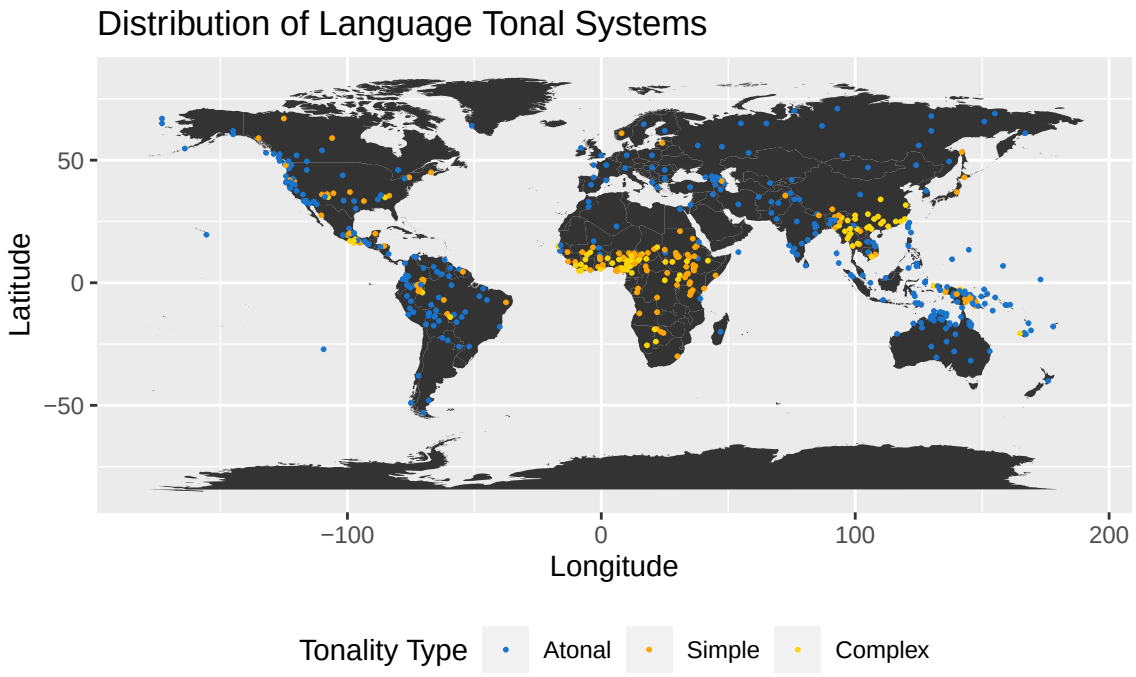
The Geographic Determinism of Complex Tone

MA576 Final Project

Ali Taqi

Abstract

In 2015, a paper by Everett and Moran showed that there is a positive correlation between reigonal humidity and the emergence of (complex) tonal language. In addition to recreating the basic results from that report, in this paper, we further explore the authors' hypothesis in two ways. Firstly, in the framework of a logistic model of the complex tonality indicator, we expand on the authors' result by evaluating the performance of alternative predictors other than the mean humidity, which they use. Secondly, using our findings with humidity statistics, we construct a null survival model to illustrate the dissappearence of complex tonality with increasing aridity, in the spirit of the authors' hypothesis. In our results, we find that the inclusion of upper quantiles in constructing alternative humidity statistics reliably gives us more information in accounting for complex tonality; however, considering quantiles too close to the right-tail starts to become problematic. Both these facts carry over in our null survival model. Furthermore, we find that the addition of indicator variables for three particular language families (out of 148 in total) that *contraindicate* complex tonality is statistically significant and seems to provide better overall model outcomes, which suggests accomodation through outlier status. While we attempt to justify their inclusion using sociolinguistic phenomena, whether proper model specification actually calls for the family indicators remains an open question.



