Drawing Isogloss Lines

Harald Hammarstrom

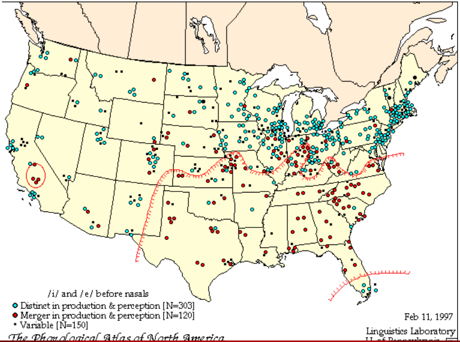
17 Sep 2014, Amsterdam

Drawing Isogloss Lines

An isogloss is the geographical boundary of a certain linguistic feature, . . . such as the pronunciation of a vowel, the meaning of a word, or use of some syntactic feature (Wikipedia 8 June 2010)

- Widely used in dialectology
- Example, pin/pen merger as of Labov (1997):

`http://www.ling.upenn.edu/phono_atlas/maps/Map3.html`



Approaches to Isogloss Lines

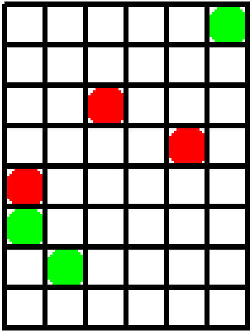
There appears to be no objective definition of an isogloss line, let alone an automated procedure for drawing one

- Dialectologists today draw isogloss lines by hand, based on intuition (p.c. Bert Vaux 2010)
- If you know of formal approaches to drawing isogloss lines, do let us know!
- This is involves a certain amount of subjectivity
- Today we will suggest a plausible definition and a procedure for actually drawing the line that fits the definition

Problem Setting #1: Input

Given:

- 2D grid map with
- rings ("red") and crosses ("green") and empty positions

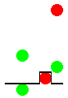


Problem Setting #2: "Line" Assumptions

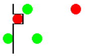
Assumptions about a "line":

- A line is not necessarily a straight line
- But, either
  - Runs from the west end to the east end on the map, crossing each column at exactly once OR
  - Runs from the north end to the south end on the map, crossing each row at exactly once

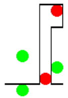
- Legal



- Legal



- NOT Legal



Some Labelling Conventions

For any line on the map separating rings and crosses, label (by majority) one of the sides the rings-side and the other the crosses-side

Let x be the number of points (either rings or crosses) on the crosses side

Let o be the number of points (either rings or crosses) on the rings side

A correctly classified point is a ring that occurs on the rings-side (c_x) or a cross that occurs on the cross-side c_o

A misclassified point is a ring that occurs on the crosses-side (m_o) or a cross that occurs on the rings-side (m_x)

Definition of an Isogloss Line

Absolute-Optimal The line that maximizes the total number of correctly classified points. I.e. the line that maximizes

$$c_x + c_o$$

(Equivalent to minimizing $m_x + m_o$)

Proportion-Optimal The line that maximizes the *proportion* of correctly classified points to the total number of points, on both sides. I.e. the line that maximizes

$$\frac{c_x}{x} + \frac{c_o}{o}$$

(Equivalent to minimizing $\frac{m_x}{x} + \frac{m_o}{o}$)

Example

Absolute-Optimal: The total number of correctly classified points

i $c_x + c_o = 2 + 2 = 4$

ii $c_x + c_o = 3 + 2 = 5$

So line (ii) is better.

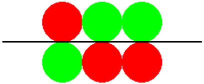
Proportion-Optimal: The *proportion* of correctly classified points to the total number of points, on both sides

i $\frac{c_x}{x} + \frac{c_o}{o} = 2/3 + 2/3$

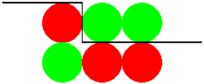
ii $\frac{c_x}{x} + \frac{c_o}{o} = 3/3 + 2/3$

So line (ii) is better.

i



ii



First Question: Equivalence

Are the two definitions equivalent?

$$\iff$$

For all possible 2D maps with rings and crosses, do the two definitions always yield the same isogloss line(s)?

