Morphology and dialectology in the Linguistic Survey of Scotland

A quantitative approach

Pavel Iosad
pavel.iosad@ed.ac.uk
Will Lamb
w.lamb@ed.ac.uk
Oilthigh Dhùn Èideann

Rannsachadh na Gàidhlig
Sabhal Mòr Ostaig
24th June 2016

Outline

- Motivation
  - State of the art
  - What is dialectometry?
  - Why dialectometry?
- LSS(G) data
- Three different analyses
  - Spatial analysis
  - Correlation analysis of features
  - Correlation analysis of varieties
- Conclusions

1 Background and motivation

1.1 State of the art

Current status

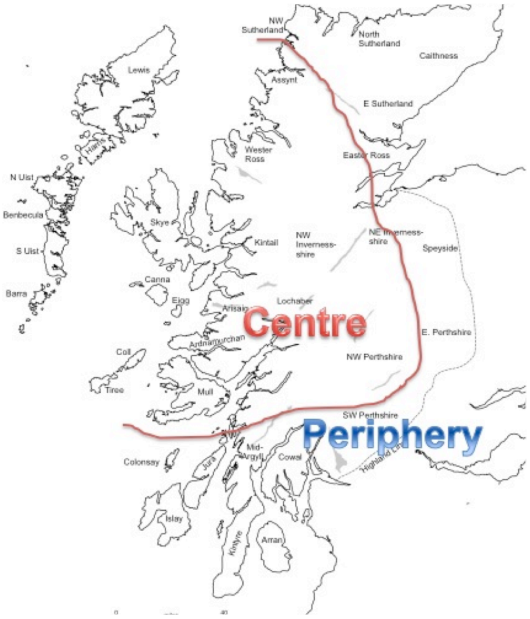
- Individual dialect descriptions
  - Pre-LSS: Borgstrøm (1937, 1940, 1941), Oftedal (1956), Holmer (1938, 1954, 1962)

  - Post-LSS: Mac Gill-Fhinnein (1966), Watson (1974), Dorian (1978), Ó Murchú (1989), Wentworth (2005)
- LSS(G) and SGDS (Ó Dochartaigh 1994–1997)
- Systematic dialectology
  - Individual features: Jackson (1967), Ó Maolalaigh (1996), Bosch & Scobbie (2009)
  - Macrodialectology: classic paper by Jackson (1968)

The division of Gaelic dialects

- Many scholars have made comments on dialectal divisions in Gaelic
- The approach is either purely historical (e. g. Jackson) or impressionistic
- No solid data:
  - SGDS exists for qualitative analysis, but not much work has been done with it
  - No quantitative data

'The *central dialect* covers the Hebrides as far south as Mull and sometimes further, Ross exclusive of the north-east corner, Assynt, Inverness-shire, western Perthshire, and mainland Argyll roughly north of Loch Awe; while the *peripheral dialects* comprise Caithness and Sutherland exclusive of Assynt, the north-east corner of Ross, Braemar, eastern Perthshire, the rest of mainland Argyll with Kintyre, and Arran. Moray and the adjacent lower region of the Spey, the wide valley of Strathspey from Rothiemurchus to the Moray border, may go with the peripheral dialects, linking up with Braemar and east Perth' (Jackson 1968: 67)



1.2 Dialectometric approach

What is dialectometry?

'Dialectometry studies dialects using exact methods, especially computational and statistical approaches' (Wieling & Nerbonne 2015)

- Focus on objective, quantitative methods
- Focus on aggregate measures not individual features
- 'Individual features are inevitably noisy'
- Covers both spatial variation and variation within a location

Common methods

- String distance (e. g. Levenshtein distance)
- Clustering methods (e. g. Ward clustering)
- Multidimensional analysis
- Correlation analysis
- Regression (including spatially adjusted methods)

Common applications

- Pronunciation distance
- Cluster analysis: alternative to traditional isoglosses
- Multidimensional analysis: identifying dialect areas from the data
- Mostly based on phonetic material!
- Wieling & Nerbonne (2015): not much has been done on morphosyntax, though increasing interest in recent years

Previous applications to Celtic

- Lexicostatistics: Elsie (1983–1984, 1986)
- Levenshtein distance for Irish dialects: Kessler (1995) based on LASID (Wagner 1958–1969): first ever application of the method to dialectology!
- Recent reevaluation for Irish by Ó Muircheartaigh (2014)
- Some work on Breton, see Brun-Trigaud, Solliec & Le Dû (2016) with references

