Modelando a dinâmica da morte de idiomas

Wellington José Leite da Silva

[1]Escola de Matemática Aplicada da FGV (EMAP), Brazil

Resumo. *Um língua carrega não só uma forma de se comunicar como também séculos de histórias, costumes, lendas, ideias, canções transmitidas de geração em geração, etc. Entender como funciona o processo de morte de uma língua auxilia para evitar o mesmo, neste trabalho tratamos da modelagem de morte de línguas e encaramos as dificuldades vindas disto.*

Abstract. *A language carries not only a way to communicate, but also centuries of stories, customs, legends, ideas, songs passed down from generation to generation, etc. Understanding how the process of language death works helps to avoid it, in this work we deal with the modeling of language death and we face the difficulties arising from this.*

1. Introdução

Linguagem é um sistema estruturado de comunicação [Fromkin et al. 2003], podendo ser baseada na fala, nos gestos ou na escrita. A linguagem humana é única entre os sistemas conhecidos de comunicação animal, pois não depende de um único modo de transmissão (visão, som, etc.), é altamente variável entre as culturas e ao longo do tempo, oferecendo uma gama muito mais ampla de expressão do que outros sistemas.

Nesse contexto, a morte de uma linguagem é uma situação na qual "uma linguagem deixa de ser usada por uma comunidade" [Crystal 2003]. Além disso, ela também pode ser pensada como um processo que afeta comunidades de fala onde o nível de competência linguística que os falantes possuem, de um determinado idioma, é diminuído. Dentre os diversos fatores que ocasionam a morte de um idioma, podemos citar baixo status socio-econômico, genocídio, não repassar para as crianças, etc. Podemos caracterizar as línguas ameaçadas da seguinte forma [Asonye 2013]:

- **Extinct**: situação em que não há ninguém que fala ou se lembra da língua.
- **Critically Endangered**: uma situação na qual os falantes mais jovens são os atuais bisavós e bisavôs da sociedade, de modo que a língua não é usada para interações cotidianas.
- **Severely Endangered**: neste caso, o idioma é falado apenas pelos avós e outras gerações mais antigas, enquanto a geração dos pais ainda entendem o idioma mas não o falam com seus filhos.
- **Definitely Endangered**: nesta fase, a língua não é mais aprendida como língua materna pelas novas gerações. Os falantes mais jovens são, portanto, da geração dos pais. Nesse estágio, os pais ainda podem falar sua língua com os filhos, mas eles normalmente não respondem na língua.
- **Unsafe**: nesse caso, a maioria das crianças ainda fala a língua dos pais e a possuem como língua materna, mas isso costuma ser restrito a domínios sociais específicos, como as casas dos pais e dos avós.

Atualmente, existem mais de 7.000 línguas faladas ao redor do mundo, mas cerca de 1/3 delas têm menos de 1.000 falantes e, de acordo com a UNESCO, mais de 40% dessas línguas estão em perigo de extinção. Só no Brasil, há 190 línguas ameaçadas de extinção [1], sendo elas, 12 Extinct, 45 em Critically Endangered, 19 em Severely Endangered, 17 em Definitely Endangered e 97 Unsafe.

De acordo com uma reportagem do El Pais [2], a cada 14 dias morre um idioma. Além disso, quando uma língua morre não se perdem apenas as palavras, mas todo o seu universo cultural. O universo cultural de uma língua inclui séculos de histórias, costumes, lendas, ideias e canções transmitidas de geração em geração. Assim, com a morte de uma língua também, esse universo cultural desaparece, juntamente com diversos e valiosos conhecimentos práticos de assuntos que envolvem desde plantas medicinais e animais, até o funcionamento do ecossistema como um todo. Dessa forma, o dano da morte de um idioma pode ser comparável ao da extinção de uma espécie.

Neste trabalho será descrita a modelagem da morte de línguas, assim como apresentado em [dAbrams and Strogatz 2003], usaremos um modelo equivalente ao descrito no mesmo, porém o processo de estimação de parâmetros sera feito usando inferência Bayesiana e adicionando certos detalhamentos. O presente estudo se encontra organizado da seguinte forma: na seção 2, tem-se a modelagem do problema com a metodologia proposta; na seção 3, a modelagem trabalhada é aplicada a algumas línguas apresentando resultados ; na seção 4, a discussão sobre o que obtemos e na seção 5 as conclusões do trabalho.