Counterexample

x
x
x
x
x
x
a
o
o
x
b
o
o
o

- The optimal cut for a-isogloss lines has one misclassified x
- The optimal cut for b-isogloss lines has $3/3 + 11/13 = 1.846$ which is better than the a-line $(5/6 + 10/10 = 1.83)$

So the two definitions a/b are not equivalent

Algorithms for Finding Isogloss Lines

- For an input map of width x and height y
- A defining a line amounts to setting the height of the line at each column (east-west) or setting the breadth of the line at each row (north-south)
- There are y^x logically possible east-west lines and x^y logically possible north-south lines
- Although there are exponentially many lines, there exist
  - A linear-time $O(x \cdot y)$ algorithm for absolute-optimal isogloss lines
  - A polynomial-time $O((x \cdot y)^3)$ algorithm for proportional-optimal isogloss lines

An Algorithm for Absolute-Optimal Isogloss Lines

1. For each column C_i
   1. For each height H_j
      1. Compute the number of correctly classified instances in column C_i if the (segment of) the line were at height H_j
      2. Take the height with the max of all height-scores in column C_i
   2. Glue together the final isogloss line from the max-score heights of each column

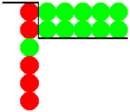
- North-south lines and opposite majority sides by simple rotations
- Computes the absolute-optimal line because every column makes an independent contribution to the score to be optimized

An Algorithm for Proportion-Optimal Isogloss Lines

- Keep a list R of c_x, x, c_o, o-combinations encountered so far, and a corresponding line (segment) for each

1. Expanding horizon rightwards, for each column C_i
    1. For each combination of a point in R and a height H_j in column C_i, calculate new c_x, x, c_o, o-combinations
    2. Set R to this (new) list of combinations
2. Take the max-scoring combination of the final total R

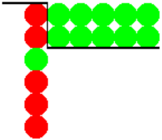
Algorithm for Proportion-Optimal Isoglosses: Example Run



<table>
<tr><th>C</th><th>H</th><th>C Counts</th><th>Aggregated Counts</th></tr>
<tr><td>0</td><td>1</td><td>0/5 1/1</td><td>0/5 1/1: 1</td></tr>
<tr><td>0</td><td>2</td><td>1/5 1/1</td><td>1/5 1/1: 2</td></tr>
<tr><td>0</td><td>3</td><td>2/5 1/1</td><td>2/5 1/1: 3</td></tr>
<tr><td>0</td><td>4</td><td>3/5 1/1</td><td>3/5 1/1: 4</td></tr>
<tr><td>0</td><td>5</td><td>3/5 0/1</td><td>3/5 0/1: 5</td></tr>
<tr><td>0</td><td>6</td><td>4/5 0/1</td><td>4/5 0/1: 6</td></tr>
<tr><td>0</td><td>7</td><td>5/5 0/1</td><td>5/5 0/1: 7</td></tr>
<tr><td>1</td><td>1</td><td>0/0 2/2</td><td>3/5 3/3: 4-1, 1/5 3/3: 2-1, 2/5 3/3: 3-1, 4/5 2/3: 6-1, 0/5 3/3: 1-1, .</td></tr>
<tr><td>1</td><td>2</td><td>0/0 2/2</td><td>3/5 3/3: 4-2, 1/5 3/3: 2-2, 2/5 3/3: 3-2, 4/5 2/3: 6-2, 0/5 3/3: 1-2, .</td></tr>
<tr><td>1</td><td>3</td><td>0/0 2/2</td><td>3/5 3/3: 4-3, 1/5 3/3: 2-3, 2/5 3/3: 3-3, 4/5 2/3: 6-3, 0/5 3/3: 1-3, .</td></tr>
<tr><td>1</td><td>4</td><td>0/0 2/2</td><td>3/5 3/3: 4-4, 1/5 3/3: 2-4, 2/5 3/3: 3-4, 4/5 2/3: 6-4, 0/5 3/3: 1-4, .</td></tr>
<tr><td>1</td><td>5</td><td>0/0 2/2</td><td>3/5 3/3: 4-5, 1/5 3/3: 2-5, 2/5 3/3: 3-5, 4/5 2/3: 6-5, 0/5 3/3: 1-5, .</td></tr>
<tr><td>1</td><td>6</td><td>0/0 1/2</td><td>3/5 2/3: 4-6, 1/5 2/3: 2-6, 2/5 2/3: 3-6, 4/5 1/3: 6-6, 0/5 2/3: 1-6, .</td></tr>
<tr><td>1</td><td>7</td><td>0/0 0/2</td><td>3/5 1/3: 4-7, 1/5 1/3: 2-7, 2/5 1/3: 3-7, 4/5 0/3: 6-7, 0/5 1/3: 1-7, .</td></tr>
<tr><td>2</td><td>1</td><td>0/0 2/2</td><td>1/5 5/5: 2-[1|2|3|4|5]-1, 4/5 3/5: 6-6-1, 3/5 5/5: 4-[1|2|3|4|5]-1, ...</td></tr>
<tr><td>2</td><td>2</td><td>0/0 2/2</td><td>1/5 5/5: 2-[1|2|3|4|5]-2, 4/5 3/5: 6-6-2, 3/5 5/5: 4-[1|2|3|4|5]-2, ...</td></tr>
<tr><td>2</td><td>3</td><td>0/0 2/2</td><td>1/5 5/5: 2-[1|2|3|4|5]-3, 4/5 3/5: 6-6-3, 3/5 5/5: 4-[1|2|3|4|5]-3, ...</td></tr>
<tr><td>2</td><td>4</td><td>0/0 2/2</td><td>1/5 5/5: 2-[1|2|3|4|5]-4, 4/5 3/5: 6-6-4, 3/5 5/5: 4-[1|2|3|4|5]-4, ...</td></tr>
<tr><td>2</td><td>5</td><td>0/0 2/2</td><td>1/5 5/5: 2-[1|2|3|4|5]-5, 4/5 3/5: 6-6-5, 3/5 5/5: 4-[1|2|3|4|5]-5, ...</td></tr>
<tr><td>2</td><td>6</td><td>0/0 1/2</td><td>1/5 4/5: 2-[1|2|3|4|5]-6, 4/5 2/5: 6-6-6, 3/5 4/5: 4-[1|2|3|4|5]-6, ...</td></tr>
<tr><td>2</td><td>7</td><td>0/0 0/2</td><td>1/5 3/5: 2-[1|2|3|4|5]-7, 4/5 1/5: 6-6-7, 3/5 3/5: 4-[1|2|3|4|5]-7, ...</td></tr>
<tr><td>...</td><td></td><td></td><td></td></tr>
<tr><td>5</td><td>6</td><td>0/0 1/2</td><td>2/5 2/11: 3-7-7-7-7-6, 4/5 7/11: 6-[1|2|3|4|5]-7-[1|2|3|4|5]-[1|2|3|4|5]-6;...</td></tr>
<tr><td>5</td><td>7</td><td>0/0 0/2</td><td>2/5 1/11: 3-7-7-7-7-7, 4/5 6/11: 6-[1|2|3|4|5]-7-[1|2|3|4|5]-[1|2|3|4|5]-7;...</td></tr>
</table>