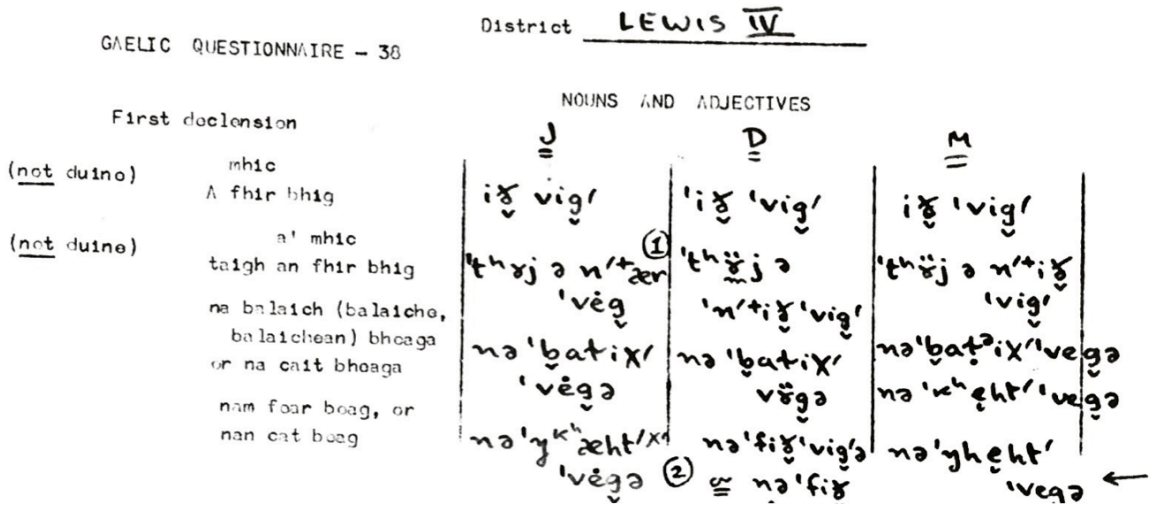
2 Data

2.1 LSS morphology data

Linguistic Survey: background

- Main collection period: 1951–63
  - Coverage very close to 18th century 'Highland Line'
  - Impressive given Jackson's famously strict criteria
- Questionnaire sections
  - Phonology: 893 headwords
    - Published as Ó Dochartaigh (1994–1997)
  - Morphophonology and syntax
    - 13.5 pages, unpublished

Example materials


GAELIC QUESTIONNAIRE – 38

District LEWIS IV

NOUNS AND ADJECTIVES

First declension

		J	D	M
(not duine)	mhic A fhir bhig	iɣ vigʲ	ˈiɣ ˈvigʲ	iɣ ˈvigʲ
(not duine)	a' mhic taigh an fhir bhig	ˈtʰɤj ə nʲtʲæɾ ① ˈvėg	ˈtʰɤ̈j ə ˈnʲtʲiɣ ˈvigʲ	ˈtʰɤ̈j ə nʲtʲiɣ ˈvigʲ
	na balaich (balaiche, balaichean) bhoaga or na cait bhoaga	nə ˈbatiXʲ ˈvėgə	nə ˈbatiXʲ vėgə	nəˈbatəiXʲˈvegə
	nam foar boag, or nan cat boag	nəˈɣᵏʰæht/Xᵗ ˈvėgə ②	nəˈfiɣˈvigʲə ≘ nəˈfiɣ	nəˈɣhęhtʲ ˈvegə ←

2.2 Our study

Coding

- Coded by hand from original field materials at the School of Scottish Studies Archives
  - 1 for presence of feature
  - 0 for absence of feature
  - Blank for no return
- Features coded using target phrase, asterisk marks feature of interest
- E. g. *na casan beag*a*: presence of suffix in feminine plural adjectives
  - 1 for *na casan beaga*
  - 0 for *na casan beag* or any other form
- Ongoing: mapping demographic data reporting in the LSS to census return to evaluate potential effects of language shift/obsolescence

Analysis

- All analysis conducted with R (R Core Team 2016)
- Methods
  - Generalized additive models with package `mgcv` (Wood 2006)
  - Cluster analysis with package `cluster` (Maechler et al. 2015)
  - Correlation analysis with R core function `cor` and `corrplot` package (Wei & Simko 2016)

