
Laura Calaza Díaz / Soraya Suárez Quintas / Rosa M. Crujeiras / Alberto Rodríguez Casal / Xulio Sousa / José R. Ríos Viqueira (2015): "A method for processing perceptual dialectology data", en *ACTAS XII Congreso Galego de Estatística e Innovación de Operacións. Lugo, 22-23-24 de outubro de 2015.* Lugo: Servizo de Publicacións da Deputación de Lugo / SGAPEIO, 282-291.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 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] e José 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.

ABSTRACT

This work presents a contribution to the question of how to develop data processing (both technically & statistically) in perceptual dialectology (PD) studies. This discipline focuses on the study of non-linguists perceptions of dialectal variation. Such tasks require a large amount of information whose treatment can be improved if suitable tools from Geographic Information Systems (GIS) for the treatment of geographic information are combined with powerful statistical data analysis methods. In order to carry out this work, simulated data, corresponding to a case study from PD research from Galicia, has been used to demonstrate the procedure.

Keywords phrases: Baddeley's distance, dialectometry, perceptual dialectology, QGIS.

1. MOTIVATION

The *Tecnoloxías e Análise dos Datos Lingüísticos* (Technologies and Analysis of Linguistic Data) network[1] was created to encourage collaborative work among linguistic, statistical and computational scientific fields. The interdisciplinary work among these areas arises from the firm desire to promote iniciatives to strengthen knowledge transfer and research results.

In particular, during the last decade special efforts have been made in the field of Perceptual Dialectology (PD), a discipline regarded as a subset of folk linguistics and related to dialectology and sociolinguistics. PD research consists of an analysis of non-linguists knowledge, beliefs and attitudes about dialectal variation. This kind of research has aroused the interest of many dialectologists worldwide. However, there is still no detailed analysis of the perceptions of non-linguists concerning dialectal variation in Galicia. Therefore, our study, which intends to discover and analyse Galician speakers beliefs and opinions regarding dialectal variation, aims to fill a lacuna in Galician language studies. Specifically, our PD survey attempts to:

- Discover if people are aware of and recognize regional variations of the Galician language.
- Identify factors which influence geographical varieties of Galician language recognition.
- Assess in which way peoples perceptions correspond to the geo-linguistic varieties traditionally recognised in Galician studies (Fernández Rei, 1990; Sousa, 2006).

The main goals of this work are therefore twofold: the identification and determination of geographical positions of different Galician dialects, and of subjective judgments regarding the different features each dialectal variety has.

[1]TecAnDaLi (`http://ilg.usc.es/tecandali/`)

This work was partly funded by TecAnDaLi (which is co-funding by the European Regional Development Fund (FEDER) under the Galician Operational Programme 2007-2013).

Such tasks require information collection about respondents in a defined geographic region. We must therefore be able to produce a large database, which collects, among other information, respondents' perceptions maps. This means that data from PD studies must be processed in a computerized way. In addition, this type of studies requires intensive field work which in turn requires that information be gathered in a cost-efficient, practical and portable manner.

The main aim of this communication is to present techniques which allow for the automated collection of PD data and to introduce statistical tools in the PD field which enable the realization of proper statistical analysis.

2. METHODS

2.1. *Theoretical foundation*

PD is a linguistic discipline that focuses on the study of how speakers perceive variation in language. Unlike traditional dialectology, which is almost exclusively concerned with identifying and analysing the variables and variants that define the geographical varieties of a language, this area of study focuses on analysing the way in which people perceive geo-linguistic variation. What is therefore interesting about PD is its focus on non-linguists' insights into language (folk linguistics). Studies within this framework reverse traditional roles, bringing to the foreground speakers knowledge about language variation in the spatial dimension.

