Exploring methods to predict future rates of Irish speakers using census data

Diarmaid Hehir

March 28, 2024

Abstract

Language modelling is an important area of research in a world where linguistic diversity is increasingly under threat. This paper explores the theory and application of language competition modelling with historical census data for Co. Galway.

Acknowledgement

I would like to express my gratitude to my supervisor, Patrick Mannion. His continuous support in our regular meetings throughout final year was immensely helpful and without it I could not have finished this project. I would also like to thank the other lecturers in the school of computer science who have offered feedback and advice, both in relation to the project itself and on keeping stress levels manageable while doing it. Finally I would like to thank my friends and family for their support.

Plagiarism declaration

DECLARATION: I, Diarmaid Hehir, hereby declare that this final year project, titled "Exploring methods to predict future rates of Irish speakers using census data", and the work presented in it are entirely my own except where explicitly stated otherwise in the text, and that this work has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

List of Figures

1 Inter-generational language transmission, UNESCO 2003 9
2 Proportion of speakers, UNESCO 2003 10
3 Trends in existing domains, UNESCO 2003 10
4 Responses to new domains, UNESCO 2003 11
5 Availability of education materials, UNESCO 2003 11
6 State attitudes to language supports, UNESCO 2003 12
7 Community attitudes towards languages, UNESCO 2003 12
8 Amount and quality of documentation, UNESCO 2003 13
9 Ability to speak Irish data for Nun's Island electoral division, Census 2016. Source: CSO 21
10 Frequency of speaking Irish data for Nun's Island electoral division, Census 2016. Source: CSO 21
11 Irish electoral divisions. Source: Wikimedia 22
12 Proportion who reported ability to speak Irish by electoral division, 2006. 24
13 Proportion who reported ability to speak Irish by electoral division, 2011. 24
14 Proportion who reported ability to speak Irish by electoral division, 2016. 24
15 Proportion who reported ability to speak Irish by electoral division, 2022. 24
16 Proportion of daily Irish speakers by electoral division, 2006. 25
17 Proportion of daily Irish speakers by electoral division, 2011. 25
18 Proportion of daily Irish speakers by electoral division, 2016. 25
19 Proportion of daily Irish speakers by electoral division, 2022. 25
20 First rows of the combined daily speakers dataset 26
21 Stock and flow model. Wyburn. J, 2019. 27
22 Feedback loops in the Epidemiology model. Wyburn. J, 2019. 27
23 Console output during optimization after multi-threading 28
24 Error values comparing 2022 prediction and actual 2022 data 29
25 Error values for epidemiology model after tuning initial parameter values 30
26 Parameter tuning heatmap showing mean absolute error 30
27 K-fold cross validation. Source: scikit-learn.org 31
28 Regression model absolute error predicting 2022 data by electoral division 32
29 Regression model, proportion of daily Irish speakers prediction, 2027 . . 33
30 Regression model, predicted change in proportion of daily Irish speakers, 2022-2027 33
31 Epidemiology model prior to tuning, predicted proportions of daily Irish speakers in 2027. 34
32 Epidemiology model after tuning, predicted proportions of daily Irish speakers in 2027. 34
33 Epidemiology model prior to tuning, predicted change in proportion of daily Irish speakers, 2022-2027. 34
34 Epidemiology model after tuning, predicted change in proportion of daily Irish speakers, 2022-2027. 34
35 Proportion of daily Irish speakers over time in Leitir Móir 35

36 Proportion of daily Irish speakers over time in Ballybaan (with 0 to 0.1 axis) 35

Contents

1 Introduction **8**
1.1 Why are languages important? . . . 8
1.2 What is language modelling? . . . 8
1.3 What are the aims of this project? . . . 8

2 Background research **9**
2.1 Theories of language transmission . . . 9
2.2 Different types of language transmission . . . 13
2.3 Geo-demographic parameters . . . 14
2.4 Socio-linguistic parameters . . . 14
2.5 Levels of analysis . . . 15
2.5.1 Macroscopic . . . 15
2.5.2 Microscopic . . . 15
2.5.3 Mesoscopic . . . 15
2.6 Types of equations used . . . 15
2.6.1 ODE . . . 15
2.6.2 Lotka-Volterra (LV) Models . . . 15
2.6.3 Reaction-Diffusion (RD) Models . . . 16
2.6.4 System dynamics Models . . . 16
2.6.5 Agent-based models . . . 16
2.6.6 Conversation Game . . . 16
2.7 Other relevant topics . . . 16
2.7.1 Machine learning . . . 16
2.7.2 Multi-layer Perceptron . . . 17
2.7.3 Epidemiology modelling . . . 17
2.8 Notable studies . . . 17
2.8.1 Modelling the dynamics of language death . . . 17
2.8.2 Quantifying the driving factors for language shift in a bilingual region . . . 18
2.8.3 The spatial distribution of clusters and the formation of mixed languages in bilingual competition . . . 18
2.8.4 Proficiency, attitude, and conventions in minority languages . . . 18
2.8.5 Agent-based models of language competition . . . 18
2.8.6 What is India speaking? Exploring the “Hinglish” invasion . . . 18

3 Data **20**
3.1 Census Questions . . . 20
3.2 Selecting censuses to use . . . 21
3.3 Electoral divisions . . . 21
3.4 Graphing data for maps . . . 22
3.5 Feature selection . . . 23
3.6 Geographic breakdown . . . 23
3.6.1 Proportion of Daily Speakers . . . 24
3.7 Formulating dataset for model use . . . 25

4 Methodology **26**
4.1 Epidemiology Model . . . 26
4.1.1 Primary influences . . . 26
4.1.2 Removing contact variable . . . 27
4.1.3 Parameter optimization . . . 27
4.1.4 Multi-threading . . . 27
4.1.5 Testing and error measurement . . . 28
4.2 Regression Model . . . 29
4.2.1 Building the model . . . 29
4.2.2 Testing and error measurement . . . 30
4.3 Comparing and contrasting the models . . . 32

5 Results **32**
5.1 Future Irish usage projection by model . . . 32
5.1.1 Regression Model . . . 32
5.1.2 Epidemiology Model . . . 32
5.2 Comparing predictions for two areas . . . 34

6 Conclusion **36**
6.1 Error . . . 36
6.2 Future trends . . . 36
6.3 Value of models . . . 36
6.4 Societal effects . . . 36
6.5 Possible future improvements . . . 36
6.5.1 Inclusion of 1996 and 2002 census data . . . 36
6.5.2 An agent-based approach . . . 37
6.5.3 A network theory approach . . . 37
6.5.4 Expanding the geographic area . . . 37

7 External resources **39**

"Tír gan teanga, tír gan anam. A country without a language is a country without a soul"

1 Introduction

1.1 Why are languages important?

Language is the medium by which humans connect with one another. According to the linguist Noam Chomsky, language is "perhaps, the central aspect of human nature."[1] Languages - like organisms - display competition. Globalization has led to massive growth in the number of speakers of the most widely spoken languages. This same process has also contributed to a decline in the number of speakers of those languages with the least speakers, to the point that linguists warn of an ongoing mass extinction in languages[11].

Language competition and diversity is especially relevant in Ireland, where the native Gaeilge has been largely displaced by English as the primary medium of communication although a significant portion of the population still retains the ability to speak Irish thanks to robust support in the education system.

Linguistic diversity is important because language has long been recognised as a huge factor in influencing identity and culture. In order to maintain the unique identity of the many different groups around the world, it is important to learn more about how languages spread.

1.2 What is language modelling?

Language modelling is using computational and mathematical models to simulate the behaviour of human languages. There are many different ways to do this, but all share the overall aim of gaining an insight into how competition between languages shapes the language environment and to inform the work of researchers and authorities working in the area.

1.3 What are the aims of this project?

The minimum achievable goal of this project was to develop an understanding of language modelling techniques and to build a model that allows predictions to be made for future langauge trends in Galway.
The preferred goal was to fulfil the minimum achievable goals and then go further, developing more than one model and comparing the performance of these models.
The stretch goal was to build a model using a novel approach that has not been done before anywhere to add to the body of research.

2 Background research

The primary study used for research was "A systematic and interdisciplinary review of mathematical models of language competition"[7] by Boissoneault and Vogt, published in 2021. This is a meta analysis of over 100 different studies on the topic of mathematical language modelling. In the opinion of the author this is the single best "jumping off point" to start learning about the area of language modelling.

2.1 Theories of language transmission

UNESCO's 2003 report "Language Vitality and Endangerment"[10] is one of the first attempts to quantify the global problem of language loss and call for action to protect endangered languages. In this report, a 9 factor approach is outlined for determining the vitality (health) of a language.

1. Inter-generational language transmission - Whether the language is being passed to the next generation or not. This is the most commonly used factor to evaluate the vitality of a language. A healthy language will be spoken by young and old alike, while less healthy languages will have a lower proportion of young speakers and a higher proportion of older ones.

Degree of Endangerment	**Grade**	**Speaker Population**
safe	5	The language is used by all ages, from children up.
unsafe	4	The language is used by some children in all domains; it is used by all children in limited domains.
definitively endangered	3	The language is used mostly by the parental generation and up.
severely endangered	2	The language is used mostly by the grandparental generation and up.
critically endangered	1	The language is used mostly by very few speakers, of great-grandparental generation.
extinct	0	There exists no speaker.

Figure 1: Inter-generational language transmission, UNESCO 2003

2. Absolute number of speakers - The total number of people who speak a language. It's hard to provide a strict interpretation of what absolute number of speakers is healthy, but a small population is more vulnerable to damage from factors like disease and emigration as well as to being assimilated into a larger group.

3. Proportion of speakers within the total population - What percentage of the population speak the language? The population group in question is usually an ethnic or cultural group. A language may have a large number of speakers but be considered endangered because only a small portion of the ethnic group in

question speaks it. Conversely, a language may have a smaller number of speakers in total but be spoken by almost all of the population.

Degree of Endangerment	**Grade**	**Proportion of Speakers Within the Total Reference Population**
safe	5	All speak the language.
unsafe	4	Nearly all speak the language.
definitively endangered	3	A majority speak the language.
severely endangered	2	A minority speak the language.
critically endangered	1	Very few speak the language.
extinct	0	None speak the language.

Figure 2: Proportion of speakers, UNESCO 2003

4. Trends in existing language domain -What the language is used for. A healthy language will be used in all walks of life, with representation in institutions, education, entertainment. Conversely an unhealthy language may see use in progressively less fields. The most endangered languages may see use only in ceremonial purposes.

Degree of Endangerment	**Grade**	**Domains and Functions**
universal use	5	The language is used in all domains and for all functions
multilingual parity	4	Two or more languages may be used in most social domains and for most functions.
dwindling domains	3	The language is in home domains and for many functions, but the dominant language begins to penetrate even home domains.
limited or formal domains	2	The language is used in limited social domains and for several functions
highly limited domains	1	The language is used only in a very restricted domains and for a very few functions
extinct	0	The language is not used in any domain and for any function.