3 Results

3.1 Spatial variation

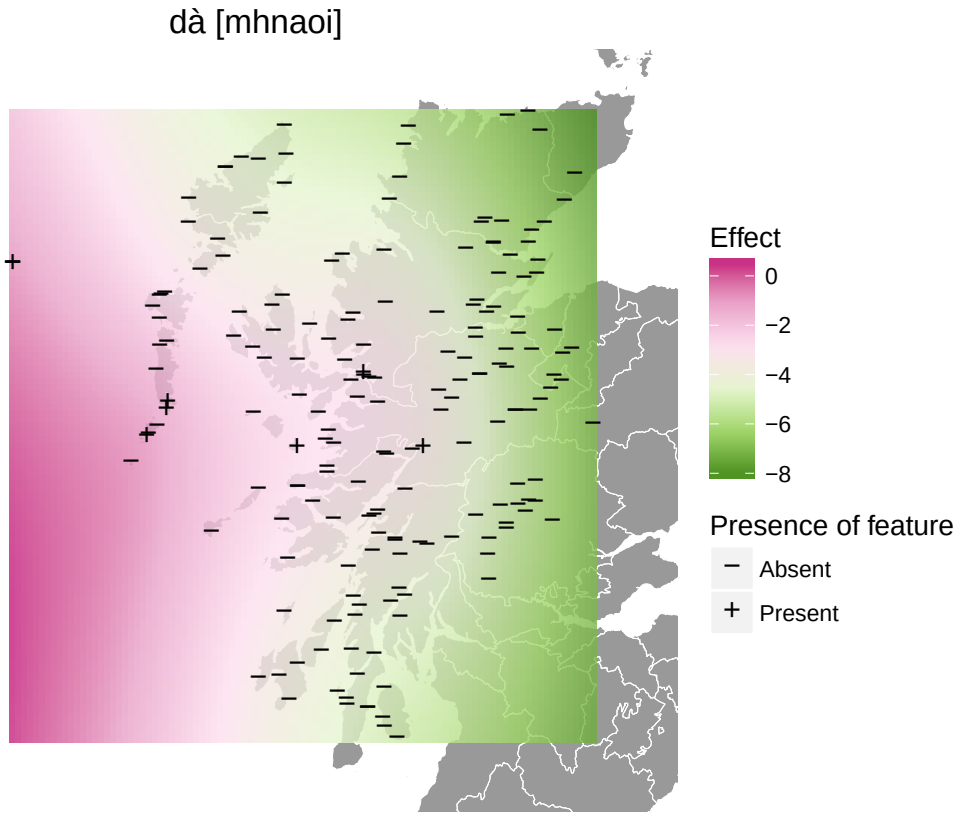
Method

- Logistic regression: probability of feature being present depending on latitude and longitude
- Non-linear regression: generalized additive models (Wood 2006)

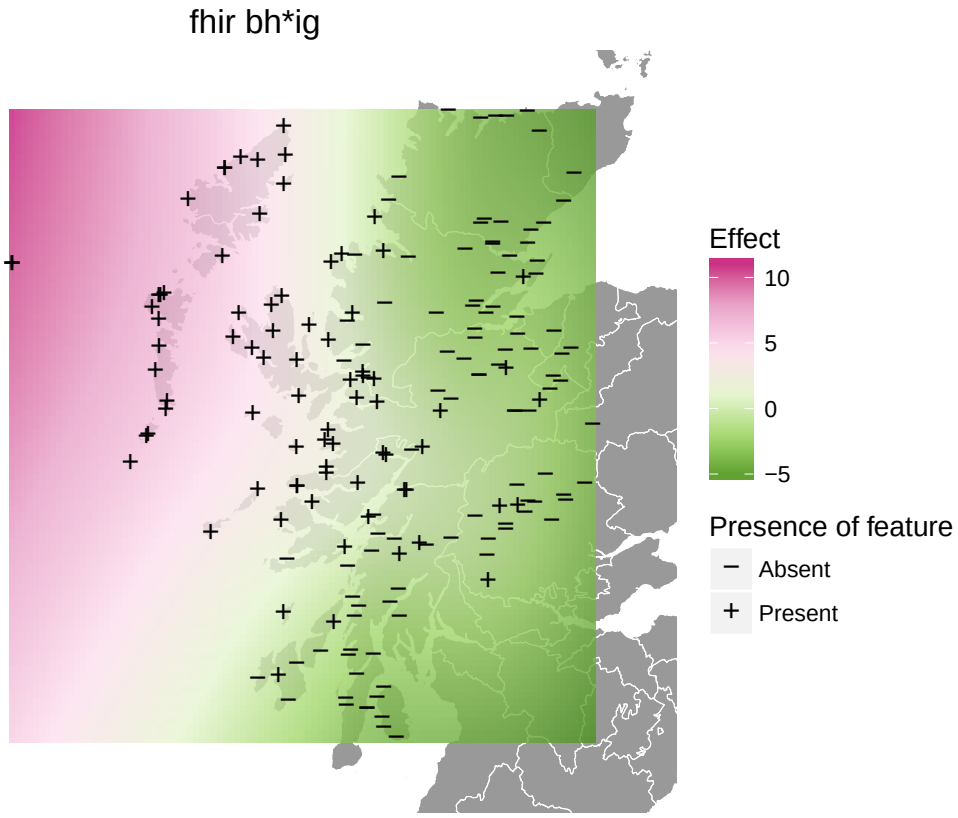
☞ Currently more a visualization method than a predictive analysis

- But can be combined with explanatory variables to adjust for them: current plan to do this with demographic data