PD studies began in the early twentieth century in Holland, where the first dialect maps based on speakers' geolinguistic perceptions appeared. Until the mid-twentieth century, this country and Japan played the main role in studies within this field. However, it was not until the 1980s that this discipline was developed. At that time, Dennis Preston (1989) reviewed and modernized the study of speakers perceptions on geolinguistic variation, redefining the field and adapting it to modern linguistics. Since Preston, who is regarded today as the most representative figure of perceptual dialectology, many studies have emerged which are inspired by some of the five techniques developed by him (Preston, 1999) for PD studies:

- Draw-a-Map: informants have to draw borders in a blank map, identifying the locations where they believe different dialects exist.
- Degree of Difference: speakers are asked to rate the similarity or difference between two dialectal regions.
- Correct and Pleasant: informants rate regions according to how correct or pleasant they think the variants spoken there are.
- Dialect identification: informants listen to recorded speech samples from a given dialect continuum and must assign each voice to a specific territorial area.
- Qualitative data: respondents must answer a series of open questions related to the identifying features of the different dialects they recognized, connected with the characteristics that they attribute to the speakers of each one of those dialects.

In relation to these five methodological possibilities, there are at least two basic lines of PD work. One of the alternatives focuses on a draw-a-map task for dialectal identification in order to detect places where people recognize different perceptual varieties (geographical perspective), whilst the other approach considers qualitative and quantitative descriptive analysis (avaliative perspective). Traditionally, they are treated separately. Therefore, the innovative aspect of our study lies in the combination of the methods offered by Preston, with the adjustments employed by Montgomery (2007, 2010), placing at its core a perceptive test to study Dialect identification by using Draw-a-Map. In addition, quantitative analysis will be supported by recent statistical methods with a geometrical flavour. This procedure enables the diagnosis and inference of PD data beyond a simple descriptive analysis based on numbers.

In that context, we must emphasize that the knowledge of speakers' (non-linguists') opinions on dialectal variation is of utmost importance both for the study of variation and linguistic change as for language planning. On the one hand, the research into attitudes and perceptions can help us achieve a better understanding of language variation, change, maintenance and decay. On the other, it helps to understand certain attitudes and behaviours and promotes the implementation of language planning adapted to the needs of a particular speech community.

In a situation like that of Galician, characterized by the contact of two languages with unequal status, the importance of discovering speakers opinions, beliefs and perceptions in relation to the language variety they speak and those spoken by others is, perhaps, even greater. In this sense, it is essential to supplement the lack of work carried out in this field so as to advance in knowledge and contribute to halting and even pushing back unfavourable dynamics for the language.

2.2. *Technical support*

The following points are not aiming at completeness, but rather provide an overview of technical support for PD. Firstly, data collection is presented, with the implementation of an app for this task. Secondly, we show how PD data can be treated and exploited by employing tools provided by a Geographic Information System (GIS). Finally, we illustrate the procedure followed in order to perform a proper statistical analysis for such data.

- **Data collection**
  An offine application was built to allow automated collection of data. The sample below illustrates its simple design (Figure 1) and how it edits information.

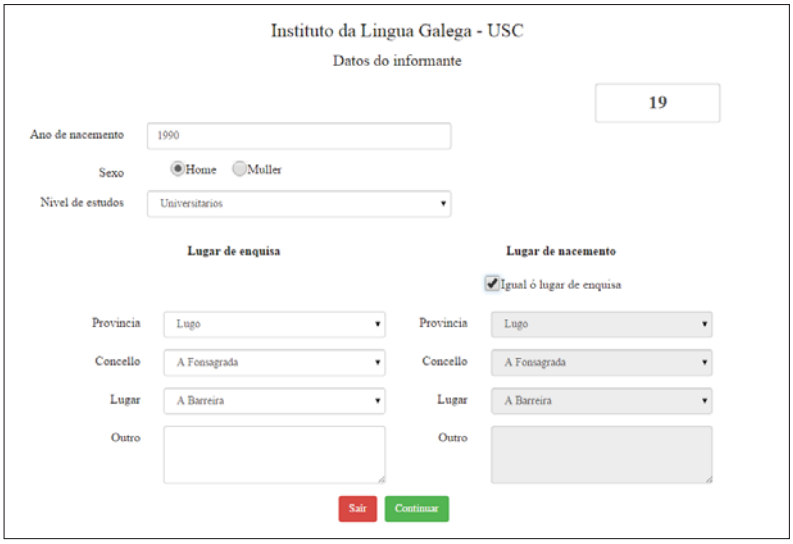


Figure 1: Front view of the data collection application.

