Tony Ogaldez and Jonah Stephens

Dr. Swanson

MAT 261

Modeling an Endangered Language

1. Objective

Language is an integral part of humans and their culture, when a language dies a part of what it means to be human is lost as well. There are thousands of languages all over the world that are at risk of becoming lost and one of these is Okinawan. Okinawan is a perfect example of a language growing on its own in a unique way. Although a member of the Ryukyuan family of languages and being itself a Japonic language, it incorporates many parts of Japanese that have been lost and forgotten in standard Japanese. This is more than just case by case, thousands of other languages with rich history and grammar such as this are also at risk of becoming extinct. With better knowledge of the lifetime left for a language then better steps may be taken in order to preserve that language. Okinawan is a dialect of Japanese found in the Okinawa prefecture of Japan. This dialect, considered by some linguists to be a different language, is currently considered critically endangered by UNESCO, meaning that this language is no longer taught in the home or to children[1]. The Abrams-Strogatz[2] and Mira-Paredes[3] studies are the foundation for the objective of our study. Generally, both studies found a decline of a less popular language when compared versus the language with more popularity with both models assuming that the less popular language will die. Using these assumptions, would it be possible to track the trend using each model and relating it to the Okinawan language and to see which model is more accurate to what is happening in the real population as well as determining the AIC and AICc. In short, our objective questions are as such:

- Can we effectively apply the models to our data set?
- Can we determine when Okinawan will die out based on the models?

[1] UNESCO Atlas of the World's Languages in Danger

[2] Abrams, Daniel M., and Steven H. Strogatz. "Modelling the Dynamics of Language Death."

[3] Mira, J., and A. Paredes. "Interlinguistic Similarity and Language Death Dynamics."

- Can we use the AICc to determine which model fits best to our data?

2. Background

We know that popular languages of the world will take up a vast majority of the speakers, the only real competition with Okinawan is standardized Japanese, and the assumptions of the Abrams-Strogatz and Mira-Paredes models hold (mostly) with ours. Using this, what we needed to find is data of the population and models to depict this. It is believed that the speakers of Okinawan would be individuals born before a variety of dates, but most would agree no later than 1970. This lack of documentation of the Okinawan language presented an issue for finding data on speakers. Using an interpretation by Zachary Read[4], we are working under the assumption that speakers of Okinawan were born during or before 1961. Utilizing 5-year census data of Japan[5] and its prefectures, we were able to compile population numbers that matched/closely resembled credible interpretations of the number of speakers by taking the population of age groups in Okinawa and directly comparing them to the total.

3. Data Assumptions

- After every 5 years of data, those 5 years will be cut (in relation to speakers born before 1961).
- The population data is accurate.
- Every person in the population acts how we want them to (speaker population is all speakers).
- Immigration and emigration are ignored.
- Okinawan is not being taught in the household.

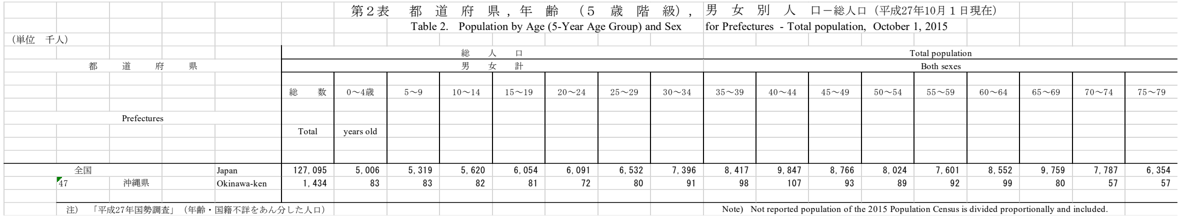
[4] Read, Zachary. Number of Central Okinawan Speakers

[5] "Population Estimates Annual Report: File: Browse Statistics." Portal Site of Official Statistics of Japan

- Standard Japanese is replacing Okinawan.

- The proportion of speakers who do not speak Okinawan speak standard Japanese.

Table 1



第２表　都道府県，年齢（５歳階級），男女別人口－総人口（平成27年10月１日現在）

Table 2.　Population by Age (5-Year Age Group) and Sex for Prefectures - Total population, October 1, 2015

（単位　千人）

都道府県 Prefectures			総人口								Total population								
			男女計								Both sexes								
			総数 Total	0～4歳 years old	5～9	10～14	15～19	20～24	25～29	30～34	35～39	40～44	45～49	50～54	55～59	60～64	65～69	70～74	75～79
全国		Japan	127, 095	5, 006	5, 319	5, 620	6, 054	6, 091	6, 532	7, 396	8, 417	9, 847	8, 766	8, 024	7, 601	8, 552	9, 759	7, 787	6, 354
47	沖縄県	Okinawa-ken	1, 434	83	83	82	81	72	80	91	98	107	93	89	92	99	80	57	57

注）「平成27年国勢調査」（年齢・国籍不詳をあん分した人口）

Note)　Not reported population of the 2015 Population Census is divided proportionally and included.

Census data example (2006-2016)[6]

Table 2

Year	Okinawan Speakers*	Japanese Population*
2006	566	127770
2007	574	127771
2008	584	127692
2009	584	127510
2010	605	128057
2011	527	127799
2012	534	127515
2013	542	127298
2014	551	127083
2015	563	127095
2016	481	126933

Data derived from Census data. *Amount of speakers/population is in thousands.