Algorithm for Proportion-Optimal Isoglosses: Example *R*

<table>
<tr><th>C</th><th>|R|</th><th>Top Items</th><th></th><th></th><th>Bottom Colour</th></tr>
<tr><td>0</td><td>14</td><td>1.60</td><td>3/5 1/1</td><td>4</td><td>red</td></tr>
<tr><td></td><td></td><td>1.40</td><td>2/5 1/1</td><td>3</td><td>red</td></tr>
<tr><td></td><td></td><td>1.40</td><td>1/1 2/5</td><td>5</td><td>green</td></tr>
<tr><td>1</td><td>38</td><td>1.67</td><td>5/5 2/3</td><td>7-[1|2|3|4|5]</td><td>red</td></tr>
<tr><td></td><td></td><td>1.67</td><td>2/3 5/5</td><td>1-7</td><td>green</td></tr>
<tr><td></td><td></td><td>1.60</td><td>3/5 3/3</td><td>4-[1|2|3|4|5]</td><td>red</td></tr>
<tr><td>2</td><td>62</td><td>1.80</td><td>5/5 4/5</td><td>7-[1|2|3|4|5]-[1|2|3|4|5]</td><td>red</td></tr>
<tr><td></td><td></td><td>1.80</td><td>4/5 5/5</td><td>1-7-7</td><td>green</td></tr>
<tr><td></td><td></td><td>1.60</td><td>5/5 3/5</td><td>7-6-[1|2|3|4|5];
7-[1|2|3|4|5]-6</td><td>red</td></tr>
<tr><td>3</td><td>86</td><td>1.86</td><td>6/7 5/5</td><td>1-7-7-7</td><td>green</td></tr>
<tr><td></td><td></td><td>1.86</td><td>5/5 6/7</td><td>7-[1|2|3|4|5]-[1|2|3|4|5]-[1|2|3|4|5]</td><td>red</td></tr>
<tr><td></td><td></td><td>1.71</td><td>5/7 5/5</td><td>1-7-6-7; 1-7-7-6; 1-6-7-7</td><td>green</td></tr>
<tr><td>4</td><td>110</td><td>1.89</td><td>8/9 5/5</td><td>1-7-7-7-7</td><td>green</td></tr>
<tr><td></td><td></td><td>1.89</td><td>5/5 8/9</td><td>7-[1|2|3|4|5]-[1|2|3|4|5]-[1|2|3|4|5]
-[1|2|3|4|5]</td><td>red</td></tr>
<tr><td></td><td></td><td>1.78</td><td>7/9 5/5</td><td>1-7-7-6-7; 1-7-7-7-6; 1-6-7-7-7;
1-7-6-7-7</td><td>green</td></tr>
<tr><td>5</td><td>134</td><td>1.91</td><td>5/5 10/11</td><td>7-[1|2|3|4|5]-[1|2|3|4|5]
-[1|2|3|4|5]-[1|2|3|4|5]-[1|2|3|4|5]</td><td>red</td></tr>
<tr><td></td><td></td><td>1.91</td><td>10/11 5/5</td><td>1-7-7-7-7-7</td><td>green</td></tr>
<tr><td></td><td></td><td>1.82</td><td>9/11 5/5</td><td>1-7-7-7-7-6; 1-7-7-7-6-7;
1-6-7-7-7-7; 1-7-6-7-7-7; 1-7-7-6-7-7</td><td>green</td></tr>
</table>