Introduction

The sound systems of human languages are generally not thought to be ecologically adaptive, as linguists commonly assume that they are immune to ecological effects. However, a paper by Everett and Moran nullified this common assumption when it showed that there is a positive relationship between regional humidity and the emergence of tonal language. This leads us to conclude that human sound systems, like those of some other species, are in fact influenced by environmental variables.

In order to fully understand the results of this paper, we first provide some background to contextualize why this hypothesis is even true. Everett and Moran summarize the procedural causes of the impact of humidity on vocalization in their paper, citing various phonetic and laryngeal studies, which we summarize in two parts in the next section. First, we give a quick overview of the linguistics behind tonal languages and define what complex tonality is. Then, we establish how there are necessary dynamics (such as precise intonation) that are necessary for complex tone to emerge. Secondly, we briefly discuss the physiology of phonation, establishing the negative impacts of aridity on phonation. By connecting these two phenomena, we arrive at the authors' hypothesis. Now, we dive into the background to understand and contextualize the authors' hypothesis. Afterwards, we discuss what statistical evidence the authors already provide, and discuss our contribution afterwards. That being said, we begin with our discussion on phonemic tone.

Background

Phonemic Tone

We begin with a brief discussion on phonemic tone. In linguistics, a phoneme is defined as a basic unit of sound in language, the smallest distinctive sound that can change the meaning of a word. In turn, *phonemic tone*, refers to the utilization of pitch to convey distinctions in lexical meaning. The use of phonemic tone is estimated to be prevalent in around half the world's languages. These languages employ modulation of the fundamental frequency for a range of purposes, including word-level stress, phrasal stress, and general pragmatic functions. However, only a smaller subset of languages adopt what are called complex tone systems, characterized by the use of three or more *pitch-based phonemic contrasts*. In other words, a complex tonal language is defined as one which includes at least three *pitch-based phonemic contrasts*. Unlike other uses of pitch modulation, complex tone systems demand a higher level of precision in pitch modulation. In other words, the auditory input judged by a listener of any of these languages can be directly affected or disrupted by any imprecision in pitch. A notable example of a complex tonal language is Mandarin Chinese, where the character *ma*, for example, can assume different meanings depending on the tone (phonemic pitch contrast); depending on which tone is used, *ma* could either mean 'mother', 'horse', or other meanings. This suggests that complex tonal languages, especially those with more than three phonemic tones, present articulatory and perceptual challenges, implying the need of specific conditions that are conducive for accurate pitch manipulation for the acquisition of tonality.

Physiology of Phonation

In their paper, Everett and Moran bring a large survey of laryngeal studies to characterize the relationship between phonation ability and the biomechanics of the vocal tract (which is in the larynx). First, they tell us that hydration can impact fundamental frequency, or the lowest sounding pitch, of vocal chords. Namely, dehydration has been shown to reduce amplitude achieved by the vocal tract. Second, we are told that the mere inhalation of dry air impacts vocal cord physiology and results in clear effects on phonation. It is mentioned that low humidity is reported to be the cause of sore throat in 30% of cases. They augment this with the surprising fact that singers tend to have a more difficult time holding a tone in dry climates. Altogether, these facts show that the functionality of the human larynx is susceptible to environmental aridity. Although less important for purposes, it is also worth mentioning the impact of temperature on humidity. Frigid air, irrespective of relative humidity, remains dry due to its diminished water vapor capacity, owing to the variation in water capacity at different temperatures. Air at -10°C reaches water saturation at 2.3 g/m^3, while air at 30°C achieves the same at 30.4 g/m^3. This signifies that a mere difference of 40° corresponds to a 13-fold increase in water vapor capacity, which is substantial. As such, being the hottest region around the Earth, the equator, also hosts the most humid regions.