For the purpose of our study, three main sections were included in the perceptive test:

1. A part to collect some sociodemographic variables.
2. A draw-a-map task to identify respondents' perceptions.
   The informants were presented with a variety of oral stimulation and asked for their perceptual judgments. Seven different audio tracks (six recordings of dialect speech and one of standard Galician) were played to each participant. In that application, they had a map of Galicia on which the seven biggest cities were marked (an example of a map is shown in Figure 2), and thus they were asked to assign each voice to a specific geographical area.
3. A questionnaire with long answers that must be processed for linguistics' experts, with questions regarding:

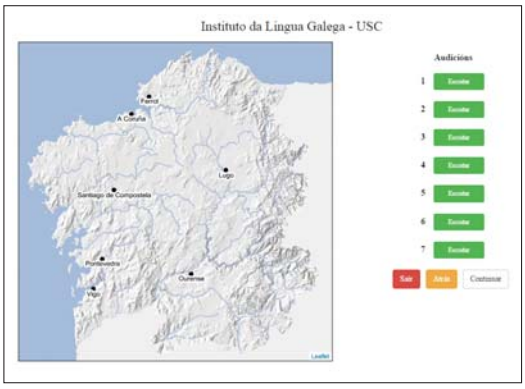


Figure 2: Map of Galicia with the seven biggest cities marked in the PD application.

- Their knowledge about dialect variation in the Galician language area.
- Their perceptions about how correct and pleasing the language varieties they identify are.
- The degree of difference or similarity between the various geographical varieties they recognize and their own variety.
- The identifying features of each of these dialects and the characteristics they attribute to their speakers.

This application allows all PD data to be stored in a relational database at the PostgreSQL (`http://www.postgresql.org/`). The storage mechanism was designed to link all the information within a Spatial Reference System. In particular, its outstanding contribution is made by the draw-a-map task (Figure 3), which allows this selection to be automatically translated into a polygon. It can therefore store the information given by respondents in a geometric pattern.

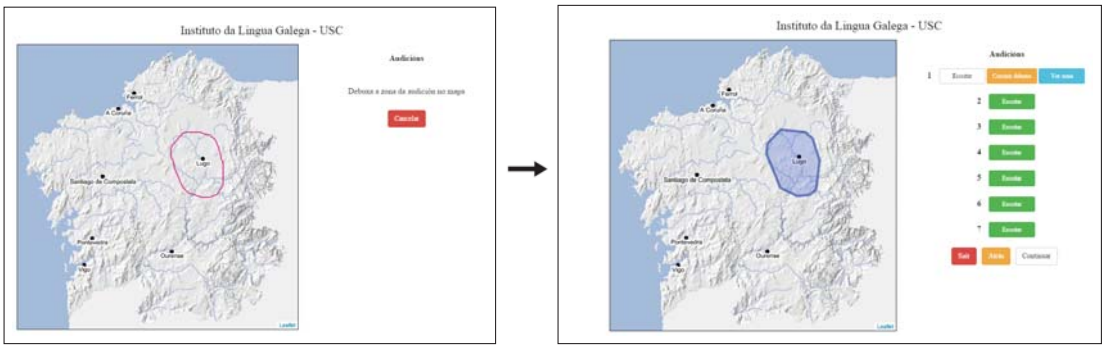


Figure 3: Draw a map example in the PD application.

- **Geographic information treatment**
  Recent research has focused on improving perceptual labels visualization (Cukor-Avila et al. 2012; Mongomery and Stoeckle, 2013). In line with these ideas, we use a Geographic Information System (GIS) for PD visualization. In particular, we use QGIS, an easy-to-use open source GIS application, ideal for data viewing, editing, and producing maps. For this reason, a clear useful approach is related to PD work in order to display geographical recognition of dialect areas. Up to this point, the method employed for processing data enables an aggregation of the draw-a-map by scanning and then incorporating it into a map as layers. However, one of the advantages of using QGIS is that it can be connected to a PostgreSQL/PostGIS database and allows us to view spatially any table containing a geometry column. As a result, we can simplify the accessibility of information.

