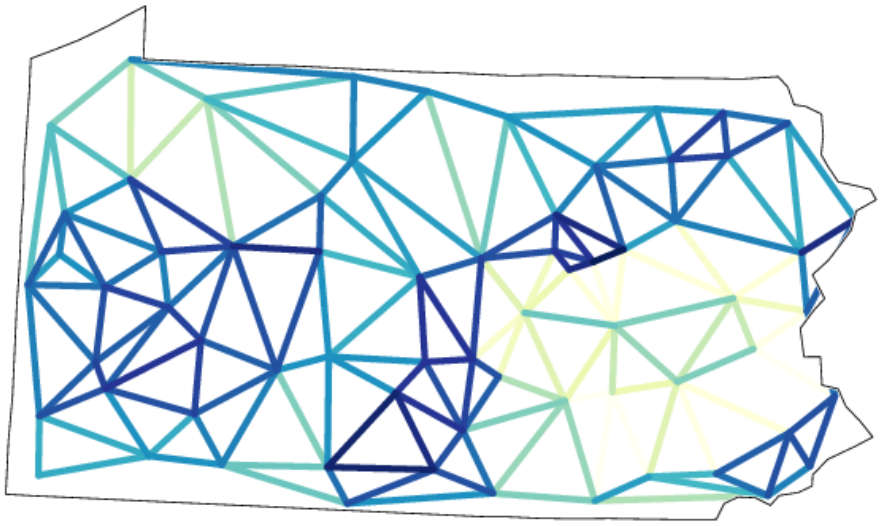


9.8 Discrete clustering

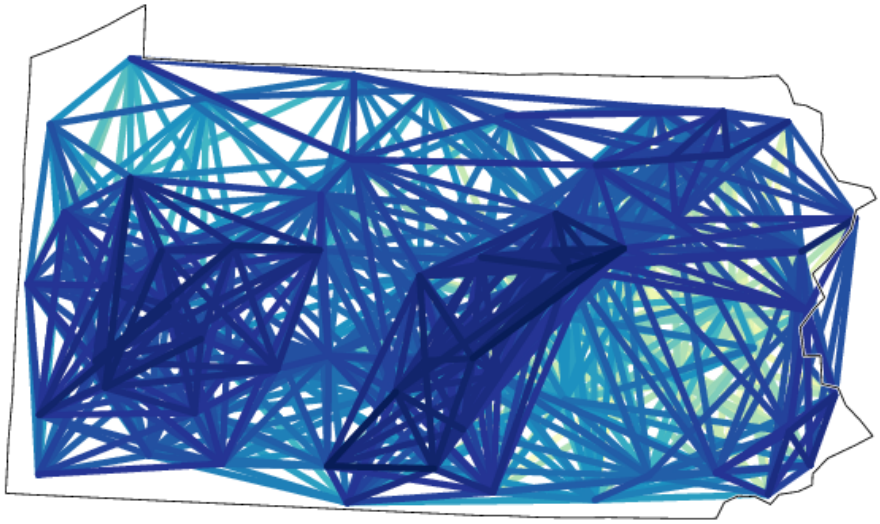
Look at the map of eight clusters with Weighted Average as clustering method. Compare this map to the difference maps (*Differences – statistics and difference maps*). Is there an agreement between the line maps on the one hand and the cluster map on the other? What is similar? What is different?