2. Metodologia

Com o proposito de modelar a morte de idiomas, assim como descrita em [dAbrams and Strogatz 2003], optamos seguir um modelo equivalente fazendo modificações quando necessário com o intuito de aplicar o modelo nas línguas em extinção (ou já extintas).

Sendo assim, gostaríamos de modelar uma função $x(t)$, que representa a porcentagem da população que fala um idioma que está morrendo em função do tempo (t). Aqui vamos considerar um sistema onde temos 2 línguas concorrentes, e vamos considerar a existência de um parâmetro s ($0 \leq s \leq 1$) que chamamos de status da língua, o quão uma língua é socialmente mais vantajosa em relação a outra, como o foco é modelar a morte dos idiomas em risco, este parâmetro é conveniente dado que diversas línguas sofrem pressão social de outras línguas, como o inglês que tem dominância em diversos países [Mélitz 1999], espera-se que haja incentivo para falantes de diversos idiomas de mudar.

Sendo $P_{yx}(x(t), s)$ a probabilidade de um falante de uma língua Y mude para uma língua X no tempo t, onde s é o status de X em relação a Y. Uma EDO que modela este sistema pode ser dada por:

$$x'(t) = (1 - x(t))P_{yx}(x(t), s) \; - \; x(t)P_{xy}(x(t), s) \tag{1}$$

[1] http://www.unesco.org/languages-atlas/index.php

[2] https://brasil.elpais.com/brasil/2016/12/26/cultura/1482746256_157587.html

Podemos assumir que ninguém adotará uma linguagem que não tenha falantes ($P_{yx}(0, s) = 0$) ou nenhuma língua com status zero ($P_{yx}(x, 0) = 0$). O artigo inicial [dAbrams and Strogatz 2003] propõe que P_{yx} é da seguinte forma $P_{yx} = cs(x(t))^a$ e $P_{xy} = c(1 - s)(1 - x(t))^a$

Logo, temos o seguinte modelo (com s, c, a e x(0) parâmetros dos dados) onde a função x(t) modela a porcentagem da população em função do tempo:

$$x'(t) = (1 - x(t))cs(x(t))^a + x(t)c(1 - s)(1 - x(t))^a \tag{2}$$

Com o modelo pretendemos modelar a morte de línguas e observar resultados, seguindo, vamos modelar o Gaelic [Withers 1984] língua tradicional irlandesa e escocesa do século um; e o Welsh [Aitchison et al. 1985] língua Galesa de mais de 1400 anos, onde o artigo usa regressão linear para estimar os parâmetros e nos usaremos inferência Bayesiana.

3. Modelagem de mortes de línguas

Os dados que obtemos para ambas a línguas encontram-se nas Tabelas 1 e 2.

Idioma	1881	1891	1901	1911	1921	1931	1951	1961	1971
Gaelic (%)	6,76	6,84	5,57	4,56	3,47	2,97	1,98	1,64	1,78

Tabela 1. Dados Gaelic

Idioma	1951	1961	1971	1981
Welsh (%)	36,8	28,9	26,0	18,9

Tabela 2. Dados Welsh

Calculamos o gráfico das posterioris para cada lingua seguindo as implementações do seguinte notebook [3], temos os gráficos das posterioris dos parâmetro Gaelic na Figura 1 e do Welsh na Figura 2. E também o gráfico da curva para cada idioma, onde para o Gaelic temos os parâmetros $c = -0.1,\ s = 0.99,\ a = 1.5\ e\ t0 = 0.092$ na Figura 3 e para Welsh temos os parâmetros $c = -0.1,\ s = 0.69,\ a = 4.3\ e\ t0 = 0.44$ na Figura 4.

Também temos o R2 referentes a modelagem optada com as observações, onde temos do Gaelic como 0.9881 e do Welsh como 0.8894. As implementações detalhadas encontram-se em jupyter-notebooks encontram-se na pasta Implementações no repositório `https://github.com/wellington36/Modelling-the-dynamics-of-language-death`.