Local pattern



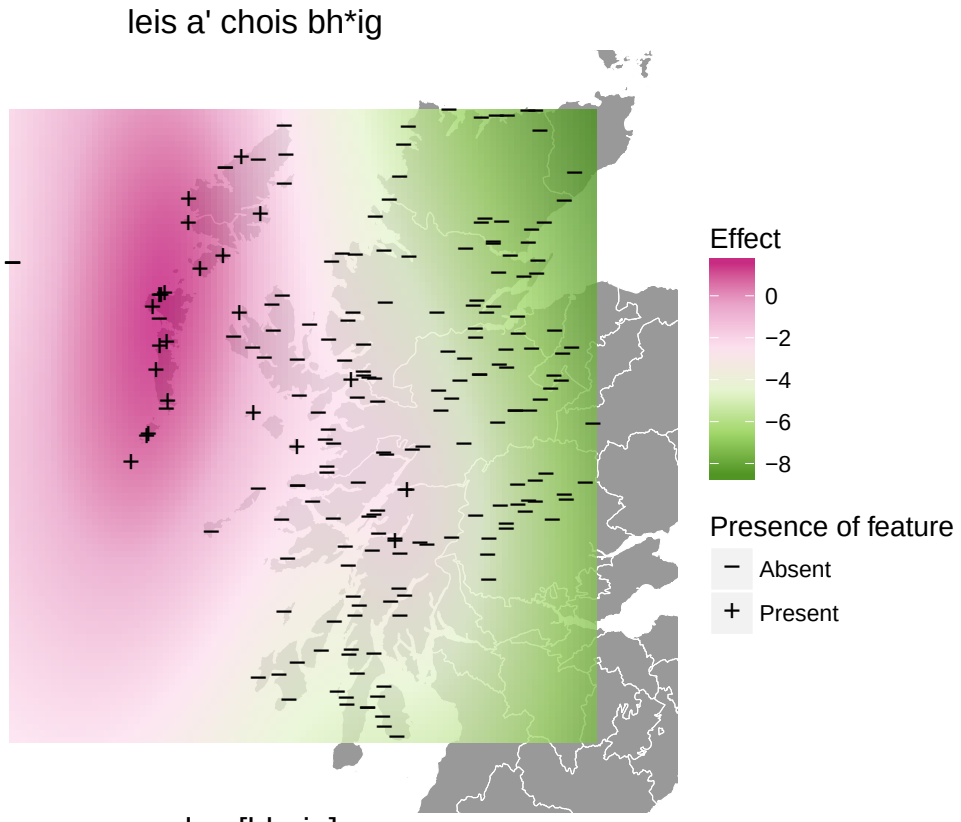
The cline

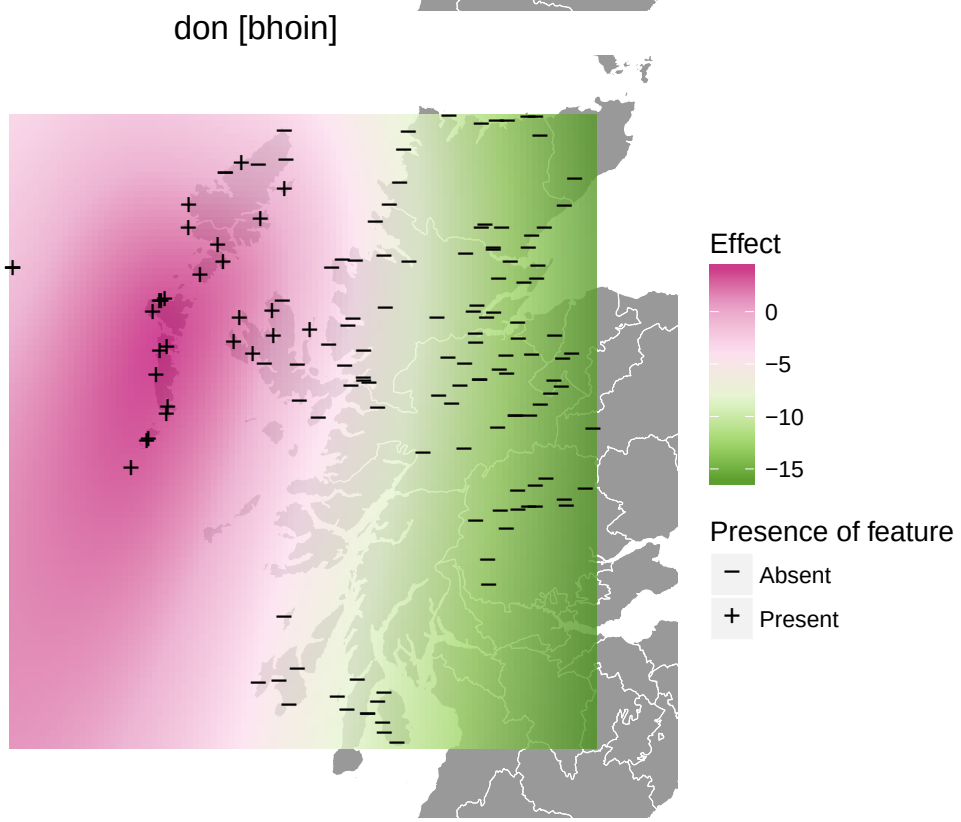


- The smoothing allows us to see the 'big picture'
- There is a southeast-northwest cline
- Could be related to language decline?