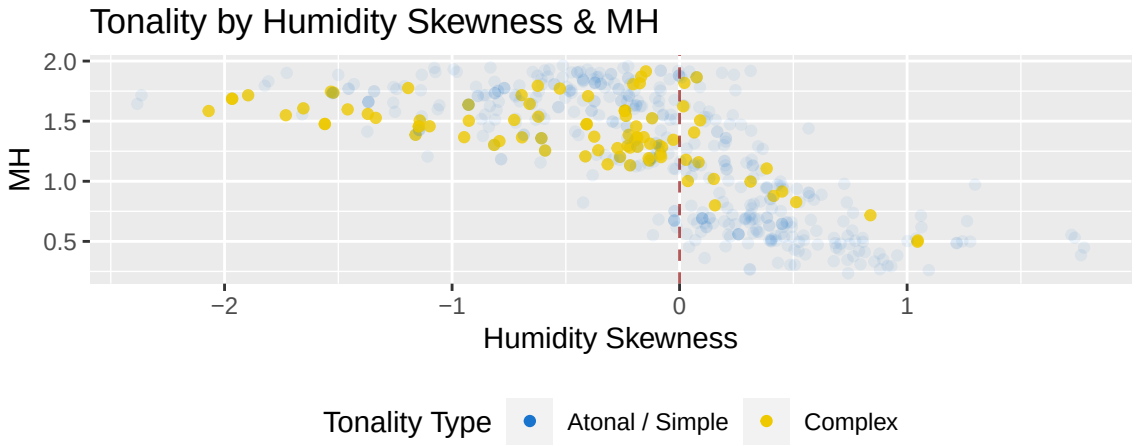
The Authors' Hypothesis

As suggested earlier, those two discussions can be linked together to give an explanation on why humid regions could better sustain the development of complex tone. The authors predict or hypothesize "that languages should not be maladaptive vis-à-vis ambient air conditions" (Everett and Moran, 2015). That is, "like other forms of human behavior, the articulation of phonemes should evolve in accordance with ecological factors that directly impinge on their production. More specifically, complex tone, requiring comparably precise manipulation of fundamental frequency, should be disfavored among extremely arid contexts" (Everett and Moran, 2015). Furthermore, the authors predict that "this tendency should be particularly apparent in frigid climates, given the extremely reduced water vapor capacity of cold air" (Everett and Moran, 2015). To summarize, the authors do not predict complex languages necessarily emerge in humid regions, they predict that complex languages do not arise in arid regions.

Statistical Evidence

In their paper, the authors provide two forms of statistical evidence. First, the more simple form, consists of a difference in means hypothesis test between the complex tonality and atonal/simple tonality group, with mean humidity (abbreviated MH hereinafter) as the predictor. The second form involves resampling methods, where they compare the MH of complex tonal languages and their non-complex counterparts at various percentiles of the MH. They find that in 89%, 88%, 43%, and 49% of samples, the complex tonal group displayed higher MH at the 15^{th}, 25^{th}, 50^{th}, and 75^{th} percentiles of MH, respectively. In context of their hypothesis, the lower percentile results are most significant, as to exclude the possibility of complex tonality in arid contexts by showing a higher baseline humidity.

At first, this result seemed surprising and intuitive. However, one statistic that could elucidate this mystery, is that $\text{corr}(\mu_H, \text{skew}_H) = -0.71$ (in our dataset, at least). That is, there is a strong negative correlation between MH and skewness, implying areas with larger MH tend to observe negatively skewed humidity distributions, and vice versa. As negative skew implies larger left-tails, this could explain why lower quantiles are reliably larger for complex languages (which we already take to observe larger MH).



With this in mind, this is what motivates our desire to explore of the usage of the larger quantiles, in the same inferential framework of using humidity to explain complex tonality. Namely, we would like to know whether the higher percentiles at the right-tails of the humidity distributions have comparatively greater explanatory power. As such, we construct two different humidity statistics in this paper, parameterized in such a way to capture "extremity" of quantiles, and we explore model performance as that parameter increases (and henceforth uses more extreme quantiles).