- **Data analysis**
  A focus on perceptual data visualization means that there has been little sophisticated statistical analysis framework developed in this field. This examination has been so far limited to a qualitative analysis of open-ended answers, or to descriptive analysis of degrees of difference or even correctness and pleasantness of speech. This technical gap has reduced the analytical potential of the data supplied by respondents. Therefore, our aim is to carry out this study by taking advantage of the full potential of information in order to extend our knowledge of PD. Thus, we must distinguish two key points to be treated in our study.

  On the one hand, as mentioned previously, we must realize that we obtain information about locations and perceptual regions in a reference space. This geographical data must be properly processed. That is, it must be used to direct comparison with data from traditional dialectology and dialectometric studies, and therefore to identify which factors influence the recognition of dialectal varieties.

  In this way, for our propose we consider maps as images, and identify locations by pixels. Any region could be represented as a composition of pixels; that is, displayed as an image. Therefore, in order to compare the respondents perceptions with geo-linguistic varieties or establish differences between points and regions, the criterion used was a distance measured to assess the diferences between two images (Baddeley, 1992). Let X be a pixel raster (Galician map), with $A, B \subseteq X$ (two identified locations or regions). Then the Baddeley distance, Δ_b, is then defined as

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

  here $d^*(x,A) = \min\{d(x,A), c\} = \min\{\inf[d(x,a), a \in A], c\}$, where $d(x,x')$ is a metric defining the distance between two points $x, x' \in X$ and the parameters c and p determine a tradeoff between misclassification error and localisation error respectively. As $c \to 0$, the value of $\frac{1}{c}\Delta_b(A,B)$ converges towards the proportion of different pixels between A and B. The index p controls the relative weight of errors of different magnitudes. We fixed for our analysis $p = 2$, to consider Δ_b as the root-mean-squared error.

  Hence, it is possible to create new variables which lead us to study the relation between locations and auditions' regions or selections, like measuring:

  – Distance between two points. For instance, between audition place and birthplace.
  – Distance between the audition place and respondents' selection.
  – Discrepances between respondents selection and dialectal varieties traditionally recognised in Galician studies.

  Once all the information was gathered, and apart from performing descriptive analysis and data visualization, we were also interested in making inferences regarding PD. In this way, having defined variables quantifying distances, we study which factors influenced in varieties recognition by using regression analysis.

  All statistical analysis were performed using `R` 3.1.3. (R development Core Team, 2015)[2]. PD treatment was conducted using `RPostgreSQL`, `rgeos`, `spatstat`, `maptools`, `rasterVis`, `MASS` packages.

3. METHOD ILLUSTRATION

[2]R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

In order to illustrate the overall process, we show an application of this methodology using simulated PD data. Six audition places, which represent different dialectal varieties in Galicia, were incorporated into the database in order of recording: Mazaricos, A Guarda, San Ramón, Xinzo de Limia, A Fonsagrada, and O Barco. A total of nine fictitious informants were included in order to exemplify the process (given that there is no interest on this data description, we only focus on the treatment overview).

- **QGIS** [3]
  Focusing on data visualization, we wish to give a basic outline here on what QGIS can be used for in relation to PD studies. As mentioned before, respondents selections were stored as a polygon, with the disadvantage that they can be off the map. Figure 4 shows a registered drawn map from our fictitious respondents.

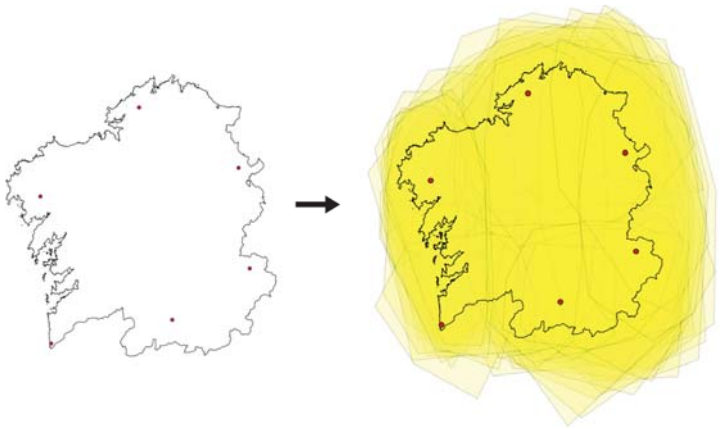


Figure 4: Map of Galicia including audition locations/ Map of Galicia including audition locations and respondents' selections.