☞ Next steps: include demographic data as explanatory variable to adjust for it

The 'Uist-Barra' effect





- Lewis is often excluded
- Often conservative features

3.2 Correlation and clustering: dialects

Correlation analysis

- We can represent each dialect as a sequence of values (a vector)

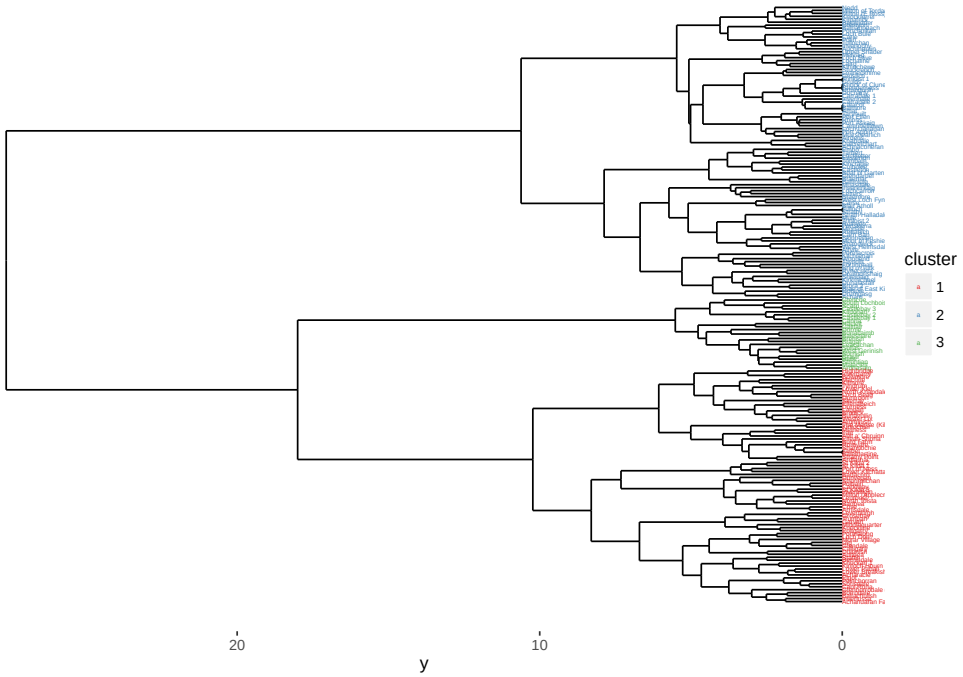
ID	Point	cas NOMlen Adj	sùil DATt Art 'leis an t-sùil'	fear VOClen N	cathair NOMlen Adj
1	Port of Ness	1	1	1	1
2	Upper Shader	1	1	1	1
3	Bragar	1	1	1	1
4	Carloway	1	1	1	1
5	Brenish	1	1	1	1
6	North Tolsta	1	1	1	1
7	Lower Pabail	1	1	1	1
8	Leurbost	1	1	1	1
9	Gravir	1	1	1	1
10	Scarp	1			
11	Ardhasaig	1	1		1
12	Grosebay	1		1	
13	Leverburgh	1	1		
14	St Kilda 1	1	1	1	1