Clustering with the Weighted Average method counts the divergence of the dialects in the east very heavily. There are many small clusters in the east, some of them comprising only one place. In comparison with the line map these results are not that surprising, the line maps show that there are very large dialect differences in the east. The western area can be identified in both the line maps and in the cluster map. The big

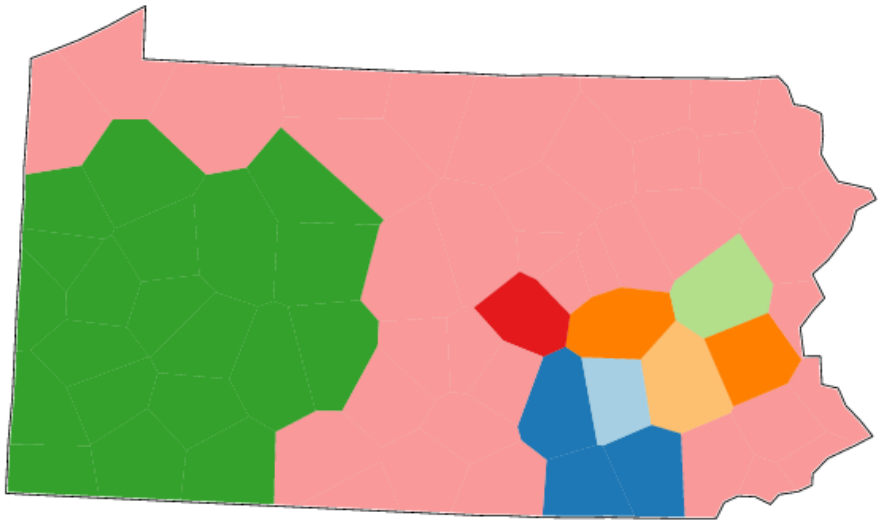
pink area in the cluster map seems to be a rest area in which no internal structure has been identified.