This can be easily corrected by using a SQL editor which allows for the intersection between these polygons and the one that defines the Galician geometry.

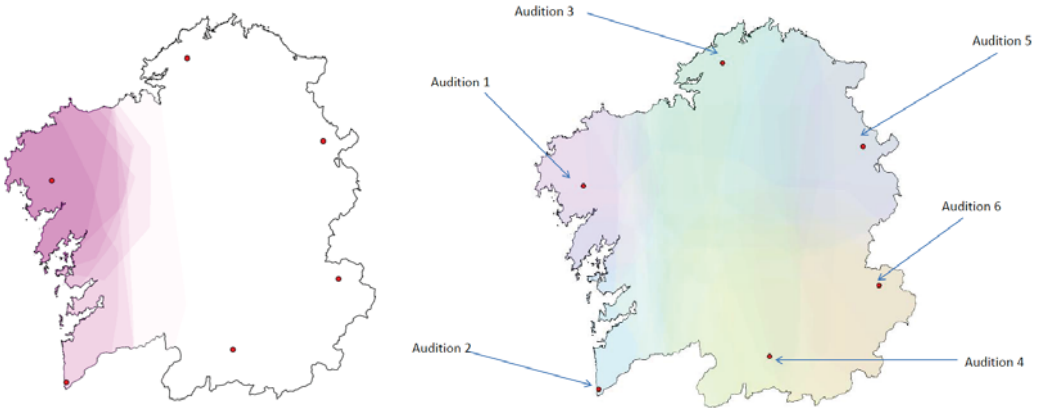
In order to obtain preliminary ideas of respondents' knowledge of dialectal varieties, respondents perceptions can be aggregated for each audition and a colour scale established according to selection frequencies; the color intensities show different degrees of agreement (Figure 5(a)). We can even merge all these representations in order to visualize the dialectal recognition of our respondents, with each color corresponding to each auditions answers 5(b)).

We must be aware that respondents' answers could overlap with different regions, so the dialectal varieties determined by linguists could be split or joined in different regions as in the example below (Figure 6). For instance, let the red area be the traditional linguistic variety for audition 3, the blue area one be the respondent answer to audition 1, and the green area, audition 3.

This enables us to configure new dialectal regions according to speakers opinions and perceptions, which can differ from linguistics studies, as can be seen in Figure 5(b)), which shows perceptual areas of our simulated PD study in Galicia, and in which we clearly differentiate a total of six perceptual dialect varieties.

- **R project**
  Although this can also be done by using `R project` this is not an inmediate process as with QGIS. First of all, we need to configure a database connection from `R` to PostgresSQL, by using RPostgreSQL package:

[3]QGIS Development Team (2015). QGIS Geographic Information System. Open Source Geospatial Foundation Project. `http://qgis.osgeo.org`.



(a) Frequencies for a single audition (b) Frequencies colormap for all auditions

Figure 5: Frecuencies maps elaborated with respondents' answers and represented using QGIS.

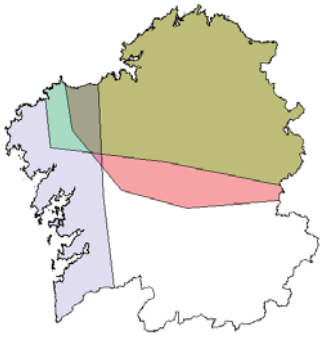


Figure 6: An example of how respondents' answers for audition 1 and 3 (blue and green respectively) overlap a dialectal variety determined by linguists for audition 3 (red).

```
require(RPostgreSQL)
m <- dbDriver("PostgreSQL")
con <- dbConnect(m,user="user_name",password="*****",dbname="database_name")
```

In this way, it is neccesary to use SQL language to extract information from the different tables in our database. An example of how to obtain information is shown below:

```
sql.galicia <- "SELECT ST_AsText(st_transform(geo,25829)) AS ShapeWKT from
+ geo_galicia;"
rs = dbSendQuery(con,sql.galicia);rs
df = fetch(rs,n=-1)
geo_galicia=readWKT(df)
```

By using the `rasterVis` library we can also represent frecuencies maps as shown in Figure 7.

