Laura Calaza Díaz / Soraya Suárez Quintas / Rosa M. Crujeiras / Alberto Rodríguez Casal / Xulio Sousa / José Ramón Ríos Viqueira (2015): “A method for processing perceptual dialectology data”. *XII Congreso Galego de Estatística e Investigación de Operacións.* Lugo, 22-24 de outubro de 2015

A method for processing perceptual dialectology data

Laura Calaza Díaz[1] Soraya Suárez Quintas[2] Rosa M. Crujeiras [1]
Alberto Rodríguez Casal[1] Xulio Sousa[2] Jose Ramón Ríos Viqueira[3]

[1]Departamento de Estatística e Investigación Operativa. Universidade de Santiago de Compostela.

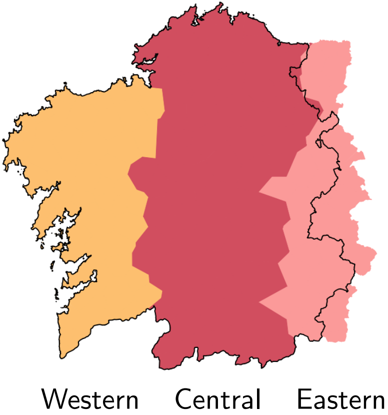
[2]Instituto da Lingua Galega. Universidade de Santiago de Compostela

[3]COGRADE. CITIUS. Universidade de Santiago de Compostela.

TecAnDaLi, Tecnoloxías e Análise dos Datos Lingüísticos

What is Perceptual Dialectology?

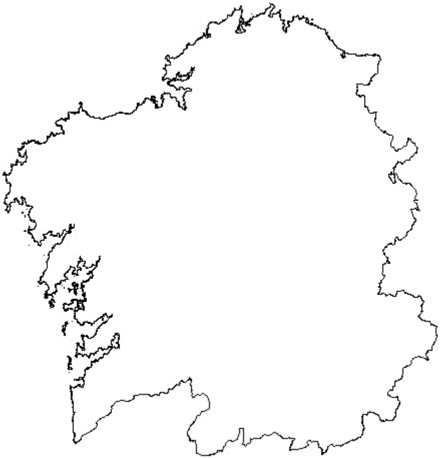
Academics postulate...



Xulio C. Sousa Fernández (2006)
Análise dialectométrica das variedades xeolingüísticas galegas. *Encontro de estudos dialectolóxicos*. Actas, M C. Rola Bernardo / H. Mateus Montenegro, Ponta Delgada: Instituto Cultural de Ponta Delgada, 345-362 - Capítulo de libro

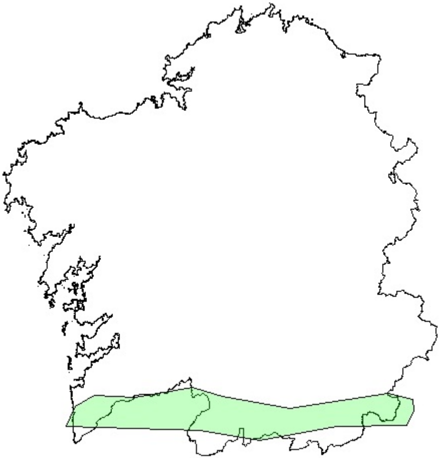
What is Perceptual Dialectology?

What I really perceive...



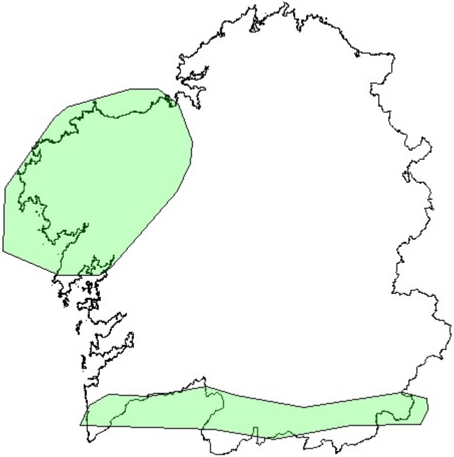
What is Perceptual Dialectology?

What I really perceive...