[6] "Population Estimates Annual Report: File: Browse Statistics." Portal Site of Official Statistics of Japan

4. Models

The three models in question are the Abrams-Strogatz model, the Mira-Paredes model, and the Minett-Wang model, each respectively increasing in complexity. Each model will be discussed, and their parameters and assumptions explored individually.

4.1. The Abrams-Strogatz Model[7]

In 2003, Abrams and Strogatz developed one of the first simple ordinary differential equation models for language decline. Consider an imaginary region in which there are two competing languages *X* and *Y*. The number of speakers of each language is a proportion of the total population of the region, so *x* is the proportion of speakers of *X*, and *y* is the proportion of speakers of *Y*. A major assumption to be made is that these are the only two languages in competition; Abrams and Strogatz also made the assumption that all individuals are monolingual, meaning they do not speak both languages. Abrams and Strogatz derived the following differential equation

Equation 1

$$\frac{dx}{dt} = yP_{YX}(x,\ s) - xP_{XY}(x, s),$$

where P_{YX} is the probability that speakers of *Y* will switch to *X*, and P_{XY} is the probability that speakers of *X* will switch to *X*. Each probability includes variables *x* and *s*; *x* has already been described, but *s* is the "status" of *x* (which will be explained shortly). Interpreting this equation into words, one might read that the rate of change of the proportion of *X* speakers (*x*) with respect to time (*t*) is equal to the proportion of *Y* speakers, which itself is proportional to the

[7] [7]Abrams, Daniel M., and Steven H. Strogatz. "Modelling the Dynamics of Language Death."

probability that *Y* speakers will switch to *X*, minus the proportion of *X* speakers, which itself is proportional to the probability that *X* speakers will switch to *Y*.

Abrams and Strogatz, through further explanation and derivation (namely $y = 1 - x$), further simplified Equation 1 to

Equation 2

$$\frac{dx}{dt} = (1 - x)cx^a s - xc(1 - x)^a(1 - s),$$

where $(1 - x)$ is the proportion of *Y* speakers (*y*), merely simplified into terms of *x*, *c* is the peak rate at which *Y* speakers switch to speak *X*, a is a scaling parameter which determines how *X* becomes more attractive as *x* increases proportionally to its status (*s*). The parameter $0 < s < 1$ represents the status of *X*, which itself is a fairly complex concept. If there are economic incentives for an individual to learn *X* (such as *X* being the main trading language of a society) or social incentives (such as *X* being the primary/official language of a nation), *s* will be closer to one; if there is legislation enforcing the teaching of *Y* in schools or in homes, *s* will be closer to zero. Given that our objective is to model the decline and eventual death of a language, *s* will generally be less than 0.5, meaning that individuals have more reasons to learn *Y* rather than *X*. Additionally, because *s* must be between zero and one, and *s* is the status of *X*, the status of *Y* would therefore be $1 - s$.

As with any model, the Abrams-Strogatz model has several accompanying assumptions:

- Everyone is monolingual.
- Speakers switch languages based on attractiveness of second language.
- "Attractiveness" is based on percentage of speakers and its status.
- Population has a uniform social structure and the interaction among individuals is constant.

- Population is constant across data.
- One language will die.

Additionally, we are forced to make an assumption in order to fit the model to our data: people are still learning Okinawan at a low rate. This assumption is very likely contrary to reality, but the model simply fails to model anything without it.

4.2. The Mira-Paredes Model[8]

In 2004, in an attempt to make the model more applicable to bilingual scenarios, Mira and Paredes added an additional parameter, k, to the Abrams-Strogatz model, resulting in the following:

Equation 3

$$\frac{dx}{dt} = (1-k)(1-x)cx^a s - cx(1-x)^a(1-s),$$

where all the parameters remain the same from the Abrams-Strogatz model, but $0 < k < 1$ is a measurement of the "mutual intelligibility" of *X* and *Y* (how well a speaker of *X* can understand a speaker of *Y* and vice versa). Unsurprisingly, this parameter is really only applicable to dialects; for example, Mira and Paredes used it to compare the Catalonian and Galician dialects of Spanish. Luckily, this model should serve our purposes well, as Okinawan is in a gray area where it can either be considered a dialect of Japanese or its own language, like how Cantonese and Mandarin are officially dialects of Chinese but are different enough that many linguists consider them to be different languages. This also brings a rather nasty consequence: when two languages are not mutually intelligible (e.g., Navajo and English), meaning *k* is very close to zero, the model devolves back into the Abrams-Strogatz model. In the event that *k* is extremely close to one, the model then assumes that *X* will eventually die out in favor of *Y*; but it would not

[8] Mira, J., and A. Paredes. "Interlinguistic Similarity and Language Death Dynamics."

matter anyway, as such a high measurement of mutual intelligibility implies that *X* and *Y* are essentially the same language.

The assumptions for the Mira-Paredes model are as follows:

- All assumptions (save for that everyone is monolingual) are adopted from the Abrams-Strogatz model.
- Some speakers are bilingual.
- Two languages are somewhat mutually intelligible.
- As mutual intelligibility decreases, the model effectively reduces to the Abrams-Strogatz model.

Similarly to our forced assumptions in the Abrams-Strogatz model, the Mira-Paredes model will not work if speakers are not learning Okinawan. Therefore, *c* must be greater than zero, even if it is very small.