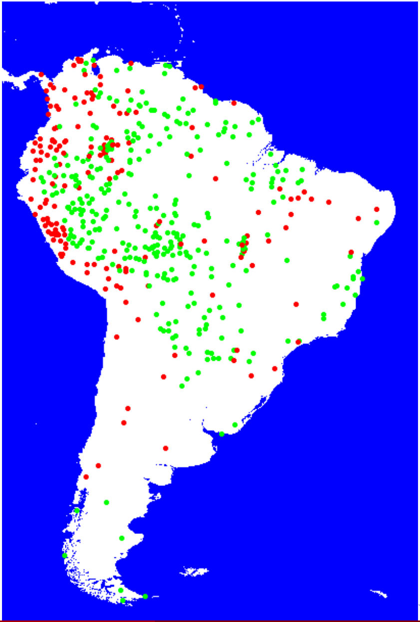


Algorithm for Proportion-Optimal Isoglosses: Analysis

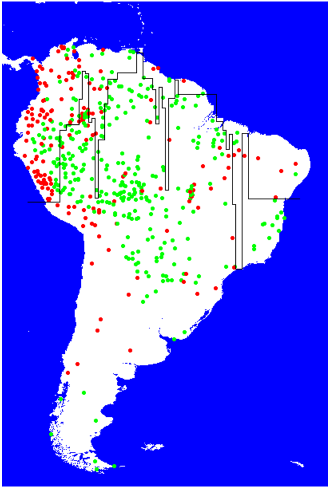
- Guaranteed to compute the proportion-optimal line
- The list R grows only polynomially in the size of the input
- Why?
  - When only the *counts* of correctly classified/total number of points are considered (not their positions) there are only polynomially many configuration
  - The size of the grid is $x \cdot y$
  - Every position in the grid is either empty or one of four possibilities: a ring on the rings side, a ring on the crosses side, a cross on the cross side or a cross of the rings side.
  - The number of ways to divide $x \cdot y$ squares into four different categories is $O(\binom{x \cdot y + 1}{3})$.

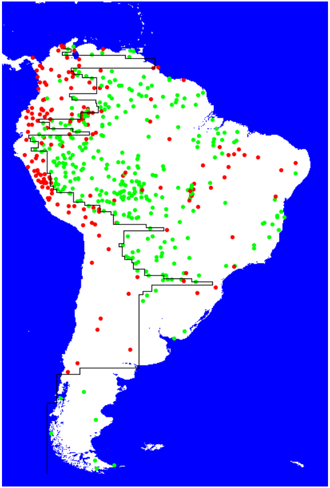
Real Example: Restricted Numeral Systems

- A numeral system is “restricted” iff
  - Monomorphemic numerals exist only up to 2 or 3 AND
  - Higher quantities are expressed orally only inexactly, or up to ca 10 with additions of 1, 2 and 3 (possibly including ad hoc use of ’hand’ for 5).
- Suppose we want to know if the border of restricted numeral system coincides with the extent of the Amazon forest



