Analysis of Catalan Knowledge by Municipalities

Enric Reverter, Pim Schoolkate

April 8, 2022

1 Introduction

Catalan language has gone through a lot during the last century, specially during the persecution after the Spanish Civil War. As such, Catalan speakers have been oscillating in numbers thorough generations. The aim of this project is to analyze the number of speakers taking into account contextual variables such as the level of education attained or the number of immigrants at a municipality level. To do so, data from Idescat [3] is used.

There do not seem to be previous studies analyzing the correlation between number of Catalan speakers and other socioeconomic indicators. However, there are other kind of studies such as the usage of Catalan by the context of different areas of personal relations [1]. Other studies can be found at Gencat [2].

2 Data

The data can be retrieved through the Idescat API, which is quite messy. The data is obtained as a semi-structured set of dictionaries, one for every municipality, each of which contain different nested indicators by subject. The cleaning will be explained on later sections. Once the data is transformed to a tabular form, each row represents a municipality while the columns describe the many indicators retrieved. In total, it has 947 rows and 369 columns. It is worth to mention that for one indicator, only its last version is available. That is, if an indicator has been measured in 2011 and 2021, only 2021 is retrieved.

Figure 1: Percentage of missing values through the variables.

Regarding the variables, all of them are numerical, so for the sake of the exercise some will be converted into categories. For example, the count of livestock is available for different kinds (i.e., avian and bovine), so a new variable encapsulating the main type of livestock by municipality will be created. Then, the different livestock count variables will be dropped. Also, a significant number of columns have more than 50% of its values missing, so only a fraction of them is going to be possible to use during the analysis.

The target variable is a continuous numerical variables, meaning that it is expected that some form of regression will be fit for this project. In order to cover the course content, some modification (like factorizing the response) might be made in order to be able to use other techniques taught in the course.

References

[1] Els usos lingüístics de la població de catalunya. https://llengua.gencat.cat/web/.content/documents/publicacions/publicacions_en_linia/arxius/eulp2018-opuscle.pdf. Accessed: 2022.

[2] Gencat, llengua catalana, dades i estudis. https://llengua.gencat.cat/ca/serveis/dades_i_estudis/. Accessed: 2022.

[3] Idescat. https://www.idescat.cat/. Accessed: 2022.