Figure 3: Trends in existing domains, UNESCO 2003

5. Response to new domains and media -New developments in technology lead to the growth of new domains of language use such as broadcast media, social media and the wider internet in general. This often can contribute to the growth of the dominant language. As such, the response of a language community to new domains is very important.

Degree of Endangerment	Grade	New Domains and Media Accepted by the Endangered Language
dynamic	5	The language is used in all new domains.
robust/active	4	The language is used in most new domains.
receptive	3	The language is used in many domains.
coping	2	The language is used in some new domains.
minimal	1	The language is used only in a few new domains.
inactive	0	The language is not used in any new domains.

Figure 4: Responses to new domains, UNESCO 2003

6. Materials for language education and literacy - Educational materials are vital for education in a language. School books, educational programs, dictionaries and everyday media are all important for education in a language.

Grade	Accessibility of Written Materials
5	There is an established orthography, literacy tradition with grammars, dictionaries, texts, literature, and everyday media. Writing in the language is used in administration and education.
4	Written materials exist, and at school, children are developing literacy in the language. Writing in the language is not used in administration.
3	Written materials exist and children may be exposed to the written form at school. Literacy is not promoted through print media.
2	Written materials exist, but they may only be useful for some members of the community; and for others, they may have a symbolic significance. Literacy education in the language is not a part of the school curriculum.
1	A practical orthography is known to the community and some material is being written.
0	No orthography available to the community.

Figure 5: Availability of education materials, UNESCO 2003

7. State language attitudes and policy - The attitude of the state is important for language vitality. The attitude may vary widely, from promotion to indifference to hostility. In some countries (like Ireland) a minority language is encouraged for cultural and ethnic reasons. In other countries, minority languages may be seen as a threat to unity or to the dominant culture.

8. Community members' attitudes towards their own language - The attitudes of a language's own speakers are often complex. In some cases, there will be a negative perception of the language as associated with "backwardness" and poverty. In other cases, a language may be viewed as an important part of cultural and ethnic identity. In either case, the attitudes of speakers are very important in the eventual fate of the language

Degree of Support	Grade	Official Attitudes toward Language
equal support	5	All languages are protected.
differentiated support	4	Minority languages are protected primarily as the language of the private domains. The use of the language is prestigious.
passive assimilation	3	No explicit policy exists for minority languages; the dominant language prevails in the public domain.
active assimilation	2	Government encourages assimilation to the dominant language. There is no protection for minority languages.
forced assimilation	1	The dominant language is the sole official language, while non-dominant languages are neither recognized nor protected.
prohibition	0	Minority languages are prohibited.

Figure 6: State attitudes to language supports, UNESCO 2003

Grade	Community Members' Attitudes toward Language
5	*All* members value their language and wish to see it promoted.
4	*Most* members support language maintenance.
3	*Many* members support language maintenance; others are indifferent or may even support language loss.
2	*Some* members support language maintenance; others are indifferent or may even support language loss.
1	Only *a few* members support language maintenance; others are indifferent or may even support language loss.
0	*No one* cares if the language is lost; all prefer to use a dominant language.

Figure 7: Community attitudes towards languages, UNESCO 2003

9. Amount and quality of documentation - documentation quality can vary wildly, some languages are documented only in a few scattered texts or not at all. Other languages have complete and comprehensive documentation with audio and visual recordings. Languages that are not documented and passed down purely orally are at a high risk of being entirely lost if they are no longer passed down.

Nature of Documentation	Grade	Language Documentation
superlative	5	There are comprehensive grammars and dictionaries, extensive texts; constant flow of language materials. Abundant annotated high-quality audio and video recordings exist.
good	4	There are one good grammar and a number of adequate grammars, dictionaries, texts, literature, and occasionally updated everyday media; adequate annotated high-quality audio and video recordings.
fair	3	There may be an adequate grammar or sufficient amount of grammars, dictionaries, and texts, but no everyday media; audio and video recordings may exist in varying quality or degree of annotation.
fragmentary	2	There are some grammatical sketches, word-lists, and texts useful for limited linguistic research but with inadequate coverage. Audio and video recordings may exist in varying quality, with or without any annotation.
inadequate	1	Only a few grammatical sketches, short word-lists, and fragmentary texts. Audio and video recordings do not exist, are of unusable quality, or are completely un-annotated.

Figure 8: Amount and quality of documentation, UNESCO 2003

2.2 Different types of language transmission

Languages spread in a number of different ways, and understanding how this happens can inform our model development[7].

- Horizontal transmission: This is peer-to-peer language transmission through interactions with others. Language is a "use it or lose it" situation, and the more interactions someone has using a particular language the better they will become at using it. The converse is also true where language skills can atrophy over time through lack of use.

- Vertical transmission: (also called inter-generational transmission) occurs from parents to their children. This is the primary method of language acquisition, with most people continuing to speak the language they were raised speaking as a child.

- Exogenous transmission: occurs via institutions such as schools. In some cases, this has been used to intentionally suppress the usage of a language, such as in Ireland where the British colonial system discouraged the use of Irish in education, the courts and all areas of public life. It has also been used to bolster a minority language, as seen in Ireland with the Gaelscoil movement. A similar model is followed in other areas, such as Brittany with "Diwan" schools and the Basque country with "Ikastolak" schools.

2.3 Geo-demographic parameters

Geo-demographic parameters concern how speakers exist in a space and reproduce, migrate, die. Many different geo-demographic factors have been covered in past research, with some being more common than others.[7]

- Almost all studies take into account the number of speakers. The size of languages both in raw numbers and in the proportion of speakers in a given area is very important. Larger languages tend to perform better than smaller ones, which can be seen in how the largest languages are growing in size while the smallest ones are shrinking.
- Many consider births and deaths as factors that introduce and remove speakers. This is relevant when considering vertical transmission.
- Many studies consider space as speakers separated by distance have less influence than those closer. Many of these also considered movement, often modelled with Reaction Diffusion models.
- Migration between countries is a factor that has not been widely researched in the context of language modelling. This is a factor that can contribute to the growth of a language or to its decline. In the Irish context, emigration has historically disproportionately affected the west of Ireland
- Other rarely considered parameters include ageing, tendency of some speakers to reproduce more, and the effect of borders.

2.4 Socio-linguistic parameters

A wide range of socio-linguistic factors have been covered in research with some being more common than others. [7]

- Prestige: Most common socio-linguistic parameter researched. Prestige refers to the perceived social value of a language. Prestige is often complex and hard to quantify.
- Loyalty: People may remain loyal to their own language for cultural reasons or convenience. Loyalty can be viewed as an opposing force to prestige, as often the reason a speaker stays using a less prestigious language is due to loyalty. Often language loyalty is linked to ideas of nationalism and cultural pride such as was seen in Ireland with the Gaelic revival in the late 1800s.
- Network topology: Different network types can be used. e.g regular, random, small world. This is especially relevant in the case of agent based networks, as these often seek to approximate how human social networks work

Some other socio-linguistic factors were covered less commonly

- Homophily: a preference for interacting with ones own ethnic group
- Environment: rural vs urban areas
- Other investigated parameters include level of competency, language similarity and languages themselves changing over time

2.5 Levels of analysis

2.5.1 Macroscopic

Macroscopic models describe population processes at a high level in terms of aggregate figures. These models give results that align with the mean and do not account for variability between individuals. An example of a study taking a macroscopic approach is Abrams and Strogatz's 2003 study[4] which is discussed in the notable studies section.

2.5.2 Microscopic

Microscopic models attempt to model the entire range of outcomes, rather than just the mean. They attempt to simulate the language environment in detail, simulating the individuals who are not represented by the average. These models also allow directly modelling interactions between individuals. Agent based models are one type of microscopic models that directly model the individuals. One study taking a microscopic approach is "Agent-based models of language competition"[8] which is discussed further in the notable studies section.

2.5.3 Mesoscopic

Mesoscopic models are similar to microscopic ones in that they deal with individuals and the interactions between them rather than the aggregate measurements in macroscopic models. However, in mesoscopic models the "individual" in question is a group of speakers of a language or an entire language itself. These models attempt to apply a similar approach to microscopic models but using a larger group in the place of single individuals. One study that uses a mesoscopic approach is a 2017 study[15] from Prochazka and Vogel which is discussed in the notable studies section.

2.6 Types of equations used

2.6.1 ODE

Ordinary Different Equations refer to simple equations that consider the proportion of the population and some parameters, which are used to predict the change over time. These just consider the proportion rather than the actual sizes. The advantage of ODE models is their simplicity, as they can work with more limited data sets. Abrams and Strogatz's seminal 2003 study[4] pioneered the use of ODE models for language modelling.

2.6.2 Lotka-Volterra (LV) Models

Lotka-Volterra models were originally developed to simulate predator-prey interactions in wildlife. In such a system a rise in the population of the predators leads to higher predation and a drop in the prey population, leading to a drop in the predator population followed by a rise in the prey population and so on. These models are similar to ODE models but also include information about the size of a group, allowing it to change separately to its proportion of the population. One study containing such predator-prey

models is 2016's 'What is India speaking? Exploring the "Hinglish" invasion'[14] by Parshad et al, discussed further in the notable studies section.

2.6.3 Reaction-Diffusion (RD) Models

RD models were originally developed to model physical processes, such as the diffusion of a gas from an area of high density to an area of lower density. RD models are useful in the context of language modelling because they allow changes in geographic space to be modelled. Such a model is used for a mesoscopic approach in one study[15] but can also be used for a macroscopic approach[6].

2.6.4 System dynamics Models

Some models use a system dynamics approach to model language behaviour. These are similar to an ODE model but expressed in stocks and flows. Often these are more accessible and easier to understand for those who may not understand the mathematical models. They also make some implicit parameters explicit. One system dynamics model is Wyburn's 2020 study "The need for language planning to address English-language media pressures on minority language survival in bilingual populations"[17].

2.6.5 Agent-based models

Agent based (AB) models are widely used to simulate the behaviour of many individuals interacting with one another under a set of rules. AB models can provide insights on an individual level that would not be found using equation based models. One study exploring an agent-based approach is "Agent-based models of language competition" by Castellano et al, which is discussed in the notable studies section.

2.6.6 Conversation Game

The conversation game is a game theory approach that focuses on the decisions made by individuals during a conversation. Individuals have imperfect knowledge (such as not knowing if the other person is a speaker of the language). An example of study using this methodology is Ireberri's 2012 study[12] which models the decisions made by a speaker of a minority language in interactions with speakers and non speakers where they have imperfect knowledge. Another study using a conversation game approach is 2019's "Proficiency, attitude, and conventions in minority languages"[13] by Laruelle et al. This study is discussed further in the notable studies section.