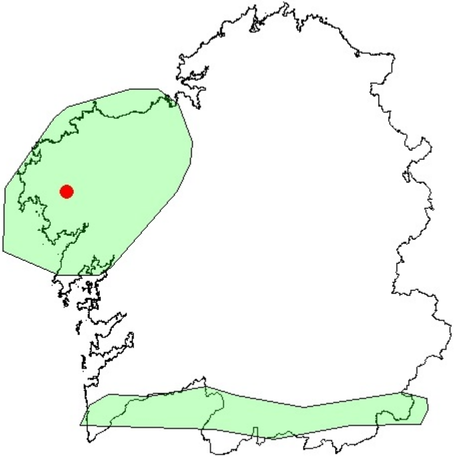
What is Perceptual Dialectology?

What I really perceive...



What is Perceptual Dialectology?

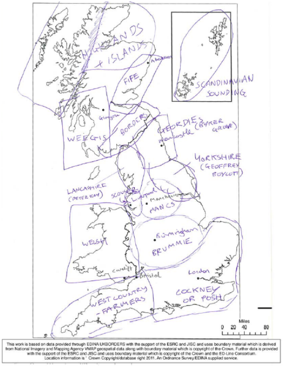
What I really perceive...



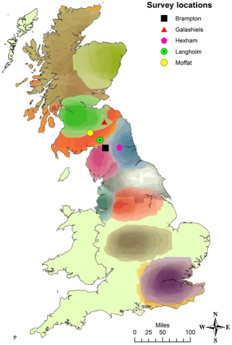
Previous studies

Analyzed issues:

- Data collection (scanning)
- Data visualization (heat maps)



This work is based on data provided through EDINA UKBORDERS with the support of the ESRC and JISC and uses boundary material which is derived from National Imagery and Mapping Agency VMAP geospatial data along with boundary material which is copyright of the Crown. Further data is provided with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown and the ED-Line Consortium. Location information is ' Crown Copyright/database right 2011. An Ordnance Survey/EDINA supplied service.



Montgomery, C. and Stoeckle, P. (2013)
Geographic information systems and perceptual dialectology: a method for processing draw-a-map data. *Journal of Linguistic Geography*, 1, 52–85.

Our purpose

Perceptions of Galician dialects

Our purpose

Perceptions of Galician dialects

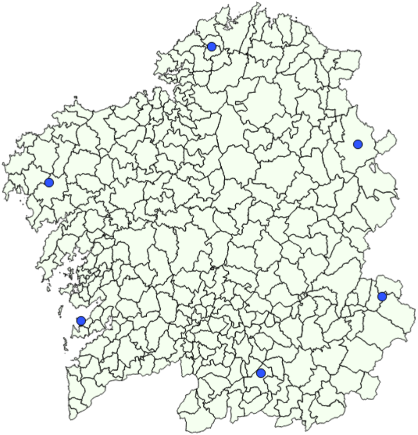
Steps...

- Collect and visualize Perceptual Dialectology data
- Discover if people are aware of and recognize regional variations of Galician language.
- Identify factors which influence geographical varieties of Galician language recognition.
- Assess in which way people's perceptions correspond to the geo-linguistic varieties traditionally recognised in Galician studies

Survey design

Material for survey:

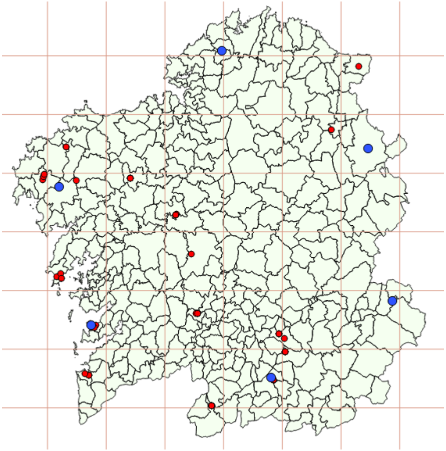
- 7 auditions (from different locations)