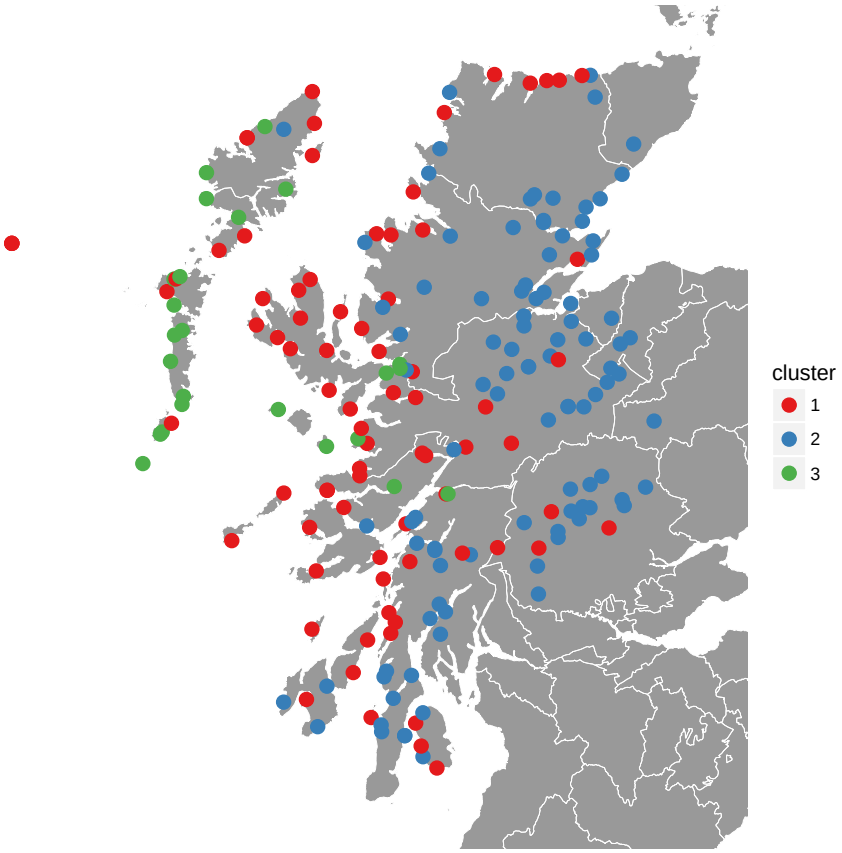
- Port of Ness = $\langle 1, 1, 1 \ldots \rangle$
- We can calculate the *correlation matrix* for a set of vectors
- The higher the correlation, the more similar the dialects are to each other
- ☞ A correlation of 1 means their behaviour is identical, a correlation of -1 means they are exact opposites

Cluster analysis

- Once we have a correlation matrix, we can rank the dialects in terms of how close they are to each other
- Based on this, we are able to conduct *clustering*
- Various methods: agglomerative Ward clustering is common
- We set the number of cuts to make in the tree
- Here: three clusters

Results





- Confirms some qualitative observations:
  - Cluster 3 (green): concentrated in Uist/Barra/Harris

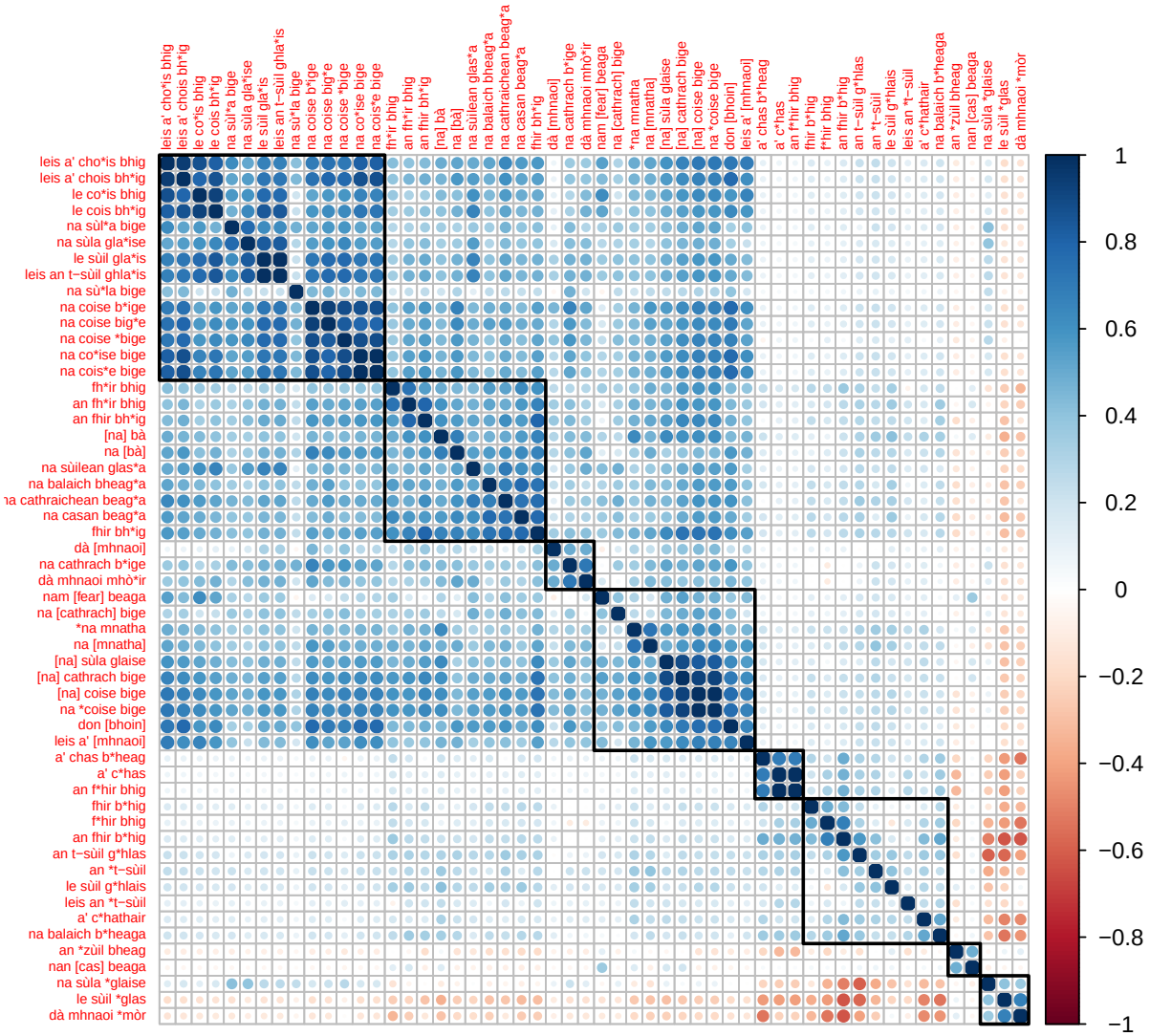
  - Cluster 2 (blue): periphery (correlation with strength of Gaelic?)
- More fine-grained analysis also possible