4.3. The Minett-Wang Model[9]

In 2005, Minett and Wang published a lengthy paper critiquing the Abrams-Strogatz model, arguing that it is an unrealistic model given the assumption that all speakers are monolingual (implying that if a speaker switches from *X* to *Y*, they forget *X*, which is extremely unrealistic, but it is simple). They put their critiques into action by using the Abrams-Strogatz model as a base for their own system of differential equations model, whose derivation was four pages long. The system follows:

System 1

$$\frac{dx}{dt} = \mu c_{ZX} s(1-x-y)x^a - (1-\mu)c_{XZ}(1-s)xy^a$$

[9] Minett, James W., and William S-Y. Wang. "Modelling Endangered Languages: The Effects of Bilingualism and Social Structure."

$$\frac{dy}{dt} = \mu c_{ZY}(1-s)(1-x-y)y^a - (1-\mu)c_{YZ}syx^a,$$

where many of the parameters remain the same, with more nuance. The variables x and y remain the proportions of speakers of *X* and *Y* respectively; however, the proportions do not add up to one, because there is a third category, *Z*, which represents bilingual speakers. There is no differential equation for *z*, because the assumption is that $x + y + z = 1$. *s* remains the status of *X*, and $1 - s$ remains the status of *Y*. a remains the scaling parameter to measure attractiveness. There are five new parameters, though. μ is the rate at which parents are replaced by their children (we had a difficult time understanding what this parameter ought to be; Minett and Wang's paper was unclear, and it was consistently very high in our modeling). c_{ZX} is the rate at which children adopt *X* from their parents. c_{XZ} is the rate at which adults whose native language is *X* become bilingual. $c_{ZY}$is the rate at which children adopt *Y* from their parents. Finally, $c_{YZ}$is the rate at which adults whose native language is *Y* become bilingual.

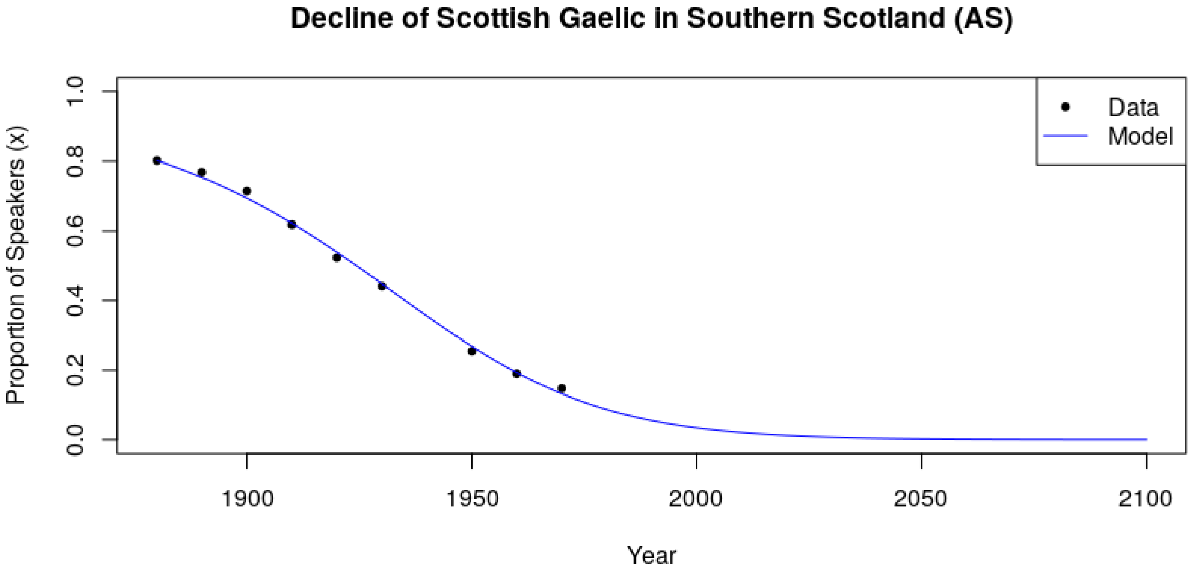
The assumptions that accompany the Minett-Wang model are as follows:

- There are three categories of speakers: monolingual speakers of one language, monolingual speakers of another language, and bilingual speakers who speak both languages.
- Speakers adopt a language based on the attractiveness of said language.
- "Attractiveness" is based on percentage of speakers and the status of the language.
- The attractiveness of being bilingual increases with both the proportion of *X* speakers and *Y* speakers.
- Bilingual individuals remain bilingual throughout their lifetimes.
- There is a uniform social structure; interactions between individuals are constant.
- Coexistence between languages is possible, but one language will overpower the other.