2.7 Other relevant topics

2.7.1 Machine learning

Machine Learning seeks to develop techniques that allow computers to improve their performance at a task without being explicitly programmed. ML models can be used to gain insight into patterns in datasets. In the context of language modelling, ML can be useful in cases where there are large datasets (such as census data). The task of predicting future data using past census data would be an example of supervised

learning which is a technique that uses labelled data sets to predict outcomes and recognise patterns. [16]

2.7.2 Multi-layer Perceptron

A Multi-layer Perceptron (MLP) is a type of artificial neural network[16]. Neural networks imitate the behaviour of neurons in a human brain. A neural network is composed of neurons, each of which has an **activation function** and a set of inputs and outputs. A neuron takes in inputs and passes them into its activation function, then passes the result to the neurons in the next layer through its outputs. The inputs and outputs are roughly analogous to the synapses (connections between neurons) in a human brain.

Neurons are located in multiple layers, with each neuron taking its inputs from the previous layer and passing its outputs to the next layer. There is an input layer, an output layer, and an arbitrary number of hidden layers in between. Hidden layers allow increasing complexity at the cost of taking a longer time to train, which is one reason that deep neural networks (those with many layers) have only become popular in recent years with advances in computing power despite being theorised for decades. MLPs are **feedforward** neural networks, meaning that information only flows in one direction. Neurons do not pass back information directly to the previous layer[2].

Learning in a MLP is done using a **loss function**, which measures the difference between the output from the MLP and the actual training data output. Each neuron has a set of weights on its outputs and inputs, and the model adjusts these weights in an attempt to reduce the loss function. This is done through a process called **back propagation**, which involves passing the updated weights back to the previous row. By repeating this process many times, the model can learn very complex relationships and trends in the data

2.7.3 Epidemiology modelling

Some studies model a language environment with two languages using an epidemiological model. Epidemiology models seek to predict patterns of infection and recovery during a pandemic in order to inform public health responses. A common epidemiological model is the SIR (Susceptible Infected Recovered) model which tracks the proportions of the population that are infected with a disease at a given time. This has been used to model language dynamics[17], where one language is represented as the "infection" and the other as the susceptible population.

2.8 Notable studies

2.8.1 Modelling the dynamics of language death

This 2003 study[4] from Abrams and Strogatz was highly influential on subsequent research. It takes a macroscopic approach to modelling language shift, and uses empirical data from several different areas around the world where a native minority language is being displaced by a larger outside language. This study looks at the cases of Quechua

in Peru and Welsh and Scottish Gaelic in the UK. It determines that the most important factor in language preservation is status and if the status of a language can be improved then language decline can be mitigated.

2.8.2 Quantifying the driving factors for language shift in a bilingual region

This 2017 study[15] from Prochazka and Vogel examines Carinthia, a region in Austria with a native Slovenian-speaking minority. The study uses highly granular historical data to show how the dominant language (German) steadily encroached in on the minority one along a geographic front. Geographic features such as mountains slowed or halted the language front, suggesting that the steady progress of the dominant language is due to the minority group being more exposed to its speakers.

2.8.3 The spatial distribution of clusters and the formation of mixed languages in bilingual competition

This 2012 study[5] used a purely theoretical approach (i.e without real world data). The focus of this study was modelling interactions in a language environment using cellular automata. A grid was randomly initialised with different languages, and in each time step a cell would maintain its language or switch to a new one based on the cells surrounding it. Over multiple time steps the spatial distribution would become delineated between larger language blocks, while over longer time periods one language would tend to dominate and eliminate the others. This study found that an equilibrium could be found in some cases with certain spatial layouts. It is an interesting study because of the novelty of its approach and because of how the behaviour shown in the cellular automata model (with a large language crowding out others) is consistent with what is seen in the real world.

2.8.4 Proficiency, attitude, and conventions in minority languages

This 2019 study by Laruelle, Navarro and Escobedo models a scenario with a bilingual minority inside a larger monolingual society. Using a conversation game approach built on game theory, this study explores how rates of language use are influenced by incomplete knowledge along with factors like the strength of personal language preference usage and the societal conventions around the use of a language. This study explores an environment analogous to the one seen in Ireland, where the indigenous minority language exists within a broader largely monolingual society.

2.8.5 Agent-based models of language competition

This 2013 study[8] from Castellano et al follows an agent based approach. It investigates the effect of different network topologies and finds that language prestige and volatility(the ease by which speakers change their language) are highly influential on the outcomes when it comes to a language shift scenario.

2.8.6 What is India speaking? Exploring the "Hinglish" invasion

This 2016 study[14] concerns an interesting scenario which is different to most covered in language modelling literature. India is seeing a large growth in those who report

being English speakers, but this study observes that many supposed English speakers are actually using "Hinglish" a combination of Hindi and English. Rather than being truly bilingual in Hindi and English, they are using both languages at the same time as if they were one language. This is a very interesting study as Hindi and English are both very large languages that could be considered dominant in their own right. The scenario of two languages seemingly merging rather than solely competing is one that merits more exploration, as the question of language change over time has not been extensively looked into in language modelling research.[7]

3 Data

The data used for this project was acquired from the Central Statistics Office (CSO)[1]. The CSO provides Small Area Population Statistics (SAPS) for all areas in Ireland. SAPS are statistics organised by different themes for each area. The following are the themes provided in the 2022 census SAPS data:

- Theme 1: Sex, age and marital status
- Theme 2: Migration, ethnicity and religion
- **Theme 3: Irish language**
- Theme 4: Families
- Theme 5: Private households
- Theme 6: Housing
- Theme 7: Communal establishments
- Theme 8: Principal status
- Theme 9: Social class and socio-economic group
- Theme 10: Education
- Theme 11: Commuting
- Theme 12: Disability, carers and general health
- Theme 13: Occupation
- Theme 14: Industries
- Theme 15: PC and internet Access

3.1 Census Questions

Right back to the first census after independence in 1926, Irish census data has always included questions about the Irish language. Until 1996, this was a simple yes/no question about ability to speak Irish. From 1996 onwards, a question was added about frequency of speaking Irish. From 2006 onwards, an additional clarification was added on whether Irish was spoken inside or outside the education system.

[1] https://www.cso.ie/en/census/

3.2 Selecting censuses to use

The decision was made to use the post-2006 data as it would be possible to select the proportions of the population who spoke Irish daily outside the education system. Given the mandatory nature of Irish in the education system, it is likely that the figures in 1996 and 2002 could overstate the frequency of Irish use. It is likely that many individuals who answered they regularly used Irish were doing so within the education system only. Any study which attempted to use the 1996 and 2002 data could do so by making an educated guess based on the proportions of speakers in the education system in later censuses, or attempt to find separate data from the time for the numbers in education. It was decided that this would be outside of the aims of the project as it was primarily focused on finding enough data to make a prediction into the future and such an amount of data was already available.

With the recent release of 2022 census data, this provided 4 complete censuses of data. 2006, 2011, 2016 and 2022. This data was used for training the models.

Irish Speaker	Value
Yes	590
No	793
Not Stated	53
Total	1,436

Figure 9: Ability to speak Irish data for Nun's
Island electoral division, Census 2016. Source: CSO

Frequency of Speaking Irish	Males	Females	Both Sexes
Daily only within the education system	23	38	61
Daily within and daily outside the education system	3	6	9
Daily within and weekly outside the education system	1	0	1
Daily within and less often outside the education system	0	0	0
Daily within and never outside the education system	0	0	0
Daily only outside the education system	18	24	42
Weekly only outside the education system	27	29	56
Less often only outside the education system	102	182	284
Never speaks Irish	57	76	133
Not Stated	2	2	4

Figure 10: Frequency of speaking Irish data for Nun's Island electoral division, Census 2016. Source: CSO

3.3 Electoral divisions

Small Area Population Statistics are available from the CSO in several forms, such as by county. This project chose to use electoral divisions (EDs) as the boundary. This is for two primary reasons. Firstly, there are 3440 EDs in Ireland, so they provide a high level of granularity. In Galway alone there are more than 200 electoral divisions. Secondly, the electoral divisions have remained mostly static for decades and are changed little, making them a good choice for data over a long period. With the Galway EDs there was a change made between 2016 and 2022. 7 rural EDs with very low populations were merged to preserve the anonymous nature of the data. The EDs in question were Derrycunlagh, DerryLea, Loughatorick, Marblehill, Cong, Ross, and Coos. As it would introduce difficulties with the dataset and each of these areas has a very small population anyway the decision was made to exclude them entirely from the dataset. On the maps they appear as grey, with no data shown.

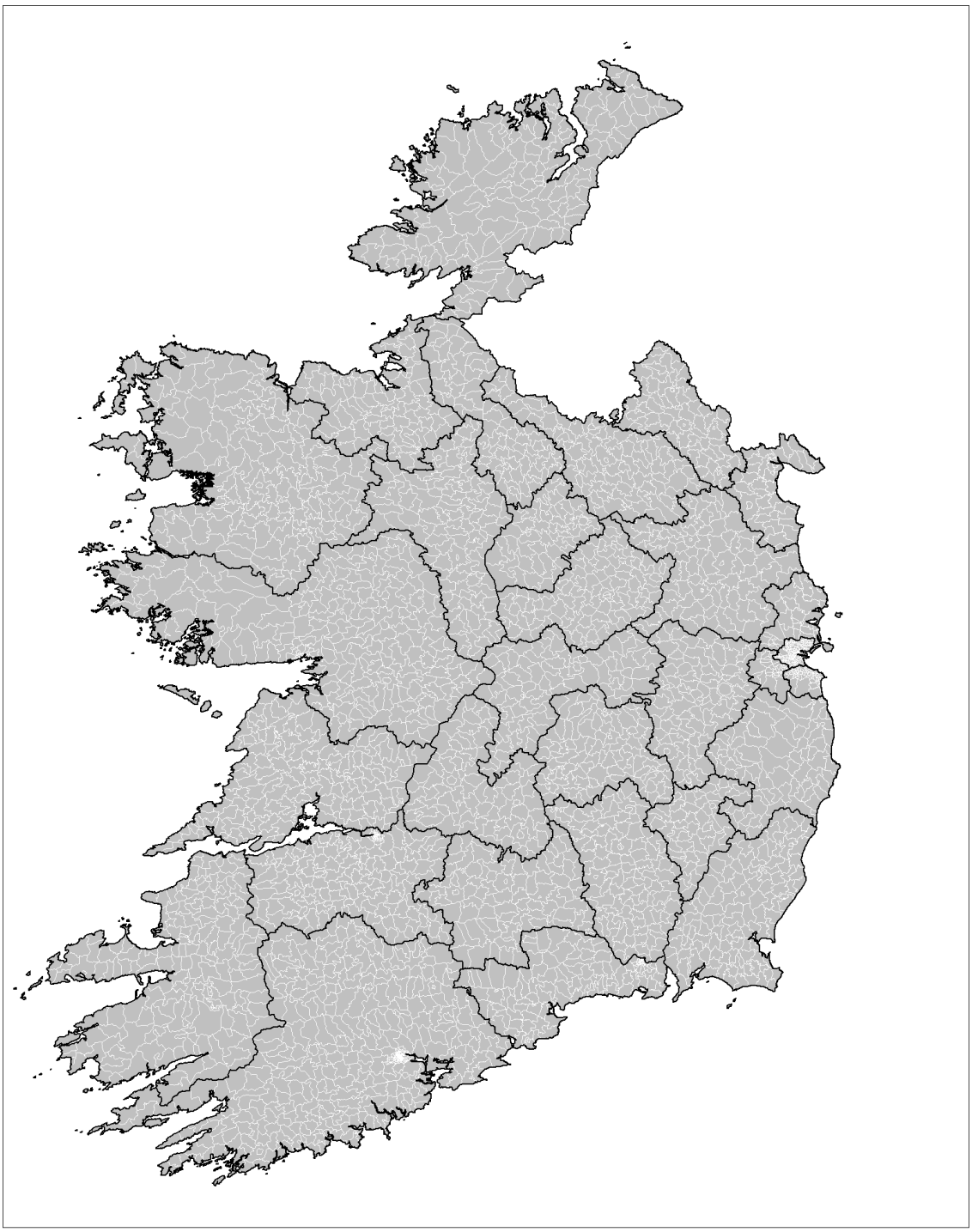

Figure 11: Irish electoral divisions. Source: Wikimedia

3.4 Graphing data for maps

In order to create maps, the library "geopandas" was used. Geopandas is a project that adds on to python's "pandas" library for data analysis. Geopandas adds support for