The Data

Lingustic Data

Dataset Description

For recreating the result in the authors' paper, the primary dataset we utilize is acquired online from the World Atlas of Linguistic Structures of the Max Planck Institute, which we abbreviate as the *WALS* dataset (Dryer and Haspelmath, 2011). This dataset contains $n = 527$ observations of a diverse range of languages. The primary columns in the raw data that we are most interested in are:

- *Family*: This indicates the name of the broader language family to which a language observation belongs to. This could be considered a `factor` variable for modelling purposes, and there are 148 language families.
- *Latitude, Longitude*: This pair of columns determines the geographic location of the language observation. They are both `numerical` variables, where latitude spans ± 90, and longitude spans ± 180.
- *Tonality Type*: A `factor` variable with 3 levels, indicating the tonality system of the language observation: *atonal*, *simple*, or *complex*.

Wrangling Process

For our modeling purposes, we use `dplyr` to wrangle our data and generate a few variables using `mutate`. Most prominently, we create a new binary indicator variable involving *tonality type* for use in modeling: `complex_tonal`, which will become our response variable. Similarly, we do the same for *language family*, creating a binary indicator variable for a particular family/group of families.

Climate Data

Dataset Description

The original humidity dataset was obtained from the American Meteorological Society, with the link in the reference (Kalnay, 1996). It is structured with a fixed grid of 94 latitude and 192 longitude points, with an observation of the humidity for every month between 1949-2022, totaling around 888 observations of humidity for each coordinate pair. We scale the humidity data a few orders of magnitude for practical visual purposes (reading estimates and axes labels), but the original data, according to the source, is unitless.

The original dataset was messy, and was wrangled and organized to a `tidyverse` format which had each row contain exactly one unique latitude, longitude, month, and year. This was necessary to make computations easier in the long-run. In the end, the final dataset had 16,062,720 rows of data, so it was saved to be read and accessed again.

One assumption we checked before we started modeling was if the average (global) humidity stays constant, since we would want our current data to match the humidity conditions back in the past. While we expected to find no change, we unexpectedly encountered a positive correlation of 0.85 between year and global humidity. Granted this was between 1949-2022, this possibly could be an indicator of climate change. Nonetheless, we stuck with using the entire range of years, because that would be better than assuming that the 1949 data is closest to the past, throw out the rest of our observations, and turn out wrong.