In addition, we can perform statistical analysis in order to study which factors influence the geographical varieties of Galician language recognition. As mentioned before, we can use Baddeley's distance to measure the discrepancies between points and/or polygons. We therefore notice if respondents are able to place and identify dialectal varieties by measuring distances. This led us to define new variables of interest:

– Dependent variable: the distance between respondents' selected region and the true region of dialectal varieties (called "theoretical distance").

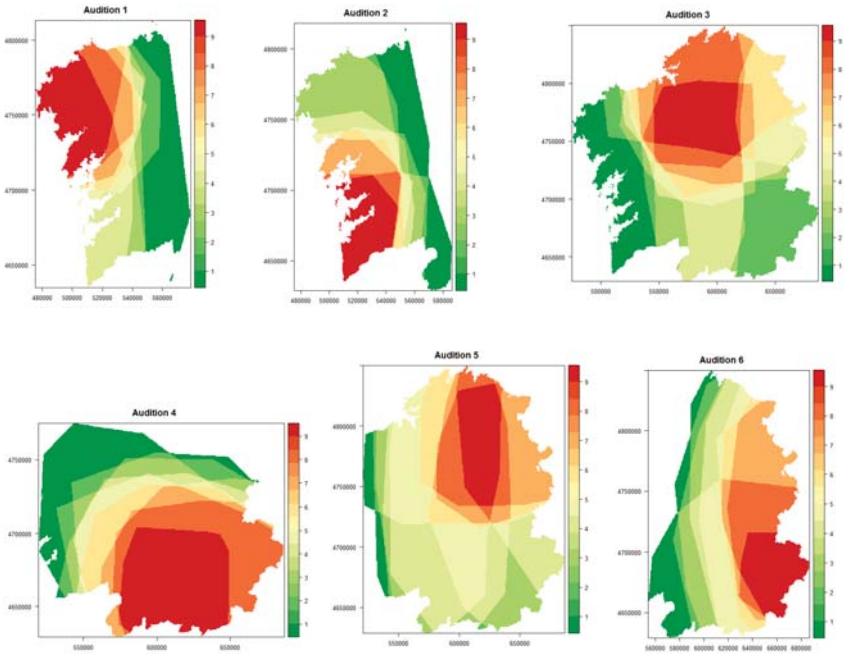


Figure 7: Frequencies maps for different auditions obtained with R project.

– Explanatory variables: distance between birthplace and audition place, distance between birthplace and survey place, distance between survey place and audition place.

These are managed together with other several variables: gender, age range, studies, etc. Under these circumstances, a regression model can be formulated in order to study the determining elements in the recognition of varieties.

The following case provides an example of the results obtained for our simulated PD study. In particular, we focus on third-region variety recognition. Firstly, the true variety region (in this case, considered as one of the polygons issued by respondents) and the selected regions by respondents are represented in Figure 8, on the left and right, respectively.

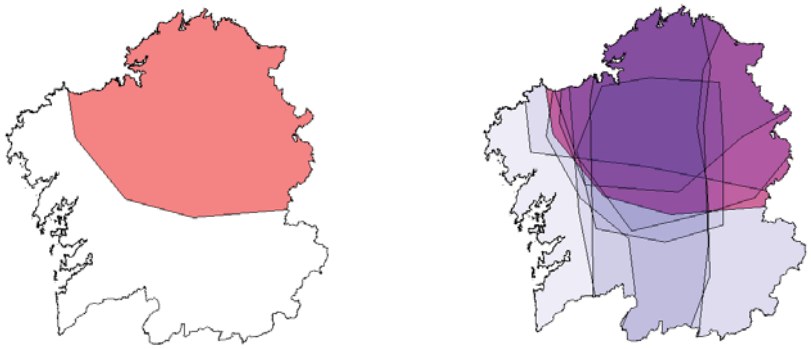


Figure 8: Dialectal region for the third audition / Respondents' selections for the third audition.

Once regions and sociodemographic information are stored in computerized data, we can directly access to obtain:

– Dependent variable: "theoretical distance" (measured in kilometres) - a total of nine distance values. The polygon selected as the true region, the second respondants selection, had a distance equal to zero.