5. Testing the Models[10]

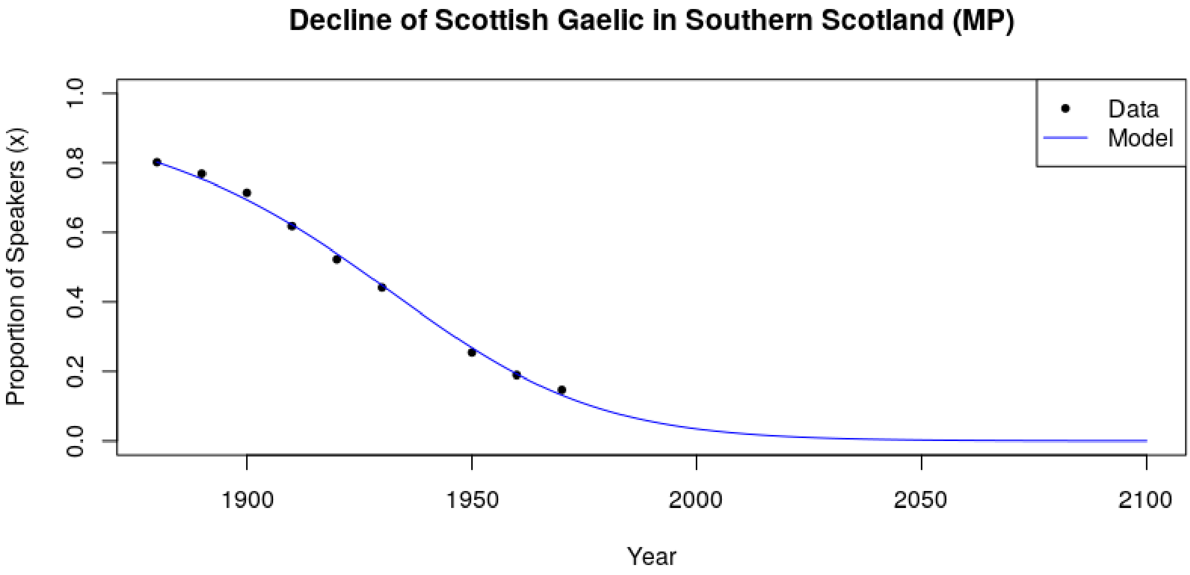
In our initial tests for the models in R-Studio, we used some sample data used by Abrams and Strogatz in their paper. The data shows the proportion of speakers of Scottish Gaelic compared to English in southern Scotland from the years 1880-2000. This data was chosen due to it being the easiest to manage and most controllable set in the paper. Both the Abrams-Strogatz (AS) and the Mira-Paredes (MP) models predicted that Scottish Gaelic would die in the mid 21st century (the proportion of Gaelic speakers becomes microscopically close to zero).

Figure 1. Abrams-Strogatz model of the decline of Scottish Gaelic in Southern Scotland.



[10] Stauffer, Dietrich, et al. "Microscopic Abrams–Strogatz Model of Language Competition."

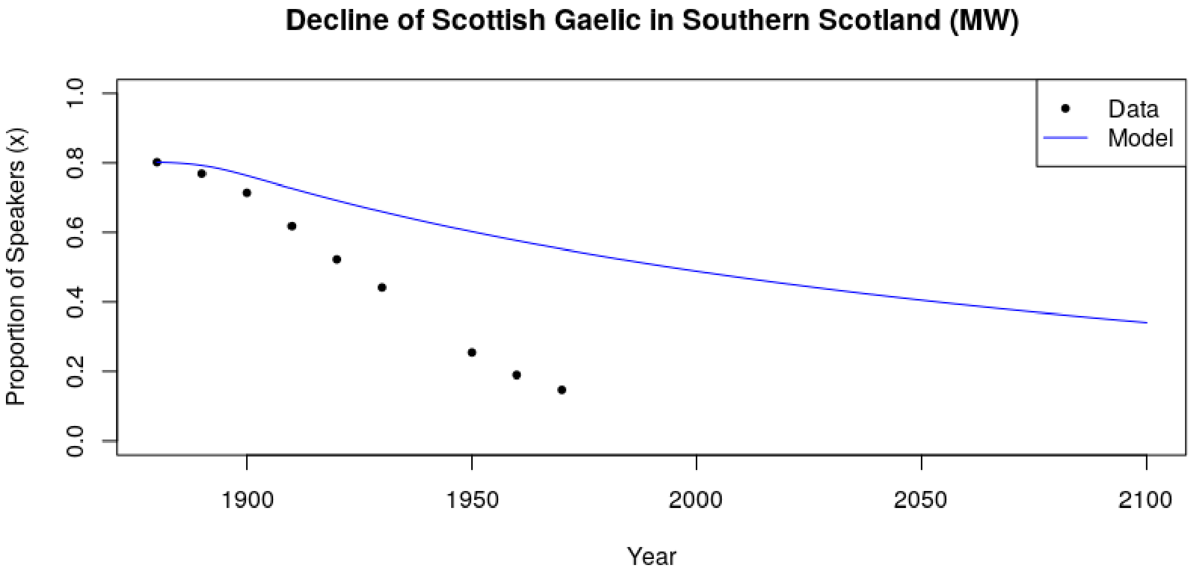
Figure 2. Mira-Paredes model of the decline of Scottish Gaelic in Southern Scotland.



Figures 1 and *2* are visually identical, which tells us that *k* is indeed very small, which is sensible, as Gaelic and English are not in the same language families. Abrams and Strogatz' paper indicated that $s \approx 0.33$, meaning that Scottish Gaelic is consistently about half as prestigious as English, though its attractiveness will continue to decrease as the proportion of speakers decreases. Their paper also indicated that across the cultures they studied, $a \approx 1.31 \pm 0.25$, which was consistent with our findings, as our Abrams-Strogatz model had an a-value of 1.12, and our Mira-Paredes model had an *a*-value of 1.13. Ultimately, $k \approx 0.1$ in the Mira-Paredes model, which is a bit surprising; ultimately, though, the full meaning of a *k* value is unclear, so the 0.1 (which should NOT be interpreted as 10%!) could simply be shared vocabulary.