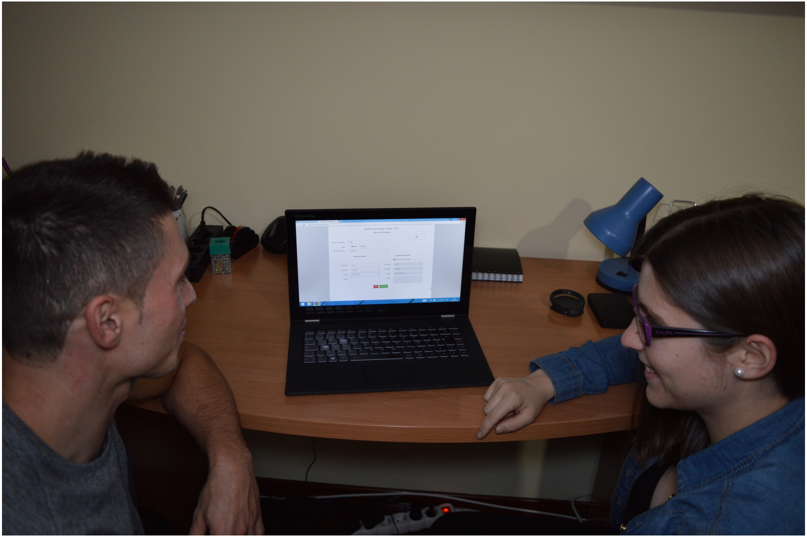
Survey design

Material for survey:

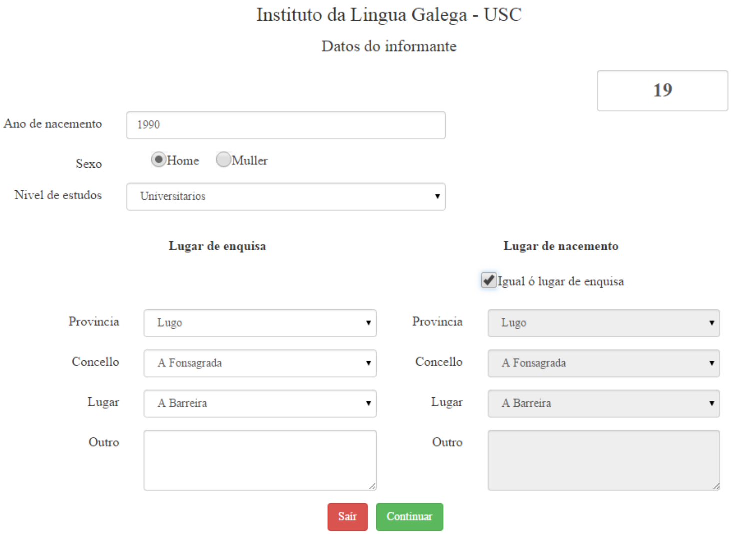
- 50 informants (from different locations)



Data collection



Data collection



Instituto da Lingua Galega - USC

Datos do informante

19

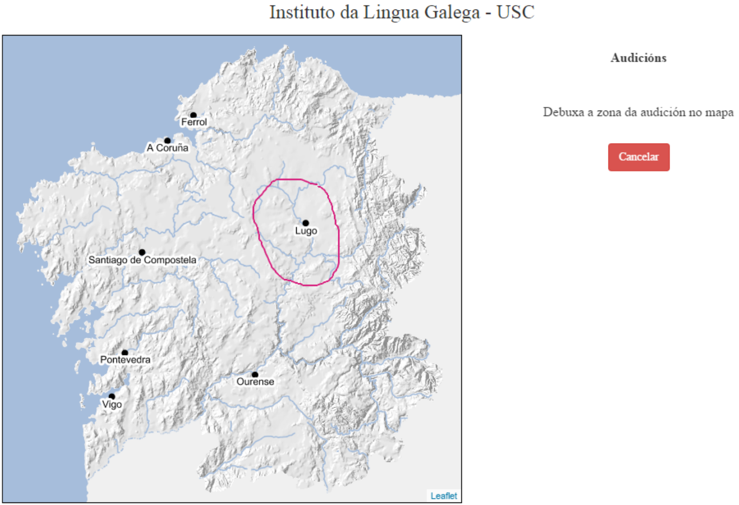
Ano de nacemento: 1990

Sexo: ● Home ○ Muller

Nivel de estudos: Universitarios ▾

Lugar de enquisa		Lugar de nacemento	
		☑ Igual ó lugar de enquisa	
Provincia	Lugo ▾	Provincia	Lugo ▾
Concello	A Fonsagrada ▾	Concello	A Fonsagrada ▾
Lugar	A Barreira ▾	Lugar	A Barreira ▾
Outro		Outro	