3.3 Correlation analysis: features

Correlation of features

- We can use the same technique to evaluate how similar the *features* are across dialects
- This can tell us about patterns of changes (and obsolescence)
- Adger (2016) suggests that simultaneous changes in apparently unrelated aspects of grammar may reveal the underlying unity of the grammatical mechanisms involved

Correlation plot



Genitive articles

- A set of correlated features is the use of *na* in the genitive
  - [*na*] *sùla glaise*
  - [*na*] *cathrach bige*
  - [*na*] *coise bige*
- Methodological sanity check
  - Different feminine lexical items lose the genitive form of the article together
  - Candidate for least surprising finding of the year, but this shows our data and methods produce at least some plausible results

Loss of lenition

- One very clear cluster is formed by 'core' lenition contexts:
  - *a' chas bheag*
  - *a' chas*
  - *an fhir*
- Lenition in these three contexts is lost simultaneously (in diatopic terms)
- But: no correlation with loss of lenition in some other contexts (e. g. (*a*) *fhir bhig*)
- No single grammatical mechanism for *all* lenition
- The simultaneity in these three contexts could show that they do reflect a single underlying mechanism

☞ See Iosad (2014) for similar reasoning on Breton spirantization

3.4 Conclusions and prospects

Conclusions

- A quantitative approach to Gaelic dialectology is possible and worthwhile
  - Produces plausible results
  - Allows us to ask new questions
- Potential for insights into diatopic variation beyond 'centre and periphery', with adjustment for other factors
- Potential for analytic insights into linguistic structure

Prospects

- Limitation of coding: currently all 0 cells are equal (count for similarity calculations) even if the forms are not identical

☞ This would need more detailed coding, but for many of our variables it doesn't really matter

- Add explanatory variables
- Combine with phonetic data (SGDS): stay tuned!
- Use insights gained to calibrate traditional/anecdotal knowledge of morphosyntactic variation: important for corpus planning (Bell et al. 2014)

References

Adger, David. 2016. Structure, use, and syntactic ecology in language obsolescence. MS., Queen Mary University of London.

Bell, Susan, Mark McConville, Wilson McLeod & Roibeard Ó Maolalaigh. 2014. *Dlùth is inneach: Linguistic and institutional foundations for Gaelic corpus planning*. Project report. Inverness: Bòrd na Gàidhlig.

Borgstrøm, Carl Hjalmar. 1937. The dialect of Barra in the Outer Hebrides. *Norsk tidsskrift for sprogvidenskap* 8. 71–242.

Borgstrøm, Carl Hjalmar. 1940. *The dialects of the Outer Hebrides* (A linguistic survey of the Gaelic dialects of Scotland 1). Norsk Tidsskrift for Sprogvidenskap, suppl. bind I. Oslo: Norwegian Universities Press.

Borgstrøm, Carl Hjalmar. 1941. *The dialects of Skye and Ross-shire*. Norsk tidsskrift for sprogvidenskap, suppl. bind II. Oslo: Norwegian Universities Press.

Bosch, Anna R. K. & James M. Scobbie. 2009. Fine-grained morpho-phonological variation in Scottish Gaelic: Evidence from the Linguistic Survey of Scotland. In James N. Stanford & Dennis R. Preston (eds.), *Variation in indigenous minority languages* (IMPACT: Studies in Language and Society 25), 347–368. Amsterdam: John Benjamins.

Brun-Trigaud, Guylaine, Tanguy Solliec & Jean Le Dû. 2016. A new dialectometric approach applied to the Breton language. In Marie-Hélène Côté, Remco Knooihuizen & John Nerbonne (eds.), *The future of dialects: Selected papers from Methods in Dialectology XV*, 135–154. Berlin: Language Science Press.