geographic information. To create maps a "shapefile" is required, which is a data format that stores geographical information. The shapefiles for this project were acquired from the CSO. [2] The electoral divisions boundary files were downloaded in geojson format. Geojson is a format based on json for storing geographical information. The geopandas dataframe was then merged with the existing Galway data frame, using the GUID value to include only the Galway electoral divisions.

3.5 Feature selection

The feature chosen both as a dependent and independent variable was the proportion of daily speakers outside the education system. This included individuals who both spoke Irish inside and outside the education system daily, as well as those who spoke it only outside the education system daily. This feature was chosen as it represents the most restrictive interpretation of Irish speakers, choosing those who use Irish very frequently. As seen in the following subsection there is a large section of the population who report ability to speak Irish, but most of those people do not speak it every day, so merely looking at the proportion who report ability to speak Irish may be misleading. For both models, the dependent variable was the year being predicted, while the independent variables were the previous years. For example, when predicting the proportion of daily speakers for 2022 using the regression model the independent variables are the proportions of daily speakers for 2006, 2011 and 2016.

The proportion of daily speakers was calculated by adding the numbers for daily speakers inside and outside the education system and daily speakers only outside the education system. The sum was then divided by the total population to get the proportion of daily speakers. This was calculated using a python script. It was then possible to use the change of the proportion of daily speakers over time to predict the future rate of daily speakers, an approach explained in later sections.

3.6 Geographic breakdown

Firstly, a geographic breakdown of ability to speak Irish is shown. The results show a very high proportion of the population reporting ability to speak Irish in Connemara. This result is unsurprising, Connemara is a Gaeltacht area where Irish has historically been widely spoken.

Galway city showed some of the lowest levels of ability to Speak Irish in the county. There are a number of possible explanations for this. Galway city has been the target of a large level of immigration in recent decades, leading to an increase in the number of people with no Irish language knowledge. Additionally, many people from other parts of Ireland have also moved to Galway city and many of those areas have a lower level of Irish speakers as Galway is the county with the highest proportion of Irish speakers [3]

The overall ability to speak Irish has remained largely constant over the 4 censuses

[2] https://www.cso.ie/en/census/census2016reports/census2016boundaryfiles
[3] https://www.cso.ie/en/releasesandpublications/ep/p-cp10esil/p10esil/ilg/

shown. Some electoral divisions showed a decline, such as in north west Galway to the north of Connemara. The mostly static level of knowledge could be due to the fact that Irish is taught as a mandatory subject in primary and secondary level education and has been . It is likely, therefore, that a similar proportion of the population will have gone through some level of Irish education in each census.

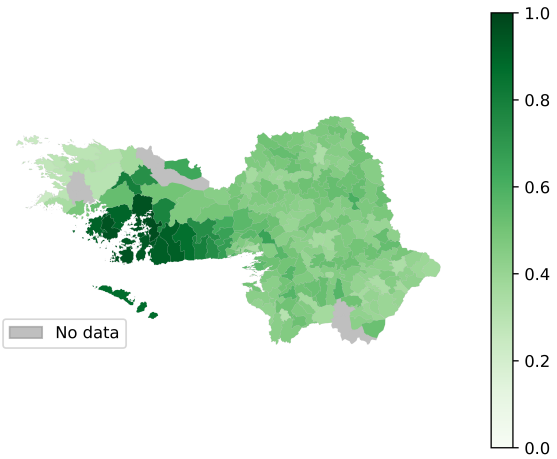

Figure 12: Proportion who reported ability to speak Irish by electoral division, 2006.

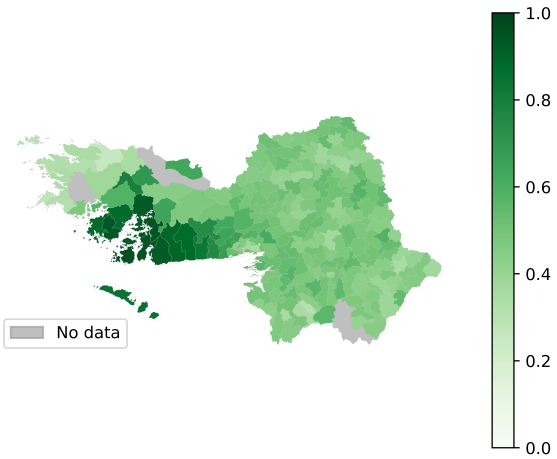

Figure 13: Proportion who reported ability to speak Irish by electoral division, 2011.

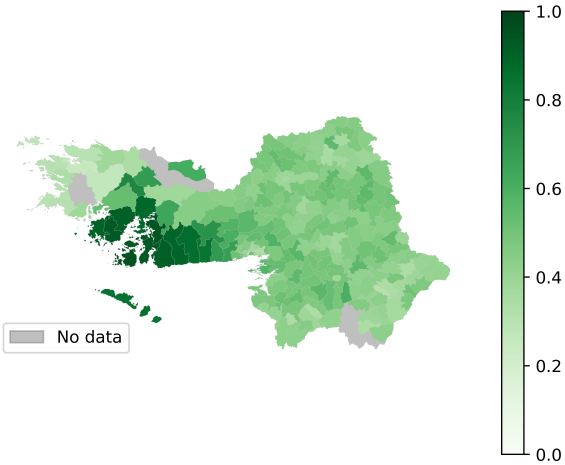

Figure 14: Proportion who reported ability to speak Irish by electoral division, 2016.

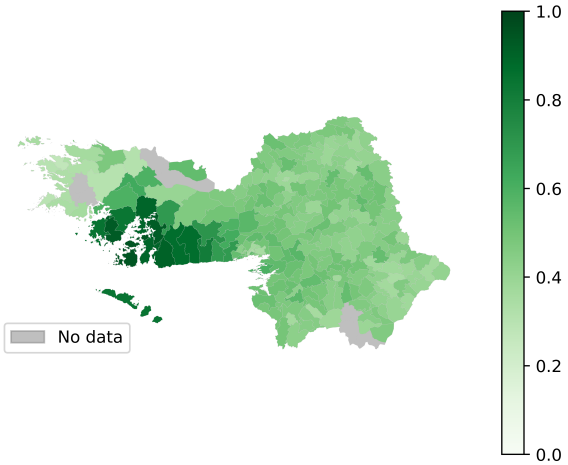

Figure 15: Proportion who reported ability to speak Irish by electoral division, 2022.

3.6.1 Proportion of Daily Speakers

Daily Speakers of Irish are far more concentrated geographically than those who report an ability to speak Irish. The traditionally Irish speaking areas in Connemara have by far the highest proportion of daily speakers, containing the only areas where the majority of the population speaks Irish daily. They can be seen in the dark green on the maps. It can be seen that there has been a mild but consistent decline in the

proportions of daily speakers in EDs around Connemara. This behaviour is consistent with that of a "language front" as explained in Prochazka and Vogel's 2017 study[15] on minority language competition in Austria.

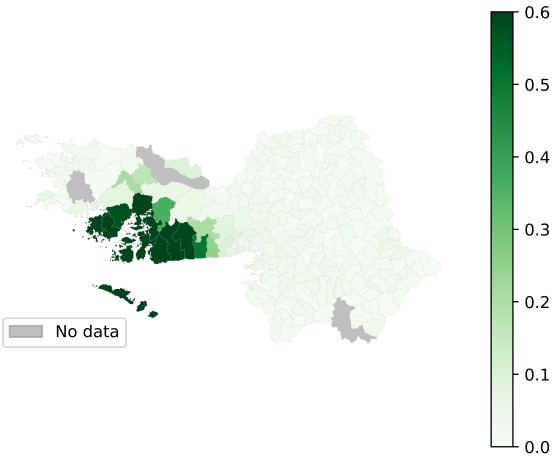

Figure 16: Proportion of daily Irish speakers by electoral division, 2006.

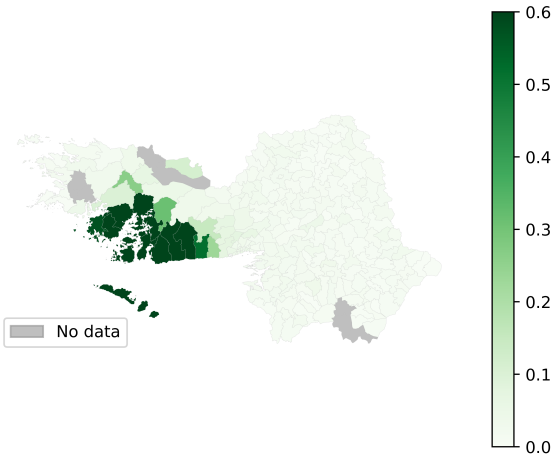

Figure 17: Proportion of daily Irish speakers by electoral division, 2011.

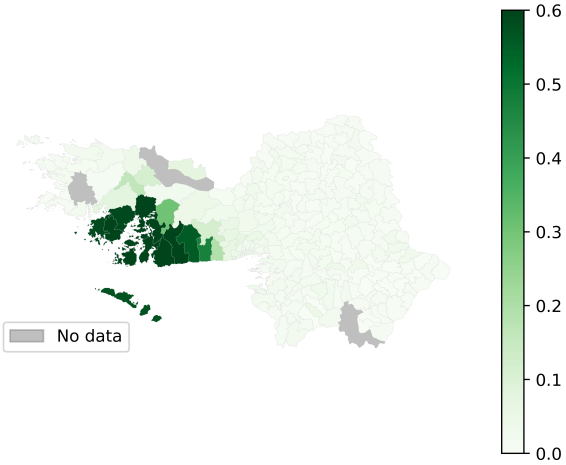

Figure 18: Proportion of daily Irish speakers
by electoral division, 2016.