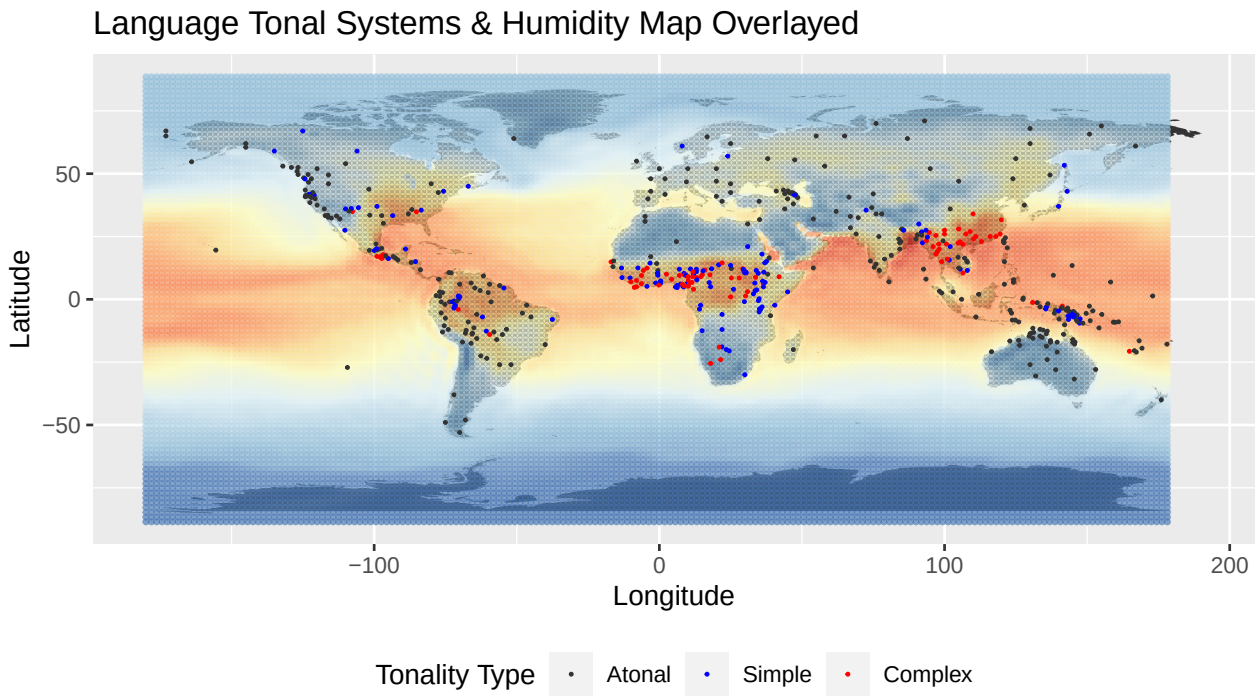
Joining the Datasets

Given these two datasets, we finally discuss the process of joining our linguistic data to our climate data. The first problem encountered was the fact that the grid in the climate dataset was fixed. To remedy this, a helper function (as shown in the pseudocode below) `nearest_coords()` is written to transform any real-valued pair of coordinates to the nearest coordinates in the humidity data. A near-optimal solution is to component-wise minimize the difference between the real coordinate in the range of possible coordinates.

Nonetheless, now, we are ready to combine the two datasets. To do so, we run a loop over our linguistic dataset, and for each language, we take the nearest coordinate pair, filter for the ≈ 900 months worth of

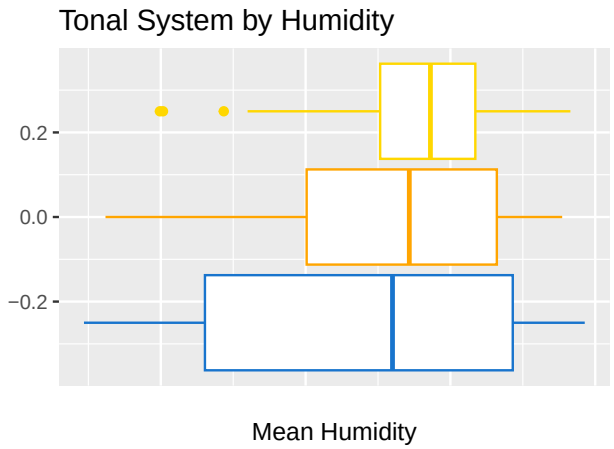
observations, pass it to some statistic function, `hum_statistic()`, and attach the result back to our linguistic dataset. The pseudocode for this process is given below.

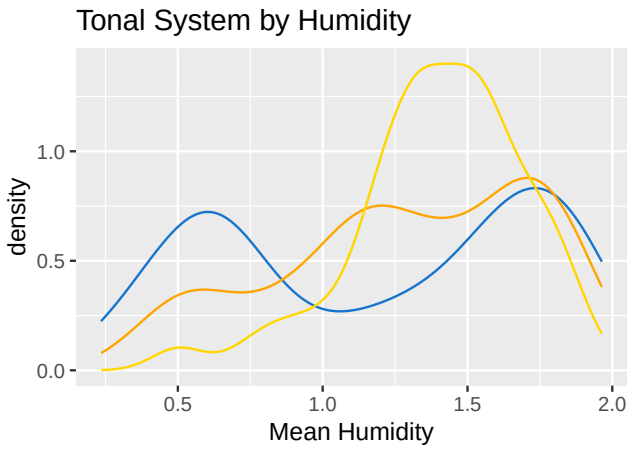
```
for lang in languages{
  # Filter the location for the relevant humidity data from the climate dataset
  lang_hum_data <- get_hum_data(location = nearest_coords(lang))
  # Attach the computed statistic to the linguistic dataset
  lang_data[lang, hum_stat] <- hum_statistic(lang_hum_data)
}
```