What of the Minett-Wang (MW) model? It produced some very unsatisfactory results for the test data.

Figure 3. Minett-Wang model of the decline of Scottish Gaelic in Southern Scotland.



The explanation for this is extremely unclear. When using sliders to approximate parameters, the model somewhat fit the data, and it remotely coincided with our results from the other two models. However, when using the modFit command in R-Studio, the parameters found were incredibly inconsistent, and they rarely made sense, resulting in plots like *Figure 3*. An obvious decline is shown, but the decline does not fit with the data nor the assumptions for the Minett-Wang model. Regardless of where the decline initiated, the proportion of speakers of Gaelic consistently approached values less than 0.5, but greater than 0.2. Perhaps this is a problem with the differential equation itself; perhaps it is a problem with our code; perhaps it is a problem with modFit's ability to handle seven parameters. If the model allowed for ultimate coexistence, these results would be somewhat possible to understand and interpret, but it does not. Strangely, though, this model works fine for our data on Okinawa.

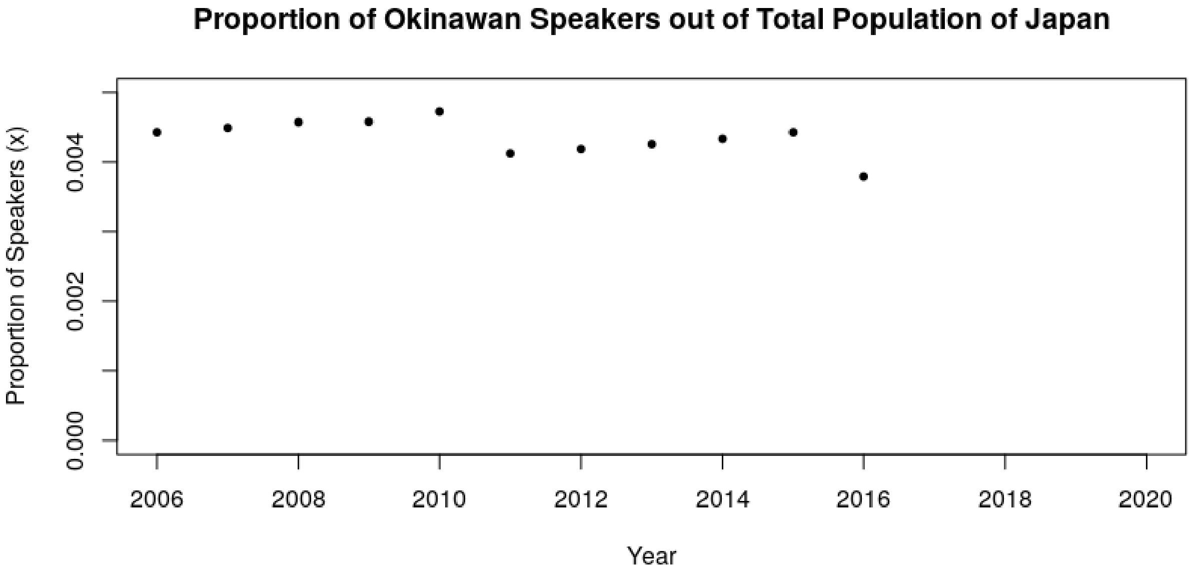
6. Applying the Models

After mostly successful testing, the next step was to apply the models to our Okinawan data and attempt to answer our questions, briefly restated:

- Can we effectively apply the models to our data set?
- Can we determine when Okinawan will die out based on the models?
- Can we use the AICc to determine which model fits best to our data?

6.1. Data Plot

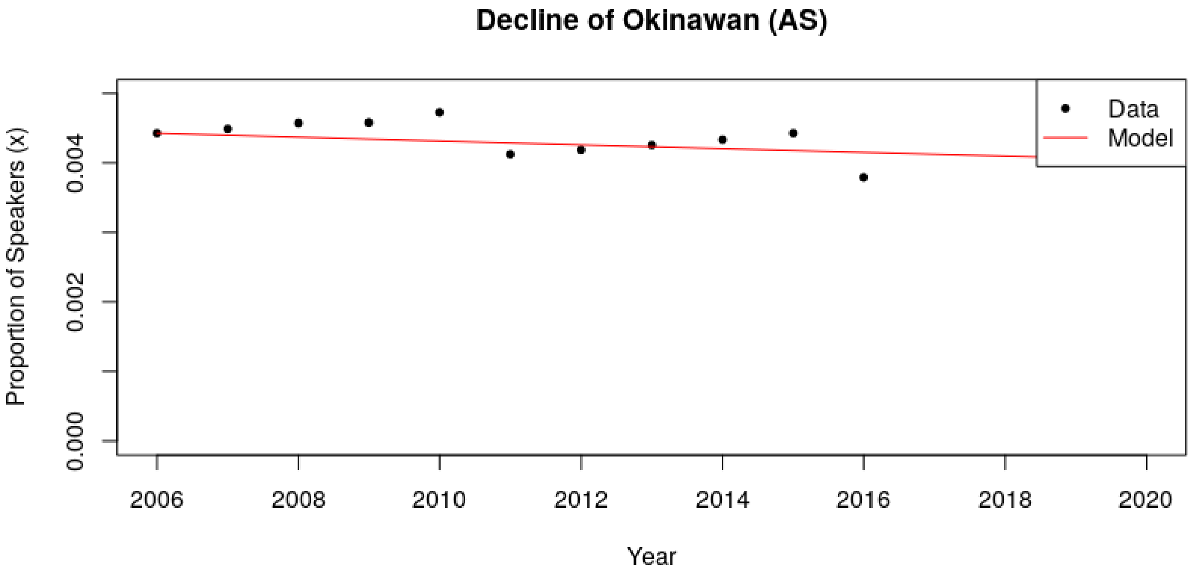
Figure 4. Plot of the proportion of speakers of Okinawan out of the whole population of Japan.