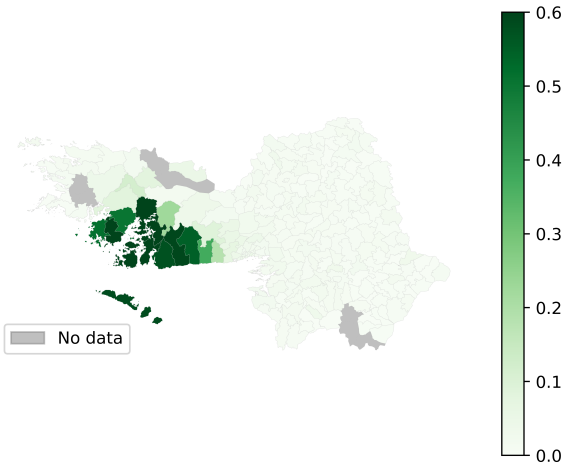

Figure 19: Proportion of daily Irish speakers by electoral division, 2022.

3.7 Formulating dataset for model use

Processing of data was done using python scripts. As explained earlier the proportion of daily speakers was determined to be representative of the most dedicated speakers. The data for daily speakers was calculated for each electoral division 2006,2011,2016 and 2022. The data was then combined into a dataframe containing the division name, GUID, and combined speakers data. This combined dataset was then used in the models.

Electoral District	GUID	Speakers2006	Speakers2011	Speakers2016	Speakers2022
BALLYBAAN	2ae19629-235a-13a3-	0.020414776	0.016112176	0.01768397	0.012842202
BALLYBRIT	2ae19629-235b-13a3-	0.036127168	0.030057803	0.027322404	0.023026316
BARNA	2ae19629-235c-13a3-	0.05003319	0.050501327	0.048198603	0.04403418
CASTLEGAR	2ae19629-235d-13a3-	0.026026605	0.023261759	0.026194145	0.021469466
CLADDAGH	2ae19629-235e-13a3-	0.038506418	0.044676409	0.034104273	0.028315946
DANGAN	2ae19629-235f-13a3-	0.035331384	0.0268757	0.023874658	0.01968591
EYRE SQUARE	2ae19629-2360-13a3-	0.030927835	0.026678725	0.01988218	0.015667575
KNOCKNACARRAGI	2ae19629-2361-13a3-	0.041696621	0.037162162	0.044735031	0.029288703
LOUGH ATALIA	2ae19629-236f-13a3-	0.0511883	0.050505051	0.05359179	0.046189376
MENLOUGH	2ae19629-2370-13a3-	0.033795802	0.032552353	0.028629032	0.030187963
MERVUE	2ae19629-2371-13a3-	0.014975845	0.024571429	0.022866704	0.026126126
MURROOGH	2ae19629-2372-13a3-	0.021579629	0.022090059	0.022602989	0.016352648
NEWCASTLE	2ae19629-2373-13a3-	0.027593255	0.027374302	0.016198704	0.015159171
NUNS ISLAND	2ae19629-2374-13a3-	0.040847201	0.033858268	0.03551532	0.020759193
RAHOON	2ae19629-2375-13a3-	0.043165468	0.035038542	0.028513238	0.024499552
RENMORE	2ae19629-2376-13a3-	0.024637681	0.040381791	0.034294622	0.026836158
ROCKBARTON	2ae19629-2377-13a3-	0.038829488	0.037826319	0.02961165	0.040286976
St. NICHOLAS	2ae19629-2384-13a3-	0.03671461	0.027483314	0.027095682	0.023577491
SALTHILL	2ae19629-2382-13a3-	0.032297755	0.035108599	0.032276492	0.031054131

Figure 20: First rows of the combined daily speakers dataset

4 Methodology

Cover techniques and their application to the data. Designing experiments. Fitting models. Measuring error

4.1 Epidemiology Model

One model used an Epidemiological approach. The monolingual English Speaking population is considered susceptible, while being a bilingual Irish speaker is modelled as if it were an infection. For the sake of the model it is assumed that everyone in the population can speak English and that there are no monolingual Irish speakers or no speakers of any other languages.

4.1.1 Primary influences

The epidemiology model drew primary inspiration from a 2019 study[17] published in the Journal of the Operational Research Society. This study examined the effects of English language media influence on Welsh speakers, and used an epidemiological model with a system dynamics approach to simulate the change of speakers over time. The following formula was used to find the rate of learning Welsh (monolingual English speakers becoming bilingual Welsh speakers.

$$learning-rate = contact*prob-acquisition*prob-meeting*English-speaking \quad (1)$$

- Contact: the number of meetings each individual has per year.
- Prob-acquisition: the probability of acquiring the minority language (parameter)
- Prob-meeting: the probability of meeting a speaker of the minority language

- English-speaking: The proportion of the population who are monolingual English speakers.

This formula is used to model the change in Speakers for both the minority language and English. The only change is that the prob-acquisition parameter is replaced by abandonment-rate and English-speaking is replaced by minority-speakers.

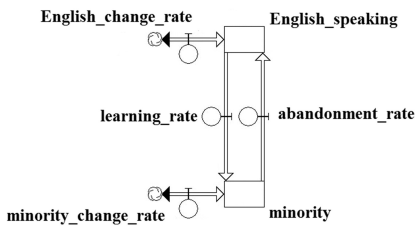

Figure 21: Stock and flow model. Wyburn. J, 2019.

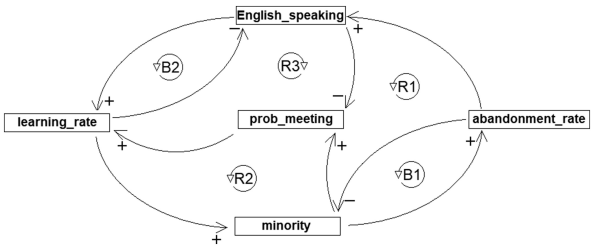

Figure 22: Feedback loops in the Epidemiology model. Wyburn. J, 2019.

4.1.2 Removing contact variable

Some changes were made to the approach taken in this paper. It was decided to eliminate the "contact" variable which represented the number of interactions per year each speaker had. This was because it was thought that this variable was out of place as a flat value of 10 while all other variables were 0 - 1 proportions. Additionally in testing the model did not give remotely plausible results with the inclusion of this variable compared to when it was excluded.

4.1.3 Parameter optimization

Another change was the decision to fit the "abandonment-fraction" and "prob-acquisition" variables using a parameter optimization approach. This involved using SciPy's optimization package. The model was run with 2016 census data as the input and values for abandonment-fraction and prob-acquisition that were initially hard coded. The result was then compared to the 2022 data, and error was calculated (the difference between the prediction and the actual data). The parameters were then adjusted to reduce the error, with the model being run repeatedly. The optimized parameters were then stored and used with the 2022 data to predict the results for 2027.

4.1.4 Multi-threading

The parameter optimization process is an intensive one, when running initially each row could take upward of 40 seconds With 230 rows in the dataset, this was an unacceptable overhead. The decision was made to implement multi-threading using python's concurrent.futures module. This allows a job (optimizing a row) to be submitted to a ThreadPoolExecutor which then selects from the collection of jobs as threads become available. As a result of these changes, the time taken dropped to consistently being under 2 seconds. After the addition of multi threading the entire time taken was far lower than just a single row was before. This improvement was surprising, as it was

expected that the improvement would merely be linear and there isn't any I/O being accessed or a comparable task that would take a very long time and be sped up by multi-threading. One possible explanation could be that the single thread was being locked for most of the time.

```
Optimized parameters:  [0.20976121 0.09010608]
Time elapsed:  1.7191612720489502
Optimized parameters:  [0.10592955 0.19469602]
Time elapsed:  1.7191612720489502
Optimized parameters:  [0.13791154 0.16373782]
Time elapsed:  1.7191612720489502
Optimized parameters:  [0.12086541 0.18000138]
Time elapsed:  1.7191612720489502
Optimized parameters:  [0.14011018 0.16002232]
Time elapsed:  1.7191612720489502
Optimized parameters:  [0.12639342 0.17509216]
Time elapsed:  1.7191612720489502
Optimized parameters:  [0.14416707 0.15662074]
```

Figure 23: Console output during optimization after multi-threading

4.1.5 Testing and error measurement

Error measurement for the epidemiology model was a challenge. It was decided that error measurement would be done by optimizing the acquisition and abandonment parameters with 2011 as the start point and 2016 as the end point. The optimized parameters were then used with 2016 as the start point, making a prediction for 2022. The 2022 prediction was then compared against the actual data for 2022 to determine the error. The lowest error was in AbbeyGormacan, where both prediction and actual data indicated that there were no daily speakers (error = 0.0). The highest error was in Gorumna, with an error of 0.50165. The mean error absolute error (MAE) was 0.02635 with a standard deviation of 0.08482.

There was a very clear trend with the error values where areas with a high proportion of daily Irish speakers (Connemara) had far higher error values than areas with less daily speakers. The error values for areas outside of Connemara are much lower in comparison. This suggested that the epidemiology model was performing significantly worse when modelling areas with a high proportion of Irish speakers.

In light of the unexpectedly high error values, it was decided to explore further tuning of the model. More specifically, the decision was made to search for better values for the "initial guess" of the abandonment fraction and probability of acquisition. Initially, little consideration had been given to these values as it was assumed the parameter optimization would find the suitable values regardless of the initial ones. A series of

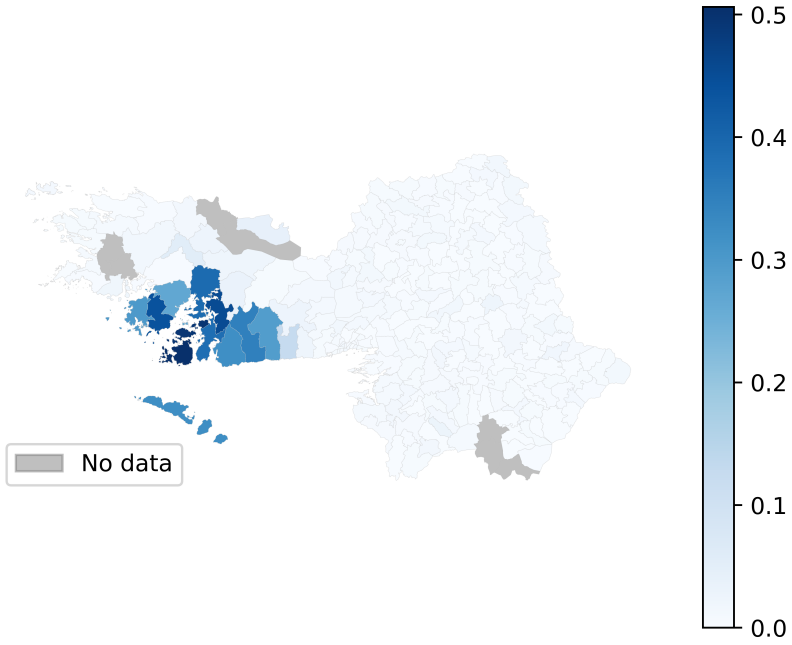

Figure 24: Error values comparing 2022 prediction and actual 2022 data

grid searches were performed with ranges of values, aiming for the lowest mean absolute error. The EDs with the lowest and highest error were the same. AbbeyGormacan as the lowest and Gorumna as the highest. The error in Gorumna was 0.124097, a significant improvement from its previous level at over 0.5. The lowest MAE was found with an initial guess of 0.02 for the probability of acquisition and 0.0000005 for the abandonment fraction. The MAE was 0.00949 with a standard deviation of 0.01524. This represented a significant reduction in both the error and standard deviation. When mapped the error values are still larger in some EDs in Connemara, but the effect is not universal across all electoral divisions there. It can be concluded that tuning the initial parameters was a success.