4. Discussão

Escolhemos primeiramente distribuições a priori como Normal porém obtivemos resultados ruins, então decidimos seguir o que é dito em [Shulman and Feder 2004] e escolhemos prioris como distribuições uniforme. Inicialmente buscamos modelar idiomas indígenas brasileiros em extinção porém devido a falta de dados optamos por outros idiomas

[3] `https://github.com/fccoelho/Modelagem-Matematica-IV/blob/master/Planilhas\%20Sage/Aula\%2014\%20-\%20Estimando\%20Par\%C3\%A2metros.ipynb`

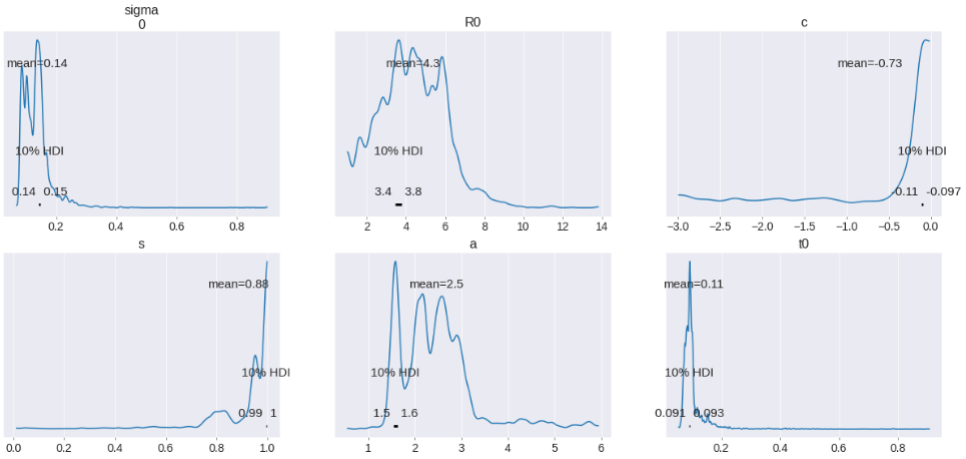

Figura 1. Posterioris dos parâmetro para gaelic

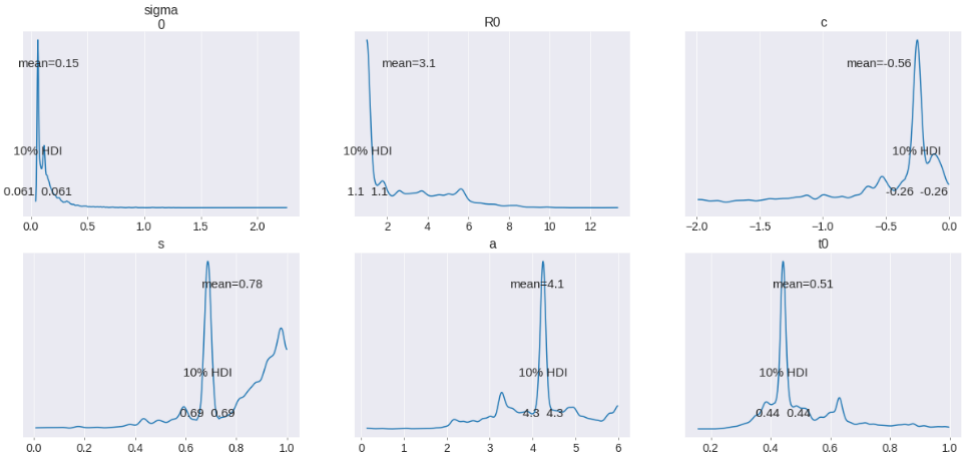

Figura 2. Posterioris dos parâmetro para welsh

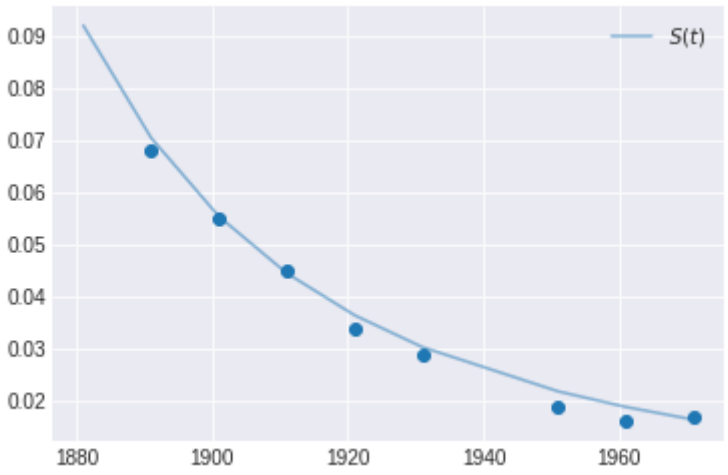

Figura 3. Modelagem gaelic

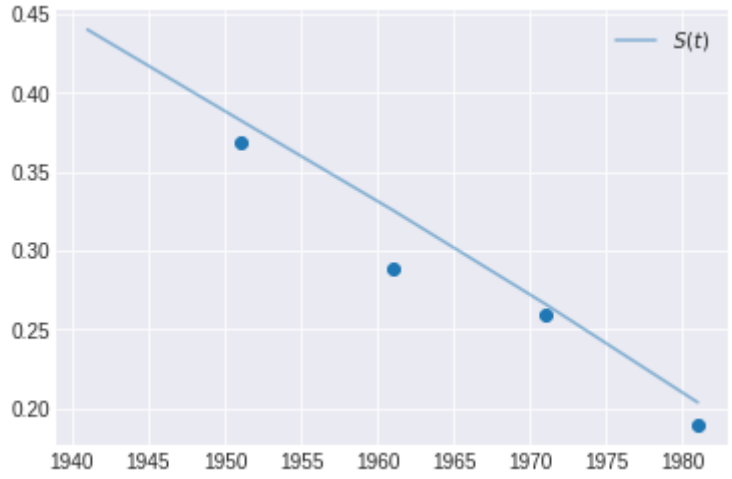

Figura 4. Modelagem welsh

mesmo assim a poucos dados e muito defasados. Os dados do idioma Welsh, como é possível identificar na modelagem dele, tem um certo ruido, o que exemplifica a dificuldade de obter dados para modelar a morte de idiomas.

Um problema é que como a modelagem está, podemos melhorar na mão, por exemplo mudando o parâmetro a da modelagem do Gaelic para 0.11 em vez de 0.1 subimos o R2 para 0.9908, significa que não estamos atingindo o melhor ponto, provavelmente devemos entender melhor a biblioteca usada e melhores formas de incorpora-la aos nossos usos. Apesar disso, o R2 foi relativamente bom.

5. Conclusão