Data collection

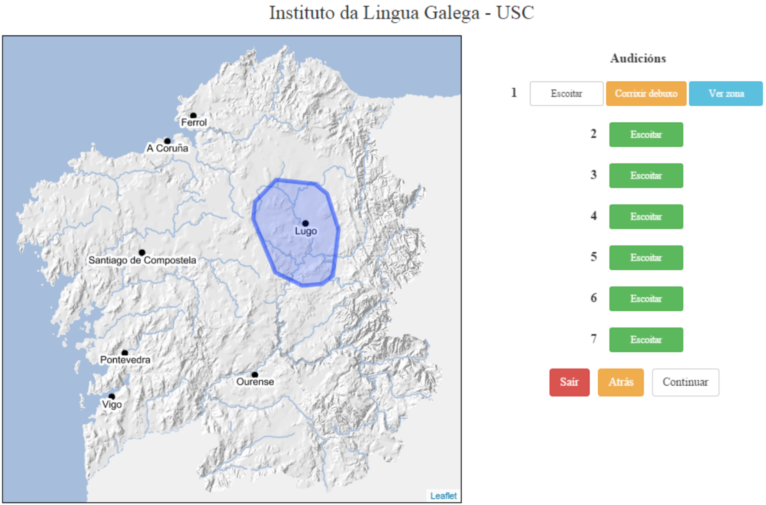


Instituto da Lingua Galega - USC

Audicións

Debuxa a zona da audición no mapa

Data collection



Instituto da Lingua Galega - USC

Audicións

Data collection

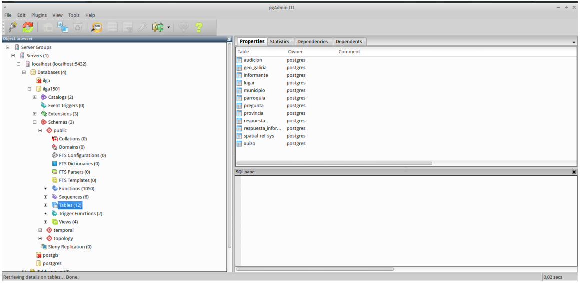
Database summary:

- Audition locations
- Respondents information:
  - gender
  - age
  - educational level
  - birthplace
  - 7 dialects geographic perceptions
- 3 divisions of Galician map (Western, Central, Eastern)

How to process information?

Data acquisition:

PostgreSQL (Structured Query Language)


pgAdmin III
File Edit Plugins View Tools Help
Object browser
Properties Statistics Dependencies Dependents
SQL pane

```
Server Groups
  Servers (1)
    localhost (localhost:5432)
      Databases (4)
        ilga
        ilga1501
          Catalogs (2)
          Event Triggers (0)
          Extensions (3)
          Schemas (3)
            public
              Collations (0)
              Domains (0)
              FTS Configurations (0)
              FTS Dictionaries (0)
              FTS Parsers (0)
              FTS Templates (0)
              Functions (1050)
              Sequences (6)
              Tables (12)
              Trigger Functions (2)
              Views (4)
            temporal
            topology
          Slony Replication (0)
        postgis
        postgres
      Tablespaces (2)
```

Table	Owner	Comment
audicion	postgres	
geo_galicia	postgres	
informante	postgres	
lugar	postgres	
municipio	postgres	
parroquia	postgres	
pregunta	postgres	
provincia	postgres	
respuesta	postgres	
respuesta_infor...	postgres	
spatial_ref_sys	postgres	
xuizo	postgres	

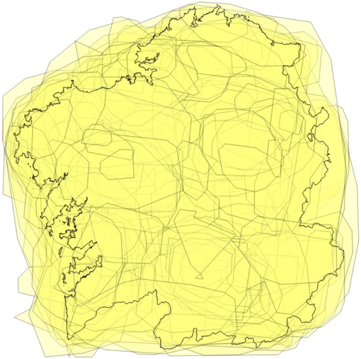
Retrieving details on tables... Done. 0,02 secs

Possibilities:

- Geographic Information System (GIS) ⇒ QGIS
- R project ⇒ Package `RPostgreSQL`

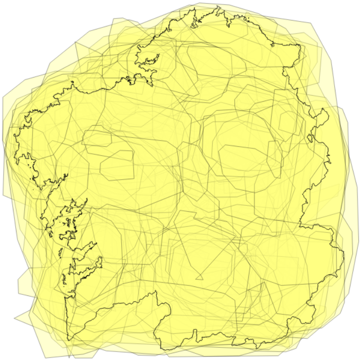
Data Visualization - QGIS

Respondents' selections

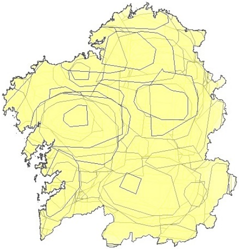


Data Visualization - QGIS

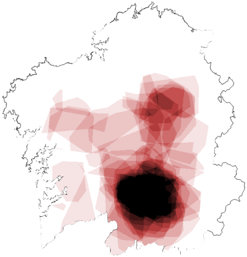
Respondents' selections



Intersection of perceptions

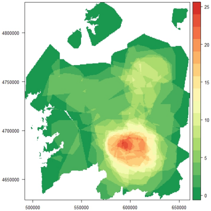


Heat map (1st audition)

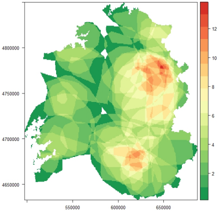


Data Visualization - R (rasterVis)+ QGIS

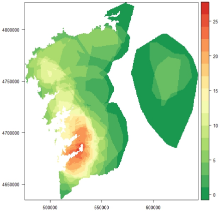
Audition 1



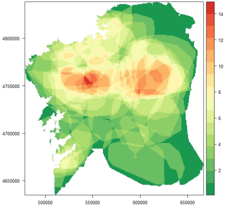
Audition 2



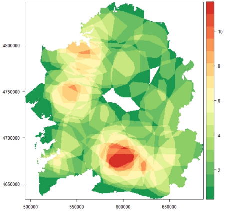
Audition 3



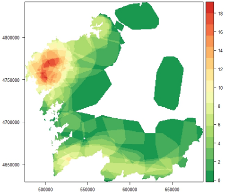
Audition 4



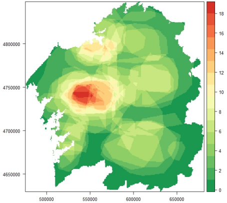
Audition 5



Audition 6



Audition 7



How to process information?

Handling geographical data to:

identify which factors influence the recognition of dialectal varieties

comparison with data from traditional dialectology and dialectometric studies

How to process information?

Handling geographical data to:

identify which factors influence the recognition of dialectal varieties

comparison with data from traditional dialectology and dialectometric studies

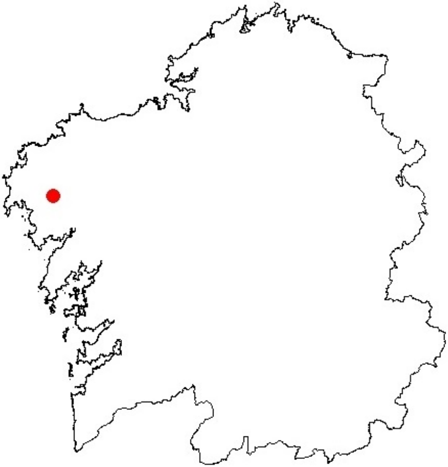
⇓

Assess diferences between sets

How to process information?

Assess diferences between sets

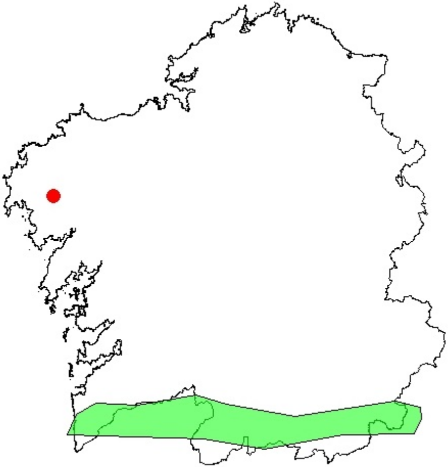
1st case:



How to process information?