4.2 Regression Model

For the regression model, it was decided to use SciKitLearn's library of regressor functions. These include linear regressors, kneighboursregressors, and Neural Networks (MLPRegressor). It was decided to use MLPRegressor for several reasons. Neural networks are effective at modelling complex non-linear relations between the features and target variables. As such they are well suited to a problem like language competition, which is not fully understood and certainly complex. Neural networks are also relatively simple to train. In this project, the neural network regressor was simpler than the epidemiological model.

4.2.1 Building the model

MLPregressor has two methods that are relevant. Fit() takes as its parameters the features being used to predict the result and the target feature being predicted. These are the independent and dependent variables respectively, or the X and Y data. Following

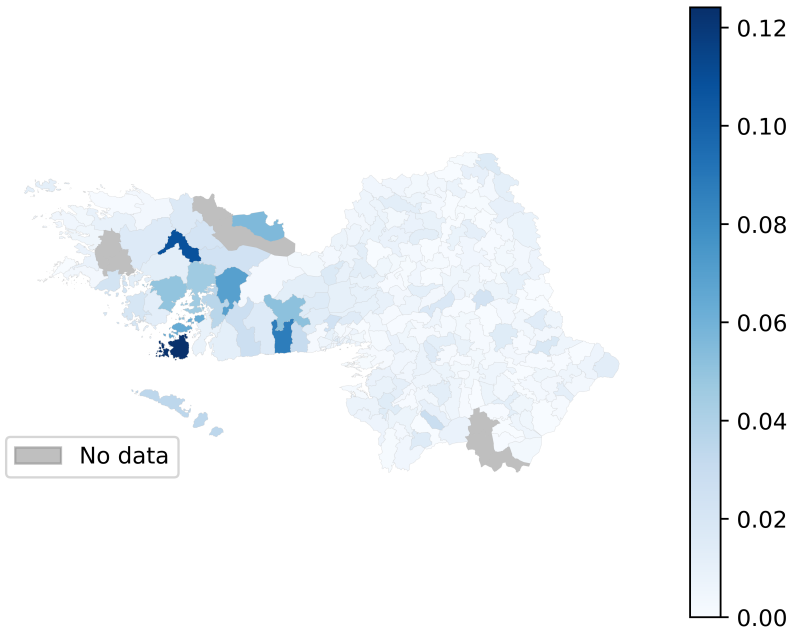

Figure 25: Error values for epidemiology model after tuning initial parameter values

		Abandonment Fraction						
		5.00E-07	1.00E-06	5.00E-06	1.00E-05	5.00E-05	0.0001	0.0005
Prob_Acquisition	0.01	0.00966339	0.0096634	0.0096637	0.009663988	0.0096665	0.0096697	0.009685
	0.015	0.00949199	0.009492	0.0094924	0.009492935	0.0094969	0.0095019	0.009542
	0.02	0.00948767	0.0094878	0.009488	0.009488333	0.009491	0.0094944	0.009521
	0.025	0.00952357	0.0095236	0.0095239	0.009524208	0.0095269	0.0095303	0.009557
	0.03	0.0097594	0.0097594	0.0097597	0.009760062	0.0097628	0.0097661	0.009794

Figure 26: Parameter tuning heatmap showing mean absolute error

fitting, the model can be passed new feature data in order to make a prediction. This is done using the predict() method which takes as its parameter the X data and returns a prediction of the Y data.

One approach would be to construct a model that individually modelled each row (electoral division). However, such a model loses out on the possible relations between electoral divisions. Such relations are likely to be complex and non-linear and it is in learning such relations that a neural network approach can excel. As a result of this, it was decided to fit the model to the entire 230 row data set in order to capture relations between rows. In order to capture as much of the data as possible, it was decided that the model would be fitted with 3 censuses of data with a 4th census being used as the target variable. For example, when predicting the value of daily Irish speakers in the 2027 census, the values for 2011, 2016 and 2022 would be used to fit the model.

4.2.2 Testing and error measurement

Error measurement for the regression model presented some challenges. The conventional approach would be to divide the data into a training/testing split. The training data would be used to train the model and then the results would be tested using the

test data. The problem is that as each row corresponds to an electoral division, the exclusion of any rows would mean the exclusion of divisions from the final prediction.

One solution would be to have included other electoral divisions outside of Galway that could be used for training/testing without needing to be included in the final prediction. However this solution would not be ideal as the new divisions would all be outside of Galway and its likely that the language environment is not the same in other parts of Ireland. With this in mind, that option was ruled out. It was decided instead to pursue the approach of using cross validation.

K-fold cross validation involves splitting a dataset into a number of folds. The process of training and testing is repeated with each fold being used as training data k-1 times and as test data once. Cross validation allows the data to be used efficiently as all data can be used for both training and testing. The model was trained on the training data with the features being 2006,2011,2016 and the target being 2022. Then the model was tested, making a prediction with previously unseen data for 2006,2011,2016. This prediction was compared against the actual data for 2022 for those rows. The absolute error was measured by taking the absolute value of the difference between the prediction and the actual data.

The electoral division with the lowest error was Clontuskert with an error of 0.000019. The electoral division with the highest error was Menlough, with an error of 0.070283. The mean absolute error (MAE) was 0.00805, with a standard deviation of 0.00975. When plotted on a map, there is no clear geographic trend for which areas have higher or lower error. The areas with high proportions of daily Irish speakers in Connemara have relatively low error values, so it could be observed that the model performs well in those areas. Aside from this, there isn't a clear trend on which areas have higher or lower error values.

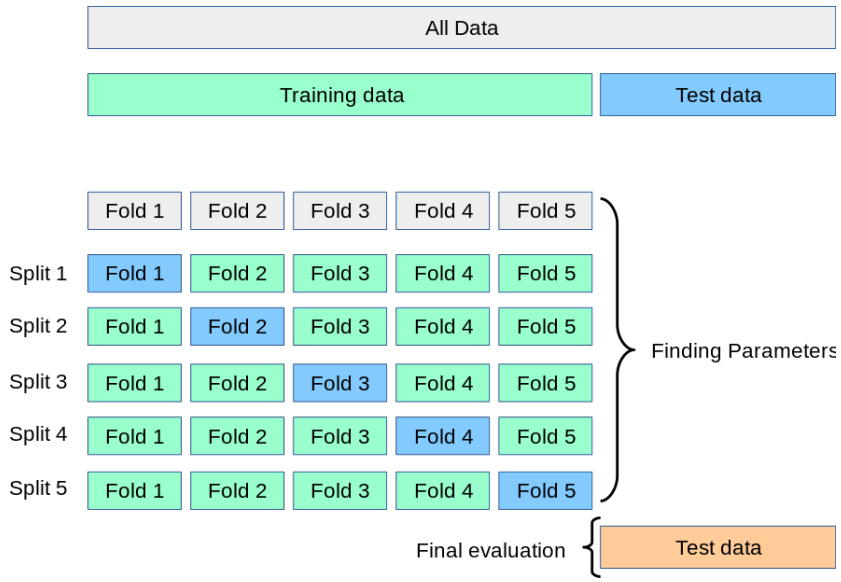

Figure 27: K-fold cross validation. Source: scikit-learn.org

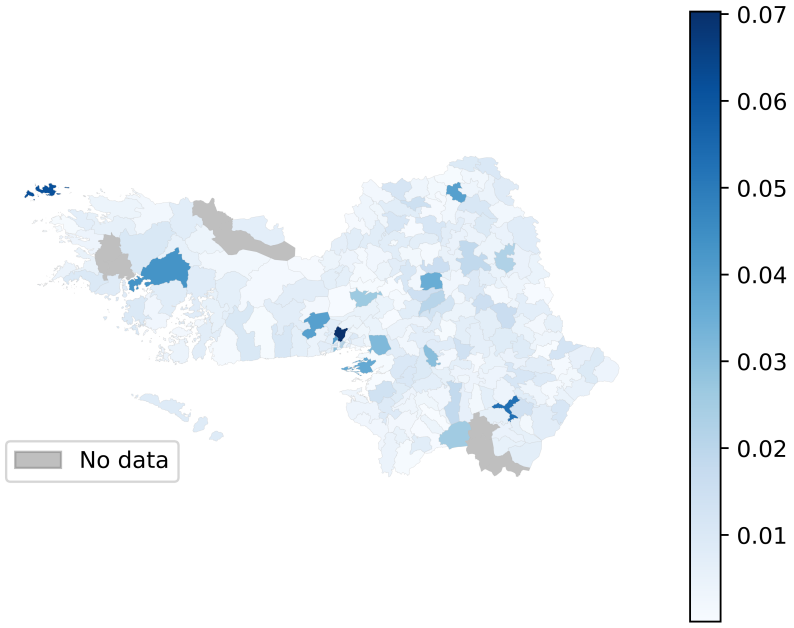

Figure 28: Regression model absolute error predicting 2022 data by electoral division

4.3 Comparing and contrasting the models

The regression model had a far lower error level than the initial epidemiology model. The epidemiology model had a MAE of 0.02635 while the regression model had a MAE of 0.00805. In this state, the regression model was likely to be the more accurate choice. After the initial parameter tuning process, the MAE in the epidemiology model dropped to 0.00949. This value was still higher than the regression model, but far closer.

5 Results

5.1 Future Irish usage projection by model

5.1.1 Regression Model

The regression model predicted modest increases in the proportion of daily speakers for some electoral divisions in the Gaeltacht area and modest decreases for others. There were only minor changes in the proportion across east Galway. The mean predicted change was 0.00200. This shows a prediction of overall modest growth, with decline in some areas and growth in others.