As one can see, the data set in *Figure 4* looks quite strange. In the first two decades of the 21st century, Japan's population has fluctuated, and our proportional data is Okinawan speakers out of the entire population of Japan (approximately 500 thousand out of 128 million). Ostensibly, the best way to present our findings is to show the plot of the model, then explain and answer our questions.

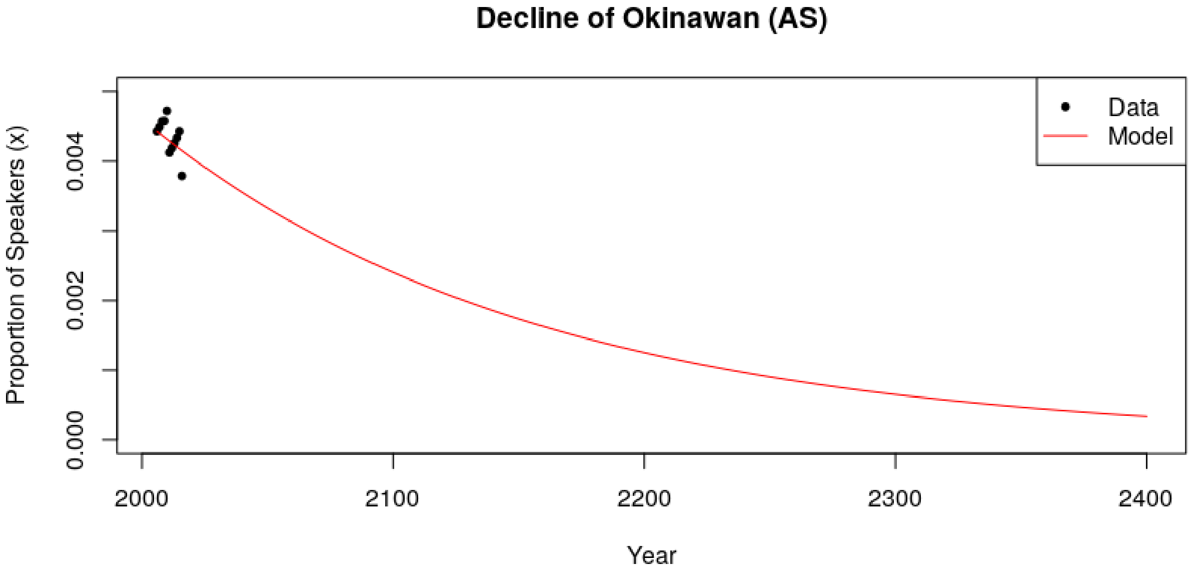
6.2. Abrams-Strogatz Model

Figure 5. Abrams-Strogatz model of the decline of Okinawan (zoomed in).



The timeframe for the data set is very narrow, and the model is not very steep, so let us extend the t-axis.

Figure 6. Abrams-Strogatz model of the decline of Okinawan (zoomed out).