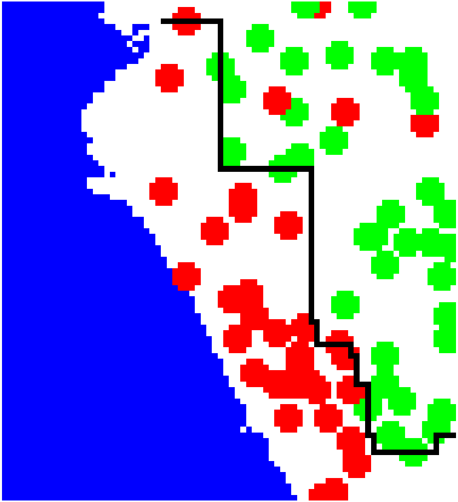
Optimal Isogloss Lines





- Absolute-optimal line east-west (left, 48 misclassifications)
- Absolute-optimal line north-south (right, 34 misclassifications).
- Proportion-optimal lines are nearly identical

Zoom on Andean-Amazon Divide



- The isogloss lines for the South American numeral data gives a fairly consistent Andean-Amazonian boundary
- But includes the Chaco and Southern Cone regions of South America in "Amazonian" part
- The border may be studies more closely for non-linguistic correlates to the boundary line

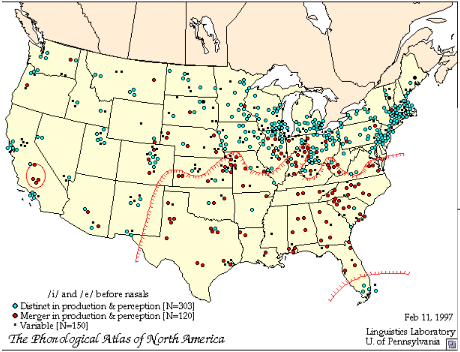
Extensions and Generalizations

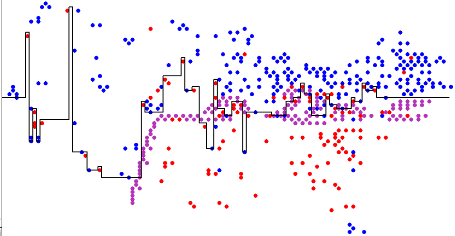
- More than one isogloss line?
  - Tractable
- More than two input colours?
  - Tractable
- More than one variable?
  - Possibly not tractable

More than one isogloss line?

- A natural extension is to draw more lines after the first one
  - Split the map into two parts according to the first line
  - Compute the optimal line in the two parts separately
  - Select the better scoring line of the two
- The more lines drawn the closer to a rote solution

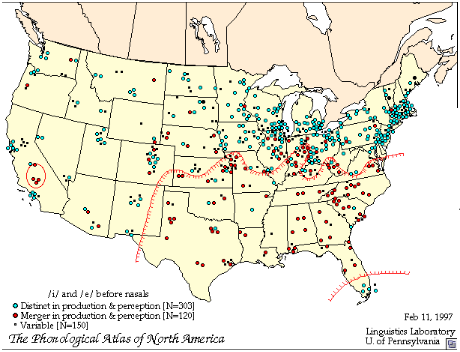
Example #1

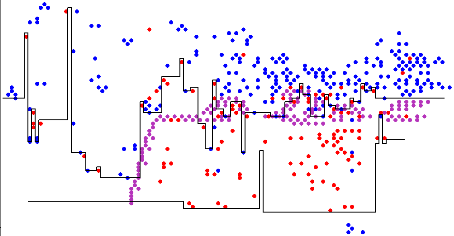




- Human drawn isogloss map has one major line, a smaller line at the bottom and a circle
- The major line is very similar to (both the absolute and proportion) optimal isogloss line

Example #2





- The smaller line at the bottom very similar to the 2nd optimal isogloss line
- The more lines drawn the closer to a rote solution

More than two input colours isogloss line?

- More than two input colours?
  - For n colours
  - Do n binarizations where all colours except one are merged
  - Compute the optimal isogloss line for the n binarizations separately
  - Keep the line which is optimal across the binarizations
- Proceed with the 2nd, 3rd, etc as per the previous slide, if needed

More than one variable?

- Suppose we have n different binary variables v_i defined for each language
- Suppose we want to find the line such that
  - The sum $\sum_{c_i}$ is maximized
  - Where c_i is the number of units of v_i that are well-classified by the line
- I suspect this introduces a complexity of $O(2^n)$ to the problem because of the indeterminacy over which is the majority side (bottom or top) for each v_i

Conclusions

- A natural definition of a the optimal isogloss line
- Algorithms for finding the optimal isogloss lines
- Impressionistically similar to isogloss lines drawn by human intuition
- Some theoretical and practical extensions unsolved/unexplored

Thank you