5.1.2 Epidemiology Model

Before tuning the initial guess parameters, the epidemiology model predicted sharp declines across all of the Gaeltacht areas, with the proportion remaining largely constant in the rest of the county. Any area with a high proportion of daily Irish speakers showed a large decline in the proportion, although the predicted figures for 2027 were still significantly higher than elsewhere in the county. The mean predicted change was -0.01785. This is a very substantial decline overall and caused by the huge predicted

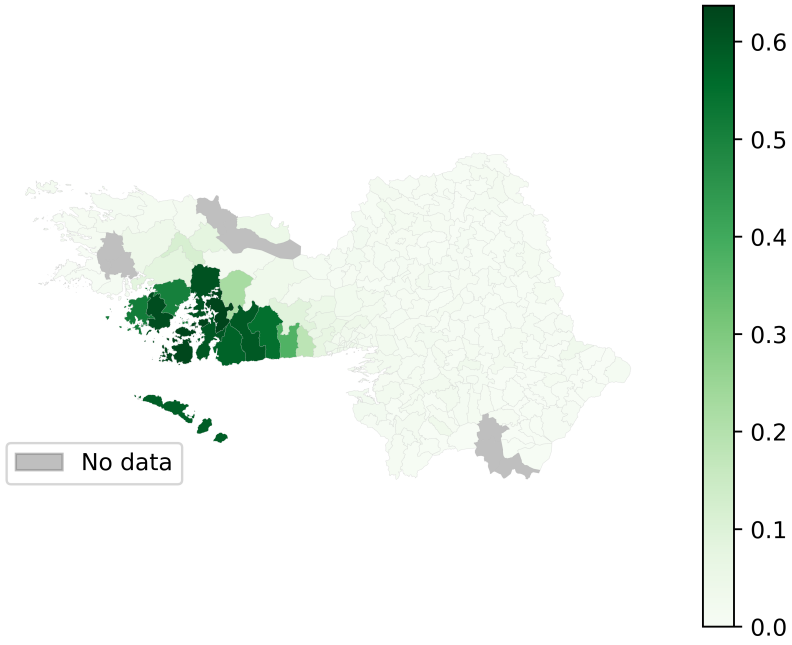

Figure 29: Regression model, proportion of daily Irish speakers prediction, 2027

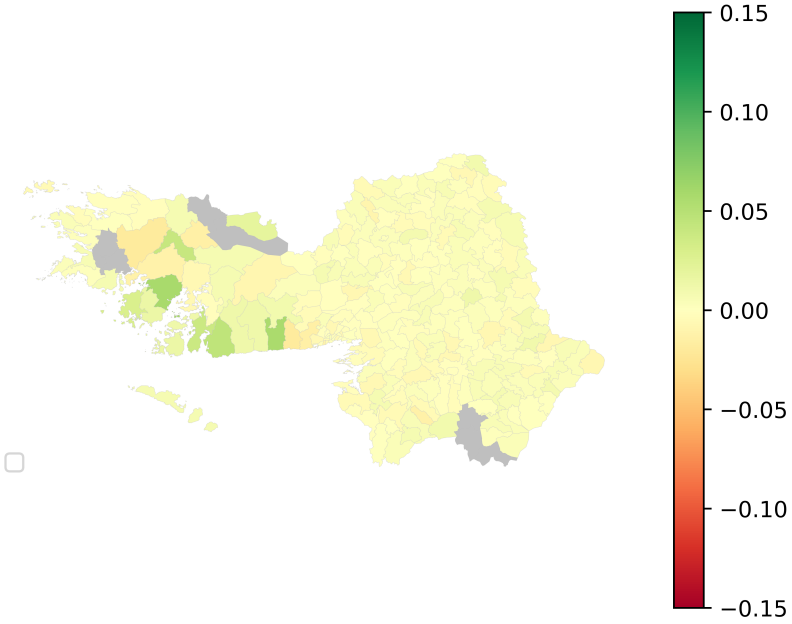

Figure 30: Regression model, predicted change in proportion of daily Irish speakers, 2022-2027

decline in Connemara.

After tuning the parameters, the epidemiology model still predicted a decline in many EDs in Connemara, but it was a far lower decline and not universal to every single ED. It also predicted moderate growth in some other areas of Connemara and in EDs outside Connemara. The mean predicted change was 0.00057. Overall in terms of purely

proportions this shows a prediction of some growth, albeit an almost negligible one.

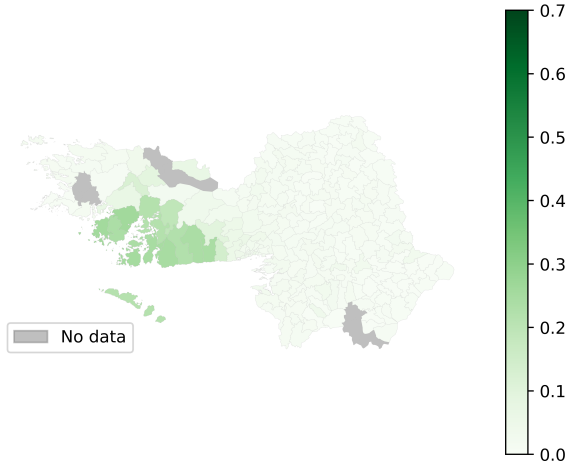

Figure 31: Epidemiology model prior to tuning, predicted proportions of daily Irish speakers
in 2027.

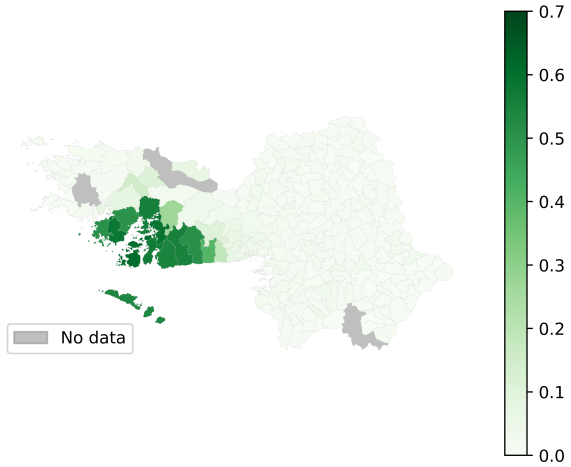

Figure 32: Epidemiology model after tuning, predicted proportions of daily Irish speakers in 2027.

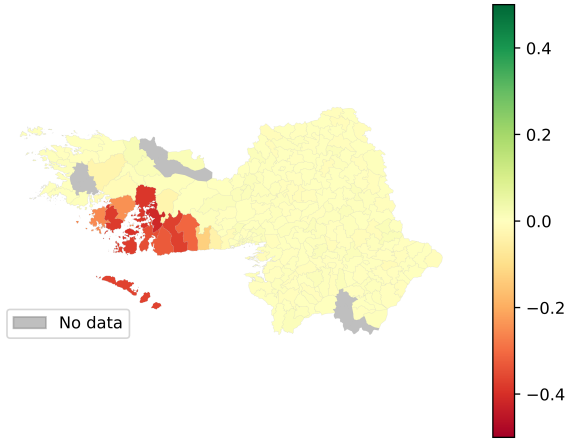

Figure 33: Epidemiology model prior to tuning, predicted change in proportion of daily Irish speakers, 2022-2027.

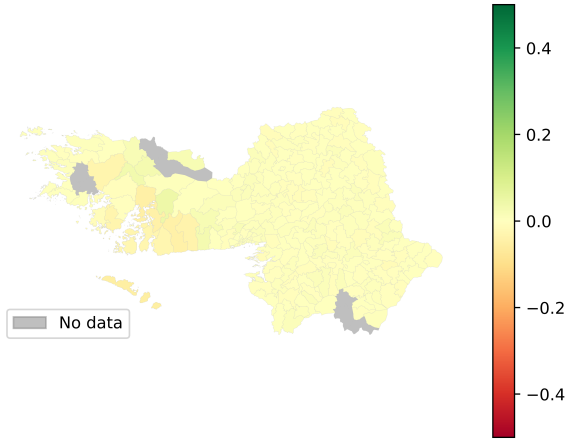

Figure 34: Epidemiology model after tuning, predicted change in proportion of daily Irish speakers, 2022-2027.

5.2 Comparing predictions for two areas

It was chosen to compare two areas with different circumstances. The first area chosen was Leitir Móir (Lettermore). This is a rural island area in Connemara with the highest rate of daily Irish speakers in Galway. The other area was Ballybaan, an urban area in Galway city with a low rate of daily Irish speakers. For Leitir Móir, the regression model (blue) predicts moderate growth from 0.63 to 0.649. The epidemiology model predicts a decline from 0.63 to 0.60. For Ballybaan, the regression model predicted

very slight growth from 0.012 to 0.013 while the epidemiology model predicted a larger growth from 0.012 to 0.017.