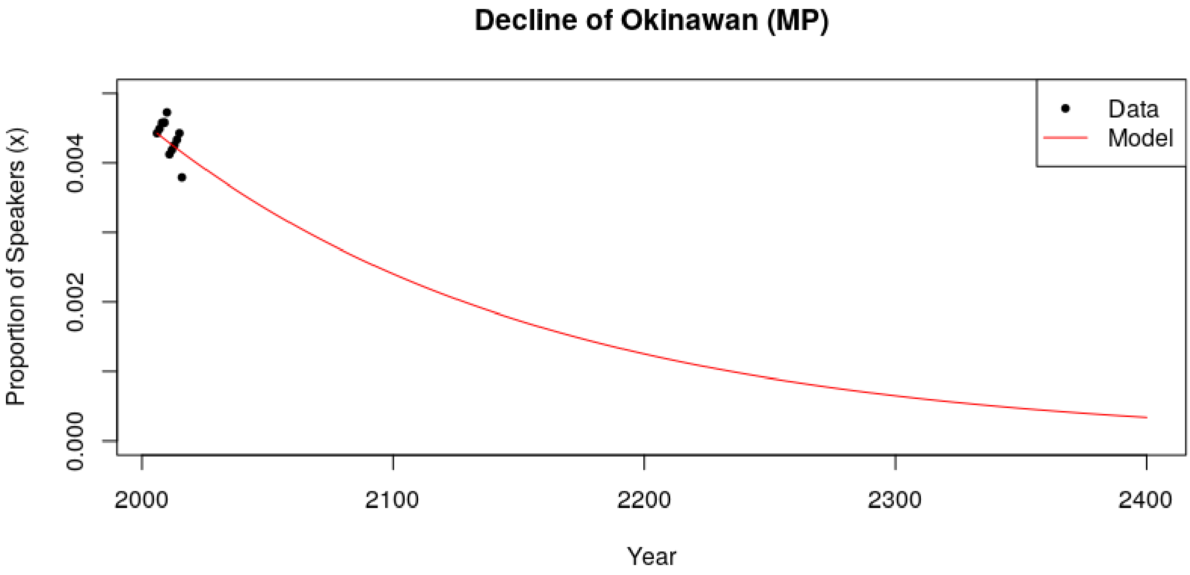


Though the decline is over many centuries, we can see some serious decline. In the model, $a = 1.98$, $s = 0.1$, and $c = 0.0073$. So, a shows that the attractiveness of Okinawan would scale quite rapidly if the number of speakers was growing. *s* being 0.1 is seemingly a very generous estimate, as Okinawan is not taught in schools, is not taught in homes, and is not known by people under the age of at least fifty. Of course, the model does not allow for such details to be implemented, unless we managed to standardize a method for manually determining status, which is totally beyond the scope of this paper. *c* makes sense; this is where our forced assumption comes into play. Because essentially no one is learning Okinawan, *c* should probably be zero; however, the model fails if that is the case, so *c* being this small already is a pleasant surprise. Additionally, that is the PEAK rate at which people learn Okinawan; that does not necessarily imply that people are learning it in any given year.

Now onto the interpretation of the plot itself. The model predicts that Okinawan will die out in either the 25th or 26th centuries. To put that time frame into perspective, now to the 26th century is a similar amount of time between Shakespeare writing his works and now. A language can change tremendously in half a millennium, so we are *incredibly* suspicious of this prediction. Again, speakers of Okinawan are at least in their fifties; virtually no one younger than that speaks it. If we were able to appropriately and effectively incorporate that into the model, we would likely see the death of Okinawan within the next century.

6.3. Mira-Paredes Model

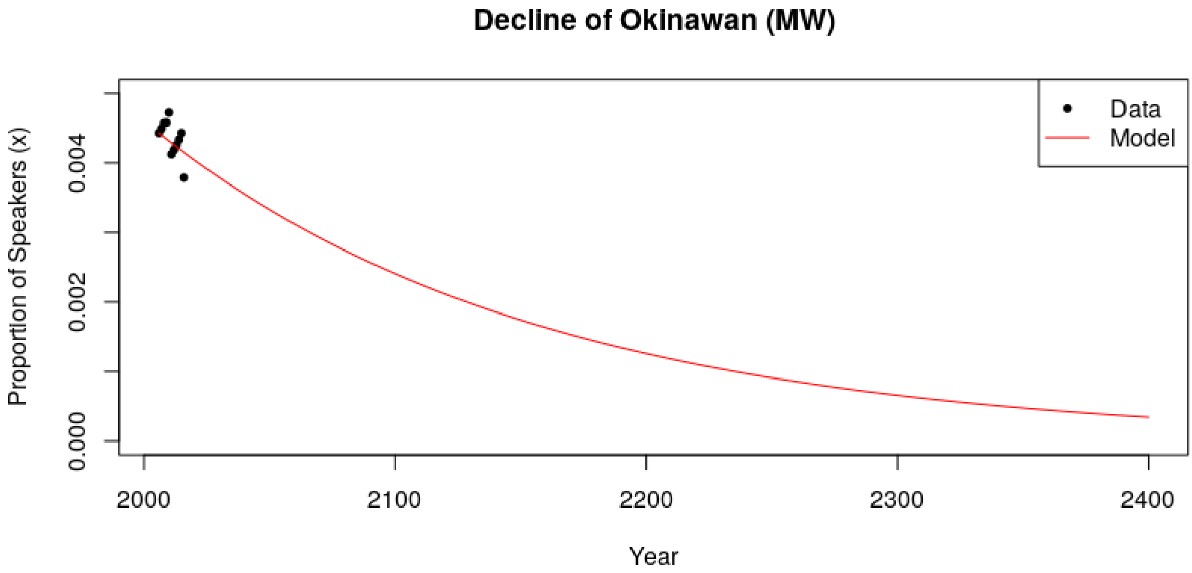
Figure 7. Mira-Paredes model of the decline of Okinawan.



The difference between *Figure 7* and *Figure 6* is subtle. The decline is slightly steeper. Some more distinctions can be seen through the parameters: $a = 1.99$, $s = 0.0002$, $c = 0.0065$, and $k = 0.99$. Both s and c are unsurprising; the status of Okinawan is almost nonexistent, and people just are not learning it. a is very similar to what it was in the Abrams-Strogatz model, so that is not too surprising. k , however, is the anomaly; it is 0.99. That value essentially means that Okinawan and Japanese are the same language (which they may simply be dialects, but that does not justify such a high value). Now, they do have a high degree of mutual intelligibility, but it cannot possibly be that close to one.

Even so, the model predicts Okinawan's death to be a bit sooner than the Abrams-Strogatz model, but it is still much too far in the future.

Figure 8. Minett-Wang model of the decline of Okinawan.