Assess diferences between sets

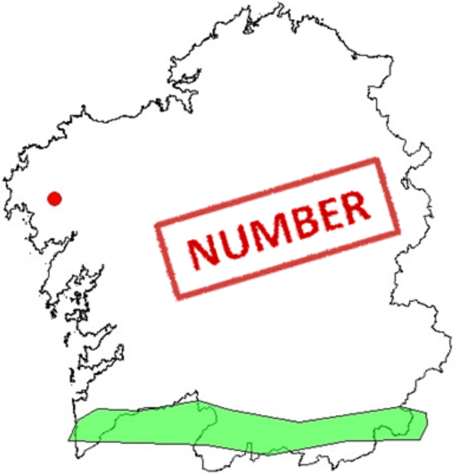
1st case:



How to process information?

Assess diferences between sets

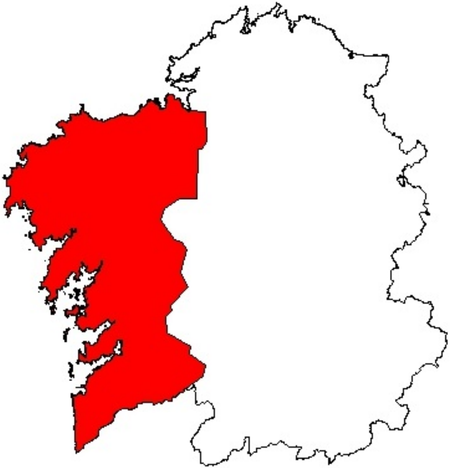
1st case:



How to process information?

Assess diferences between sets

2^{nd} case:

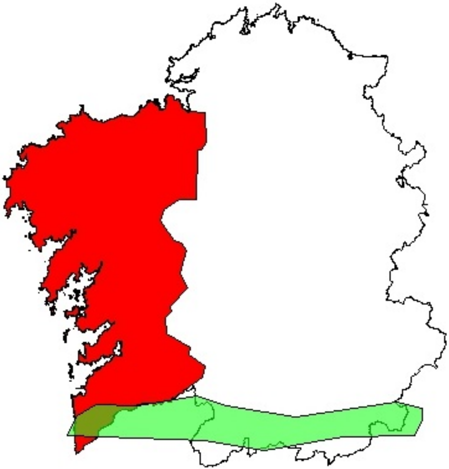


Western

How to process information?

Assess diferences between sets

2^{nd} case:

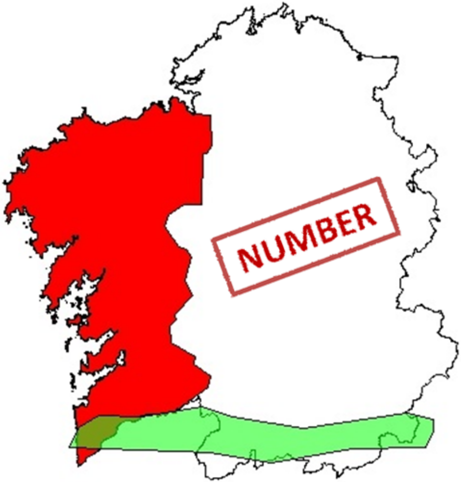


Western

How to process information?

Assess diferences between sets

2nd case:



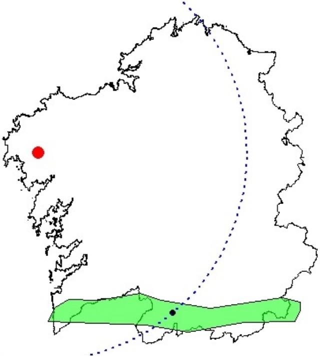
Western

How to process information?

Alternatives of distances between sets:

- **Centroid**

  Example (1st case)



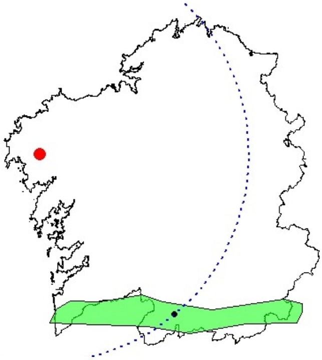
distance=140297.9

How to process information?