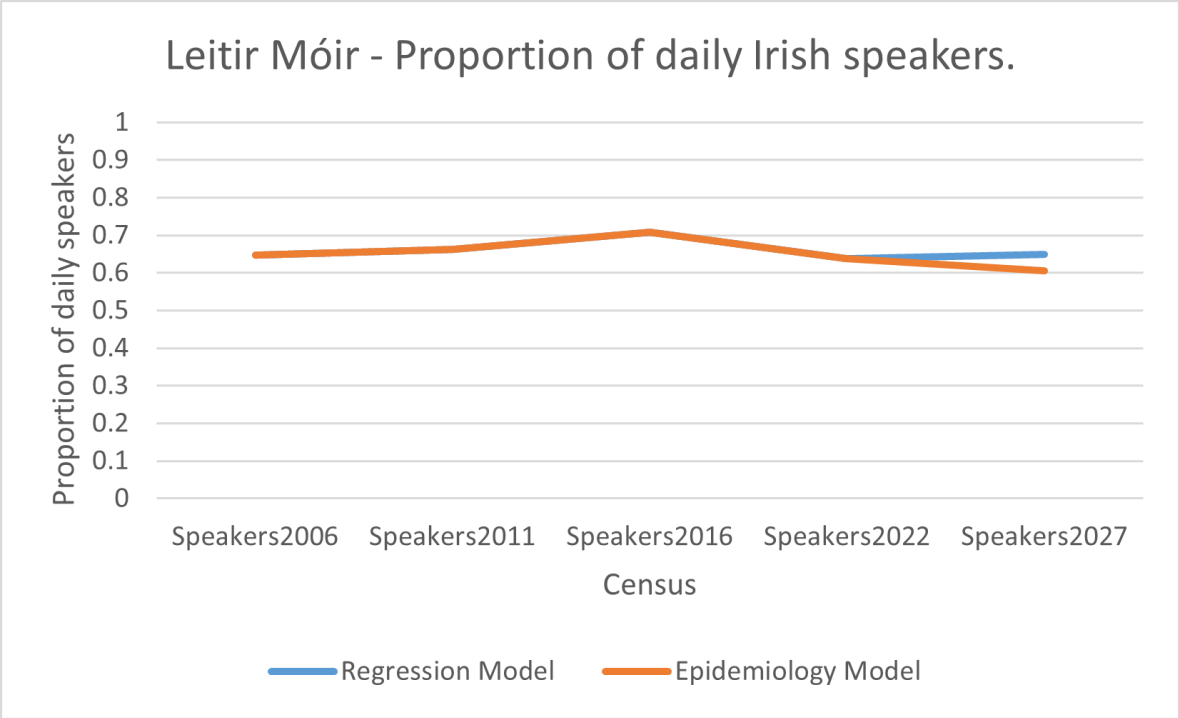

Figure 35: Proportion of daily Irish speakers over time in Leitir Móir

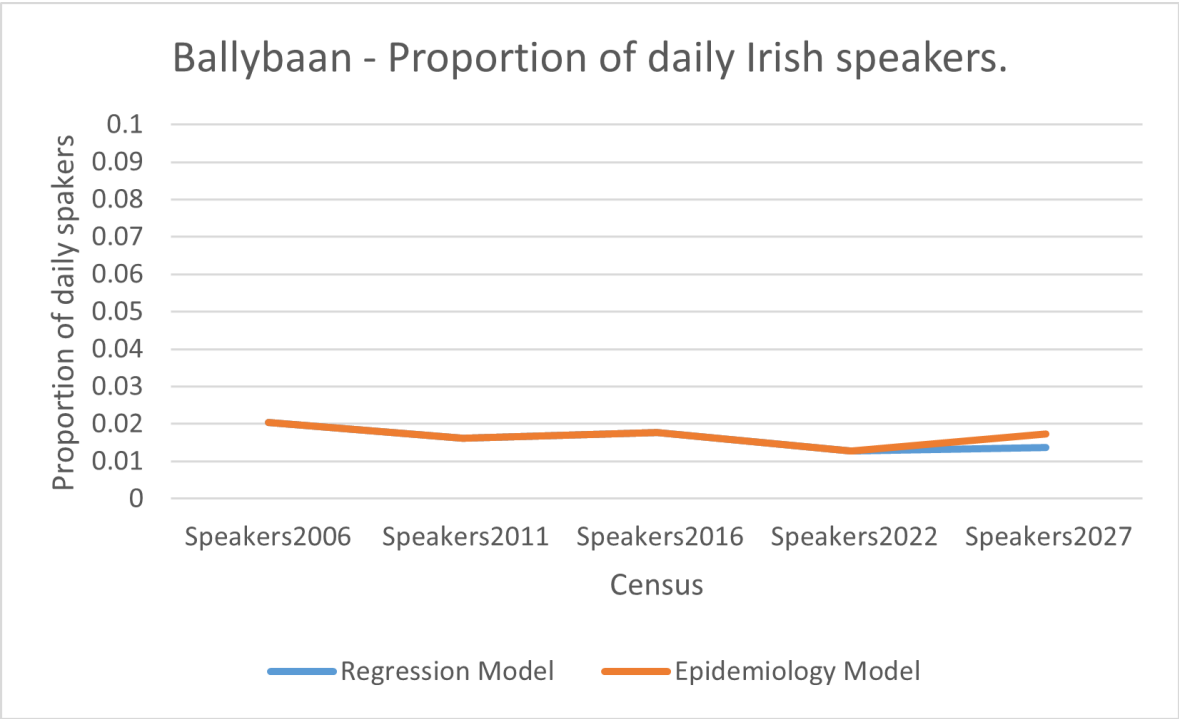

Figure 36: Proportion of daily Irish speakers over time in Ballybaan (with 0 to 0.1 axis)

6 Conclusion

6.1 Error

The MAE of both models was close. The regression model had a MAE of 0.00805. The epidemiology model had a MAE of 0.00949.

6.2 Future trends

According to both the the Epidemiology model (post tuning) and the regression model, the 2027 census is predicted to show on average a very modest increase in the proportion of daily Irish speakers. The epidemiology model predicts a mean increase of 0.00057. The regression model predicts a mean increase of 0.00200. The regression model predicts the regions with the largest increases will be in Connemara. The Epidemiology model predicts modest decreases across Connemara with very modest growth in other parts of the county.

6.3 Value of models

There is no conclusive answer on which model is "better". The upsides of the regression model included being easier to train and better at capturing complex relations, as well as having a lower mean absolute error. The upsides of the epidemiology model were explainability, with the abandonment fraction and probability of acquisition parameters being explainable factors as contrasted with the inherent difficulties that a neural network has with explainability.

6.4 Societal effects

The area of language modelling is - in the opinion of the author - vitally important. As discussed earlier in this report, linguistic diversity is under threat around the world. The world is currently in the midst of a mass language extinction. If nothing is done and smaller languages are allowed to fade into history without any supports, humanity will be worse off without them and the cultures that they support. Closer to home, Gaeilge is an important part of the Irish national identity and indeed of the identity of the University of Galway. By learning more about language modelling and implementing it as part of an innovative new approach, it may be possible to preserve the identity of groups all over the world for future generations.

6.5 Possible future improvements

A number of future improvements could be suggested for any one wishing to pursue a similar project in the future.

6.5.1 Inclusion of 1996 and 2002 census data

The 1996 and 2002 censuses were excluded due to not including the clarifying questions about whether Irish was being spoken inside or outside the education system. As a result it was determined there was a high risk of overstatement in the numbers speaking Irish

regularly as the criteria applied to the data focused on those who spoke daily outside the education system. As discussed in the data section, it would be possible to make an informed estimate of what proportion of speakers were speaking within or outside the education system. This could be done by looking at the proportions in later censuses or finding data from another source about the numbers of students and teachers in education at the time.

6.5.2 An agent-based approach

During the planning stage of the project, an agent-based approach was suggested. Agent-based approaches are prominent for mesoscopic and microscopic modelling. A problem with a microscopic agent based approach is when dealing with population sizes. Many electoral divisions have populations in the 10s of thousands, and attempting to individually model each one and its interactions with the others could introduce huge overhead. One workaround would be to scale down the number of agents by a factor of 10. Alternatively, the same number of agents could be assigned to each ED with some weight applied to their influence based on population size. A mesoscopic approach could be taken by having one agent represent each ED.

6.5.3 A network theory approach

The electoral division map (like any map) can be represented as a network of nodes and edges. Each node corresponds to an ED, while the edges show the bordering EDs. A network based approach could give a stronger weight to nodes with a higher population. Python libraries such as networkX could be used to relatively easily implement networks, although a mechanism would need to be devised to translate a map into a network if one is not already available. Network topologies are relevant for an agent-based approach, where topologies like the Watts-Strogatz small world model can closely approximate human social networks.[3]

6.5.4 Expanding the geographic area

It should be possible to extend the area being covered by such a project with minimal modification to the code base. The census information used in this project was taken from a dataset that covers the republic of Ireland. The scope of the project could be expanded to cover the entire country. One problem would be finding where electoral divisions have been changed / merged. The geographic mapping data also covers the entirety of the country so minimal code modification would be required there.

One interesting possibility would be to explore the census data available for northern Ireland. The administration in northern Ireland collects data about Irish speakers and the frequency of speaking Irish as of their most recent census. An inspiring researcher would need to explore for how many censuses the data is available and compare it to the data in the republic, as questions may not be the same. They also collect information about speakers of Ulster Scots, which could be studied in its own right or in comparison with Irish. Language in northern Ireland (as everywhere) is connected to ethnicity and identity. Ethnically Scots-Irish Protestants are more likely to speak Scots, while ethnically Irish Catholics are more likely to speak Irish[9].

There are, however, a number of difficulties with northern Irish data. For one, censuses in northern Ireland do not always coincide with censuses in the republic. Northern Ireland has censuses every 10 years while the republic does every 5. Northern Ireland also collects data on a "ward" level[4]. It is unclear if wards are comparable in size to electoral divisions. The most recent census in northern Ireland was in 2021, while that in the republic was in 2022 which may raise questions over whether the data can be put in the same dataset. Unlike extending the study to other parts of the republic, extending the study to northern Ireland would likely require a lot of adaptation of the code.

[4]Northern Irish census information : https://www.nisra.gov.uk/statistics/census

7 External resources

- Multi-threading was implemented using information from Python's official documentation
- Central Statistics Office geographical boundary files
- Cross validation
- Project github page

References

[1] 1984 interview with noam chomsky. https://chomsky.info/1984/.

[2] Feedforward neural networks. https://deepai.org/machine-learning-glossary-and-terms/feed-forward-neural-network. Accessed 26/03/24.

[3] Small world networks. http://www.scholarpedia.org/article/Small-world$_n etwork. Accessed$26/03/24.

[4] D. M. Abrams and S. H. Strogatz. Modelling the dynamics of language death. *Nature*, 424(6951):900–900, 2003.

[5] Z. An, Q. Pan, G. Yu, and Z. Wang. The spatial distribution of clusters and the formation of mixed languages in bilingual competition. *Physica A: Statistical Mechanics and its Applications*, 391(20):4943–4952, 2012.

[6] I. Baggs and H. I. Freedman. Can the speakers of a dominated language survive as unilinguals?: a mathematical model of bilingualism. *Mathematical and computer modelling*, 18(6):9–18, 1993.

[7] P. Boissonneault, M; Vogt. A systematic and interdisciplinary review of mathematical models of language competition. *Humanit Soc Sci Commun 8, 21*, 2021.

[8] X. Castelló, L. Loureiro-Porto, and M. San Miguel. Agent-based models of language competition. *International journal of the sociology of language*, 2013(221):21–51, 2013.

[9] D. M. G. Chríost and J. Aitchisont. Ethnic identities and language in northern ireland. *Area*, 30(4):301–309, 1998.

[10] S. Drude, I. C. H. U. A. H. E. Group, et al. Language vitality and endangerment. 2003.

[11] K. e. a. Hale. Endangered languages: On endangered languages and the safeguarding of diversity. *Linguistic Society of America*, 1992.

[12] N. Iriberri and J.-R. Uriarte. Minority language and the stability of bilingual equilibria. *Rationality and Society*, 24(4):442–462, 2012.

[13] A. Laruelle, N. Navarro, and R. Escobedo. Proficiency, attitude, and conventions in minority languages. *Sociological Methods & Research*, 48(2):350–368, 2019.

[14] R. D. Parshad, S. Bhowmick, V. Chand, N. Kumari, and N. Sinha. What is india speaking? exploring the "hinglish" invasion. *Physica A: Statistical Mechanics and its Applications*, 449:375–389, 2016.

[15] K. Prochazka and G. Vogl. Quantifying the driving factors for language shift in a bilingual region. *Proceedings of the National Academy of Sciences*, 114(17):4365–4369, 2017.

[16] S. J. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.

[17] J. Wyburn. The need for language planning to address english-language media pressures on minority language survival in bilingual populations. *Journal of the Operational Research Society*, 71(9):1357–1374, 2020.