No, this is not a visual trick. The Mira-Paredes and Minett-Wang models resulted in the exact same plot. *Figures 7* and *8* are visually identical, but they are the results of different models. Of course, the distinctions are in the parameters: $a = 2$ (cap) , $s = 0.00003$, $\mu = 0.99$, $c_{ZX} = 1$ (cap), $c_{XZ} = 1$ (cap), $c_{ZY} = 0.000006$, $c_{YZ} = 0.0000001$. The parameter values with ”(cap)” next to them are ones where they always approached their upper limit. If we extended the limit, they would rise to it; if we artificially forced the limit to be lower, it would simply be that limit. *s* was, again, sensible, as Okinawan hardly has any status. μ makes sense if we were working on the assumption that the population remained constant, but we are not, and the model would not have a replacement rate if the population was constant; therefore, that value for μ does not make sense. If *a* was going to continue approaching the upper limit, there is not much commentary to be made on it. Okinawan would be more attractive if more people spoke it, but the status is so low that it hardly matters. The conversion rates seem very off as well. Assuming our understanding of the conversion rate parameters are correct, c_{ZX} should be quite low, c_{XZ} should be quite high, c_{ZY}should be quite high, and c_{YZ}should be extremely low. Only two of those fit our understanding. However, the model is identical to the Mira-Paredes model, so surely there is an aspect that we are neglecting.

7. Finding the AICc

Because our methods of numerically fitting the models already include the sum of squared residuals (RSS), finding the AICc for our models and data should be quick and simple. AICc is an arbitrary measurement of how well a model fits to data; it is done so by plugging in the RSS, the number of data points, and the number of model parameters into a formula. The values for the AICc are not independently significant; they only have meaning when compared to other AICc values. Generally, the model whose AICc is lowest is the one with the best fit, and models whose values have a difference less than or equal to two are considered identical.

To efficiently display the AICc values, they will be inserted into a table.

Table 3

Model	# of Data Points	# of Parameters	RSS	AICc
Abrams-Strogatz	11	3	$5.47 \cdot 10^{-14}$	-352.85
Mira-Paredes	11	4	$5.47 \cdot 10^{-14}$	-347.62
Minett-Wang	11	7	$5.47 \cdot 10^{-14}$	-310.95

Based on the information in *Table 3* (obtained by the method previously described), we see that the Abrams-Strogatz model was the best-fitting model for our data set. Now, had we more data points, and had our proportions been out of the population of just Okinawa, we likely would have gotten different results.

8. Brief Conclusions

Returning again to our three objective questions, we can model our data somewhat effectively, and we can use the AICc to determine which model best fits our data, but we cannot accurately predict when Okinawan is going to die. As previously stated, we have information that is useful in our analyses, like the fact that nearly all Okinawan speakers are over fifty years old, but it was not useful in our modeling, and it could not have been. Such nuance cannot be managed with a differential equation without the risk of becoming obtusely complex; the Minett-Wang system is complex enough.

In a way, it both is and is not surprising that the Abrams-Strogatz model turned out to be our best-fitting model. It was the simplest, which complements our small dataset, and it comes with some extremely hefty assumptions, but its simplicity allowed for it to be easily understood, and it kept it from becoming bogged down in its own complexity. The Mira-Paredes model probably has a lot more potential for the Okinawan problem; the main issue is that we do not have enough data, and we do not have a great way of approximating its unique parameter. Finally, the Minett-Wang model is a fascinating and frankly genius system to model language dynamics (which is its intent), but it is unbelievably easy to become wrapped up in its parameters, both their quantity and meaning, which inclines us to say that it is not well-suited for modeling language death.

Though we were unable to answer our questions with an emphatic "yes!," we still managed to replicate the Abrams-Strogatz model and apply all three models, even if with some unsatisfactory results. Perhaps most importantly, we managed to develop such an intricate understanding of the models that we were able to predict and identify the faults of the models before we even finished modeling.

Sources

Abrams, Daniel M., and Steven H. Strogatz. "Modelling the Dynamics of Language Death." Nature News, Nature Publishing Group, https://www.nature.com/articles/424900a.

Minett, James W., and William S-Y. Wang. "Modelling Endangered Languages: The Effects of Bilingualism and Social Structure." Lingua, North-Holland, 29 June 2007, https://www.sciencedirect.com/science/article/pii/S002438410700071X.

Mira, J., and A. Paredes. "Interlinguistic Similarity and Language Death Dynamics." ArXiv.org, 18 Jan. 2005, https://arxiv.org/abs/physics/0501097.

"Population Estimates Annual Report: File: Browse Statistics." Portal Site of Official Statistics of Japan, https://www.e-stat.go.jp/en/stat-search/files?page=1&layout=datalist&toukei=00200524&tstat=000000090001&cycle=7&month=0&tclass1=000001011679&cycle_facet=tclass1%3Acycle&tclass2val=0.

Read, Zachary. Number of Central Okinawan Speakers – Jlect. https://www.jlect.com/downloads/Number-of-Central-Okinawan-Speakers.pdf.

Stauffer, Dietrich, et al. "Microscopic Abrams–Strogatz Model of Language Competition." Physica A: Statistical Mechanics and Its Applications, North-Holland, 14 Aug. 2006, https://www.sciencedirect.com/science/article/pii/S0378437106008181.

UNESCO Atlas of the World's Languages in Danger, http://www.unesco.org/languages-atlas/index.php?hl=en&page=atlasmap.