– Explanatory variables: gender, age, educational level and some locations, such as birthplace. To deal with this type of information, for example, we can quantify the distance between audition place and the birthplace, measured in kilometres (see Figure 9).

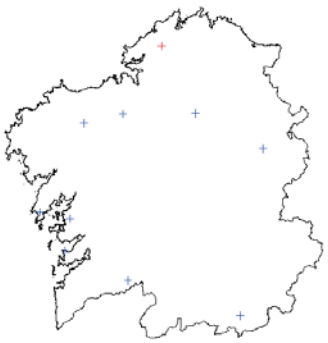


Figure 9: Locations of the audition place (red) and respondents' birthplace (blue).

Thus, we can merge data of interest from the third audition and create our base to work which to work, as shown in Table 1.

Table 1: Database for the third audition in our ficticious PD study.

id	theoretical distance (km)	gender	age	educational level	birthplace (km)
1	35.574	h	45	primary	52.015
2	0.000	m	43	primary	72.774
3	6.303	m	41	secondary	96.135
4	30.342	h	48	secondary	151.062
5	19.042	h	53	primary	137.039
6	9.340	m	39	secondary	50.424
7	27.842	h	51	secondary	130.229
8	2.229	m	36	universitary	187.196
9	30.121	h	55	secondary	157.935

After defining the variables and the data set, a multiple regression analysis can be performed in order to study which factors are influencing the dialectal varieties recognition. This can be written as:

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

where X includes both categorical and continuous variables and ϵ is the error term. For our particular case,

$$Theoretical\ distance = \beta_0 + gender\beta_1 + age\beta_2 + educational\ level\beta_3 + birthplace\beta_4 + \epsilon.$$

Model variables have been selected by the Akaike Information Criterion (AIC) using a backward procedure. Final model regression includes gender, educational level and birthplace as explanatory variables, results obtained are shown in Table 2.

We can observe that the average of the theoretical distance, without considering explanatory variables, is equal to 40.71 (km). In addition, our results showed that being a woman and born far from the audition place decrease the dialectal identification in 31.24 and 0.14 points respectively. However, a higher level of education increase the third audition recognition.

The previous results are obtained with simulated PD data. However, over a period of three months, from July to September 2015, the fieldwork to obtain real data will be performed and subsequently analyzed with these tools.

Table 2: Final regression model.

Variable	Coefficients	p-value
Intercept	40.7055	< 0.05
gender:woman	-31.2349	< 0.05
educational level:secondary	8.8054	0.03
educational level:universitary	18.6868	0.05
birthplace	-0.1385	0.03
Adjusted R-squared: 0.9464		

REFERENCES

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

Cukor-Avila, P., Jeon, L., Rector P. C., Tiwari, C. and Shelton, Z. (2012) "Texas - Its Like a Whole Nuther Country": Mapping Texans' Perceptions of Dialect Variation in the Lone Star State. Proceedings of the Twentieth Annual Symposium About Language and Society–Austin. Texas Linguistics Forum, 55, pp. 10–19.

Fernández Rei, F. (1990) Dialectoloxía da lingua galega. Vigo: Edicións Xerais de Galicia.

Montgomery, C. (2007) Northern English Dialects: A perceptual approach. PhD thesis, University of Sheffield.

Montgomery, C. (2011) Starburst Charts: Methods for investigating the geographical perception of and attitudes towards speech samples. In Studies in Variation, Contacts and Change in English (eVARIENG) Volume 7.
`http://www.helsinki.fi/varieng/journal/volumes/07/montgomery/index.html`

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.

Preston, D. R. (1989) Perceptual Dialectology: Nonlinguists' Views of Areal Linguistic. Dordrecht: Foris.

Preston, D. R. (1999) Introduction. In Preston, D. R. (ed.): Handbook of Perceptual Dialectology, vol.1: xxiii-xxxix. Amsterdam: John Benjamins Publishing Company.

Sousa, X. (2006) Aproximación á análise dialectométrica das variedades xeolingüísticas galegas: un estudo comparativo, in M. C. Rolão Bernardo & H. Mateus Montenegro (ed.), Actas do I Encontro de Estudos Dialectológicos, Ponta Delgada: Instituto Cultural de Ponta Delgada, 345-362.