Como mencionado no inicio do trabalho há muitas formas que pode levar um idioma a morte então as modelagens podem variar muito de idioma para idioma. Não há muitos trabalhos nesta área principalmente devido a falta de dados sobre, nesse ponto temos a importância deste trabalho.

Em trabalhos futuros podemos buscar formas de modelar os idiomas indígenas brasileiros e buscar dados melhores. Também podemos fazer mudanças no modelo o que aqui não foi feito apenas no método de estimação dos parâmetros. Além de que podemos entender melhor a biblioteca usada para melhorar a modelagem.

Referências

Aitchison, J. W., Carter, H., and Williams, C. H. (1985). The welsh language at the 1981 census. *Area*, 17(1):11–17.

Asonye, E. (2013). Unesco prediction of the igbo language death: Facts and fables.

Crystal, D. (2003). *A Dictionary of Linguistics and Phonetics*. The Language Library. Wiley.

dAbrams, D. M. and Strogatz, S. H. (2003). Modelling the dynamics of language death. *Nature*, 424(2):1476–4687.

Fromkin, V., Rodman, R., and Hyams, N. (2003). *An Introduction to Language*. Thomson/Heinle.

Mélitz, J. (1999). *English-language Dominance, Literature and Welfare*. Centre for Economic Policy Research London: Discussion paper series. Centre for Economic Policy Research.

Shulman, N. and Feder, M. (2004). The uniform distribution as a universal prior. *IEEE Transactions on Information Theory*, 50(6):1356–1362.

Withers, C. W. J. (1984). Gaelic in scotland 1698–1981: The geographical history of a language (donald, edinburgh, 1984).