difference map 1



difference map 2



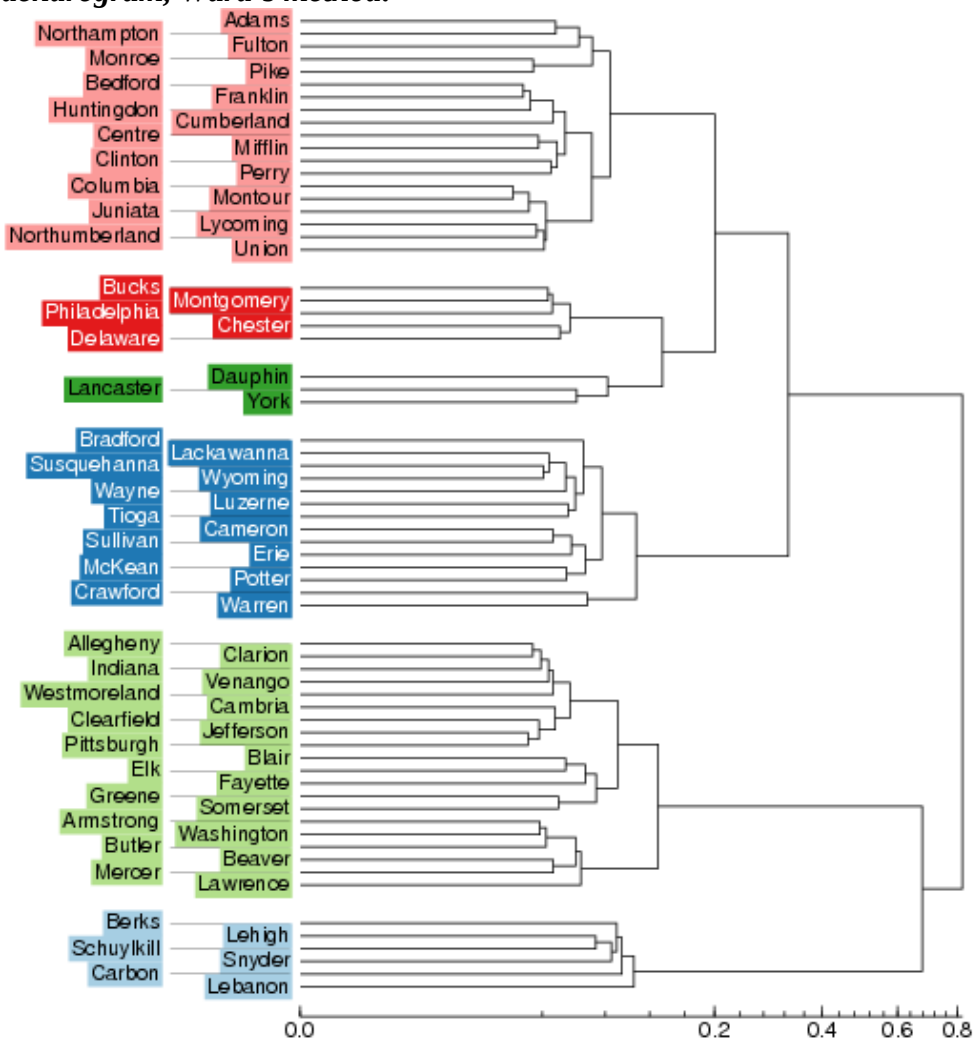
8 clusters, WA

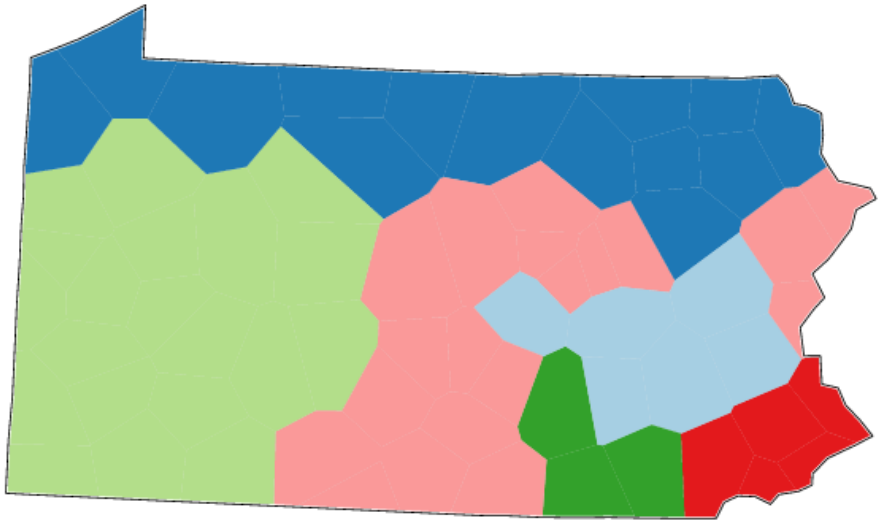
Compare the results of six clusters using Ward's Method to the results of fuzzy clustering. Are the results similar?

The cluster map with six clusters extracted with Ward's Method looks very similar to the fuzzy cluster map: The German-influenced area in Pennsylvania can be identified, there is a western, a northern, and a central cluster, the south-eastern corner of Pennsylvania is identified, and there is the small cluster with only the three counties Lancaster, York, and Dauphin.

However, the first major split into two branches in the dendrogram of Ward's Method puts the light blue German-influenced area and the light green western area into one group, and the northern and the remaining eastern area in the other. So if you would make a classification into two groups this is the grouping you would get. But the fuzzy clustering separates the German-influenced dialects from all the other ones with high confidence in the first step. The probabilistic dendrogram uses Group Average and Weighted Average as clustering methods. These methods do not have the bias of Ward's method to try to make equal size clusters.

dendrogram, Ward's method:





6 clusters, WM

Compare the map of six clusters using Ward's method to the one of eight clusters with Weighted Average as clustering method. Which one is better?

There is no right or wrong answer to this question. When you use cluster analysis you should think about why you want to classify dialects and choose the method that suits your purposes best. Ward's method produces relatively equally-sized clusters, but does not acknowledge the fact that the distances within one cluster might be larger than across some of the other groups. Weighted Average gives a strong signal of the linguistic distances, but producing classes with only one member might not be meaningful in all circumstances. It is important to validate results of cluster analysis, for example, by using fuzzy clustering or by comparing to the results of MDS.