Alternatives of distances between sets:

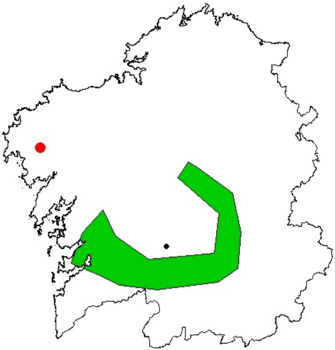
- **Centroid**

Example (1st case)



distance=140297.9

Example (2nd case)



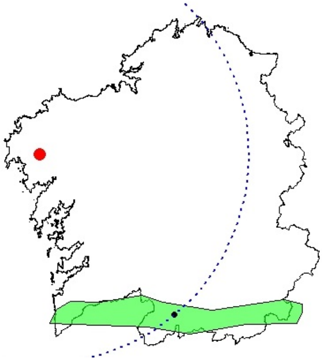
distance=102477.7

How to process information?

Alternatives of distances between sets:

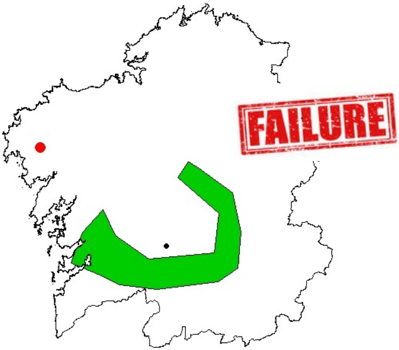
- ⋆ **Centroid**

Example (1st case)



distance=140297.9

Example (2nd case)



distance=102477.7

How to process information?

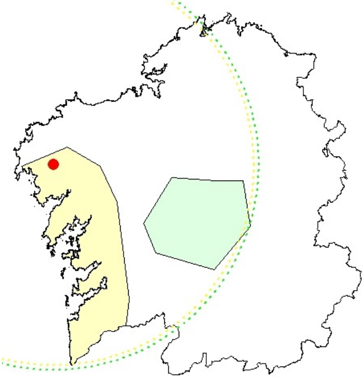
Alternatives of distances between sets:

- ⋆ **Centroid**
- ⋆ **Hausdorff**

Let be the sets $A, B \subseteq X$, then

$$H(A,B) = \sup_{x \in A} |d(x,A) - d(x,B)|$$

with $d(x,A) = \inf\{\rho(x,a), a \in A\}$, and $\rho(x,y)$ the distance between any two pixels $x, y \in X$



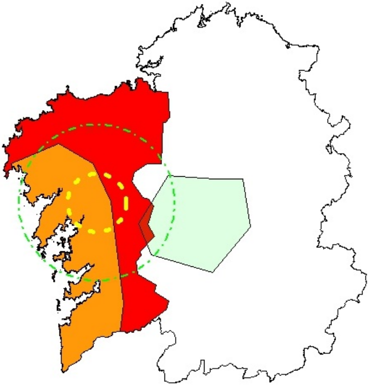
How to process information?

Alternatives of distances between sets:

- ⋆ **Centroid**
- ⋆ **Hausdorff**
- ⋆ **Baddeley**

$$\Delta_b(A, B) = \left[\frac{1}{n(X)} \sum_{x \in X} |d^*(x, A) - d^*(x, B)|^p\right]^{1/p},$$

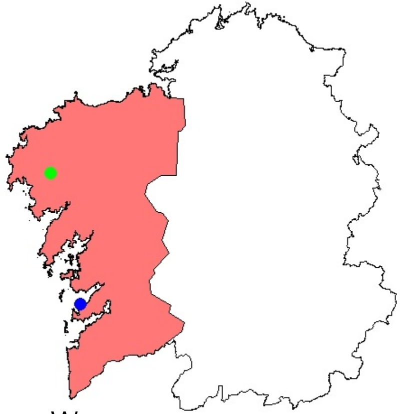
here $d^*(x, A) = \min\{d(x, A), c\} = \min\{\inf[d(x, a), a \in A], c\}$



Baddeley, A. J. (1992)
An error metric for binary images. *Robust Computer Vision: Quality of Vision Algorithms*, W. Förstner and S. Ruwiedel (Eds.), Karlsruhe, Wichmann, pp. 59–78.