Dorian, Nancy C. 1978. *East Sutherland Gaelic*. Dublin: Dublin Institute for Advanced Studies.

Elsie, Robert W. 1983–1984. Lexicostatistics and its application to Brittonic Celtic. *Studia Celtica* 18/19. 110–127.

Elsie, Robert W. 1986. *Dialect relationships in Goidelic*. Hamburg: Helmut Buske Verlag.

Holmer, Nils M. 1938. *Studies on Argyllshire Gaelic* (Skrifter utgivna av Kungliga Humanistiska Vetenskapssamfundet i Uppsala 31). Uppsala: Almqvist & Wiksell.

Holmer, Nils M. 1954. *The Gaelic of Arran*. Dublin: Dublin Institute for Advanced Studies.

Holmer, Nils M. 1962. *The Gaelic of Kintyre*. Dublin: Dublin Institute for Advanced Studies.

Iosad, Pavel. 2014. The phonology and morphosyntax of mutation in Breton. *Lingue e linguaggio* 13(1). 23–42.

Jackson, Kenneth Hurlstone. 1967. Palatalisation of labials in the Gaelic languages. In Wolfgang Meid (ed.), *Beiträge zur Indogermanistik und Keltologie: Julius Pokorny zum 80. Geburtstag gewidmet*, 179–192. Innsbruck: Sprachwissenschaftliches Institut der Universität Innsbruck.

Jackson, Kenneth Hurlstone. 1968. The breaking of original long *ē* in Scottish Gaelic. In James Carney & David Greene (eds.), *Celtic studies: Essays in honours of Angus Matheson*, 65–71. London: Routledge.

Kessler, Brett. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics* (EACL '95), 60–66. San Francisco: Morgan Kaufmann.

Mac Gill-Fhinnein, Gordon. 1966. *Gàidhlig Uidhist a Deas*. Baile Átha Cliath: Institiúid Ard-Léinn Bhaile Átha Cliath.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert & Kurt Hornik. 2015. `cluster`*: Cluster Analysis Basics and Extensions*. Version 2.0.3.

Ó Dochartaigh, Cathair (ed.). 1994–1997. *Survey of the Gaelic dialects of Scotland*. Dublin: Dublin Institute for Advanced Studies.

Ó Maolalaigh, Roibeard. 1996. The development of eclipsis in Gaelic. *Scottish Language*. 158–173.

Ó Muircheartaigh, Peadar. 2014. *Gaelic dialects present and past: A study of modern and medieval dialect relationships in the Gaelic languages*. Edinburgh: The University of Edinburgh PhD dissertation.

Ó Murchú, Máirtín. 1989. *East Perthshire Gaelic: Social history, phonology, texts and lexicon*. Dublin: Dublin Institute for Advanced Studies.

Oftedal, Magne. 1956. *The Gaelic of Leurbost, Isle of Lewis* (A linguistic survey of the Gaelic dialects of Scotland 3). Oslo: W. Aschehoug & Co.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Version 3.2.4. R Foundation for Statistical Computing. Vienna, Austria. `http://www.R-project.org/`.

Wagner, Heinrich. 1958–1969. *Linguistic atlas and survey of Irish dialects*. Dublin: Dublin Institute for Advanced Studies.

Watson, Seosamh. 1974. A Gaelic dialect of north-east Ross-shire: The vowel system and general remarks. *Lochlann* 6. 9–90.

Wei, Taiyun & Viliam Simko. 2016. *`corrplot`: Visualization of a Correlation Matrix*. Version 0.77.

Wentworth, Roy Graham. 2005. *Rannsachaidh air fòn-eòlas dualchainnt Ghàidhlig Gheàrrloch, Siorrachd Rois*. Dublin: Dublin Institute for Advanced Studies. `http://www.celt.dias.ie/publications/online/gearrloch/`.

Wieling, Martijn & John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics* 1(1). 243–264.

Wood, Simon N. 2006. *Generalized additive models: An introduction with R*. Chapman & Hall/CRC.

Appendix: Jackson on fieldwork

- On change
  - ☞ 'Remarkable that AF' s case system is so much more decayed [...] though he is the same age as JM. AM has kept it well'
- On omissions
  - ☞ 'Really this sort of thing would try the patience of a saint. Particularly since Barra is especially interesting in preserving the forms of the adjective rather well!'
- On informants
  - ☞ 'An ideal informant, a first-rate mind with natural flair for analysis. Hardly literate in Gaelic. Does not now use Gaelic much actively.'
  - ☞ 'Struck me as a crude and uneducated old man but this questionnaire suggests rather that he knows written Gaelic.'
  - ☞ '[an] ideal informant: totally unsophisticated'