Comparison with traditional dialectology

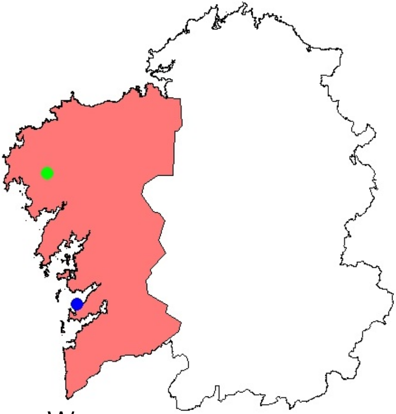
1^{st} objective: comparison with data from traditional dialectology and dialectometric studies



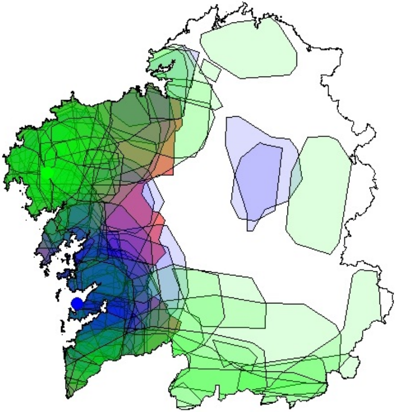
Western

Comparison with traditional dialectology

1st objective: comparison with data from traditional dialectology and dialectometric studies

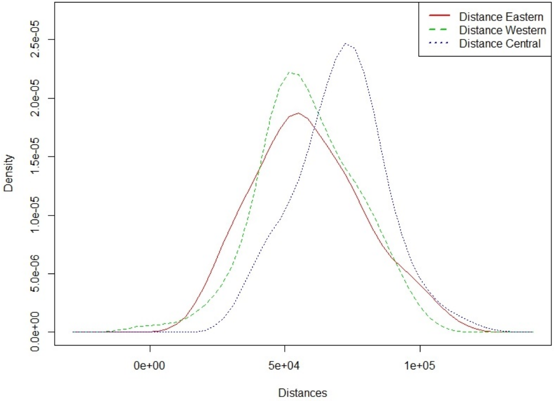


Western



Comparison with traditional dialectology

1st objective: comparison with data from traditional dialectology and dialectometric studies



- Normality tests (p.values> 0.05)

  Eastern mean: 55019.33 m
  Western mean: 54227.32 m
  Central mean: 71083.86 m

- Comparison of univariate density estimates:
  test equality of distributions
  p.value=0

Regression model

2^{nd} **objective: identify which factors influence in the recognition of dialectal varieties**

For each audition:

- Response: distance between the audition location and respondents' selections.
- Explanatory variables:
  - ⋆ gender
  - ⋆ age
  - ⋆ educational level
  - ⋆ distance between birthplace and audition location
  - ⋆ distance between birthplace and survey location

all models have been fitted and validated

Regression model

2$^{\text{nd}}$ **objective: identify which factors influence in the recognition of dialectal varieties**

For each audition:

- Response: distance between the audition location and respondents' selections.
- Explanatory variables:
  - ⋆ gender
  - ⋆ age
  - ⋆ educational level
  - ⋆ distance between birthplace and audition location
  - ⋆ distance between birthplace and survey location

all models have been fitted and validated

Regression model

2^{nd} objective: identify which factors influence in the recognition of dialectal varieties

To generalize...

$$Y = X\beta + \epsilon,$$

with,

$Y = \sum_{i=1}^{6} w_i Y_{ik}, w_i = \frac{1}{n}, \forall i = 1, \ldots, 6, k = 1, \ldots, n$

X includes:

- gender
- age
- educational level
- distance between survey place and birthplace

$\epsilon =$ error term

Results

```
lm(formula = baddeleydista ˜ estudos, data = basepromedio)

Residuals:
     Min       1Q   Median       3Q      Max
-21386.6  -5132.4    299.7   5512.6  23109.4

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)              59691       2693  22.168  < 2e-16 ***
estudosSECUNDARIOS       -5323       3517  -1.514  0.13682
estudosUNIVERSITARIOS   -11372       3376  -3.369  0.00151 **


Residual standard error: 9328 on 47 degrees of freedom
Multiple R-squared:  0.201,     Adjusted R-squared:  0.167
F-statistic: 5.911 on 2 and 47 DF,  p-value: 0.005132
```

Future research lines

- **Mixed-Effects models**

  To introduce random effects for: informants and auditions

- **(Spatial) correlation between sets**

  To avoid the simplification of our database

thanks
GRAZAS
gracias
GRACIAS
THANKS
grazas