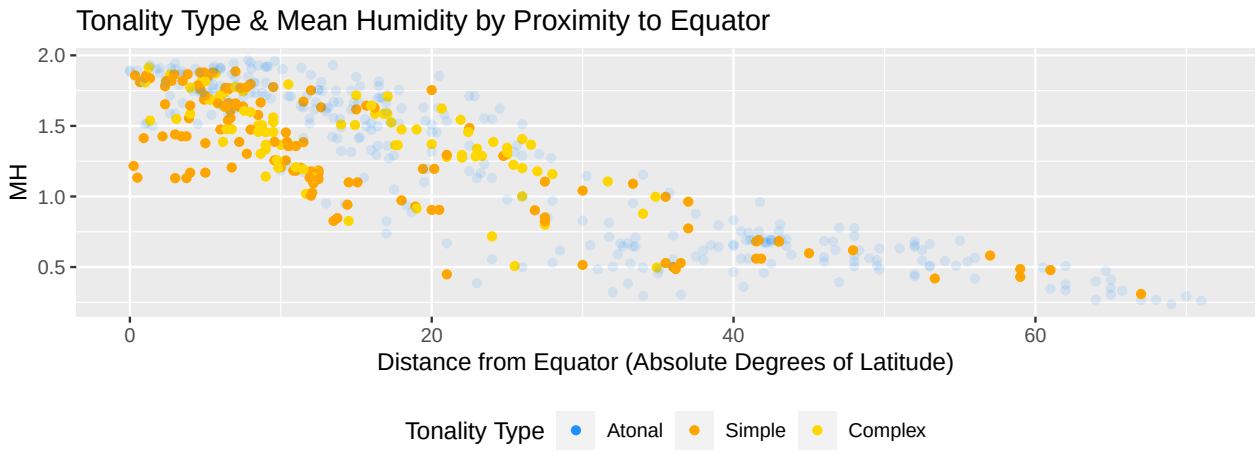


Exploratory Data Analysis

Data Visualization







Methods

Logistic GLMs

As previously mentioned, we will be exploring the ability for upper quantiles of humidity data to better explain complex tonality, doing so through the means of a logistic regression model. As such, we fit our indicator variable `complex_tonal` using some iteration of a humidity statistic with `glm`. Throughout the remainder of this paper, we will stick to using the probit link function for consistency, although performance with the logit link is nearly identical. We might prefer the probit link here due to its slightly thinner tails, and the neat framework of dichotomization our situation calls for. Using our logistic models, we observe heuristics such as reduction in deviance, dispersion, and our coefficients' performance. Using these heuristics, we interpret our results in terms of their inferential implications.

The Family Predictors

In our logistic models, we also explore the inclusion of particular language family indicators. Namely, we first vet out which families are statistically significant through modeling `complex_tonal ~ MH + family`, then select those which have negative coefficient estimates (correlations) to complex tonality, having accounted for humidity already. The purpose of this is not to find indicators solely for the purpose of deviance reduction in the model, as classification is not the goal. Rather, the goal is to isolate potential confounding variables that are causing misspecification-induced underdispersion, which impacts confidence in our inferences.

In our analysis, we simultaneously observe models that include all and none of the the family predictors to contrast their effect. To summarize, the primary angle with which we analyze the role of family predictors an attempt to better specify the model, suggesting that certain language families are to be accommodated with an outlier status. In the discussion, we outline potential rationales for the inclusion of these indicators, linking them to their impact on model dispersion outcomes.

Humidity Statistics

As mentioned, we explore statistics other than MH which leverage more of the upper humidity quantiles to model complex tonality. In this section, we define two alternative mean statistics that are parameterized by a single parameter which corresponds to the extremity of the upper quantiles used. Increasing that parameter, we study our model's behavior, hoping to see if there are any patterns or effects.

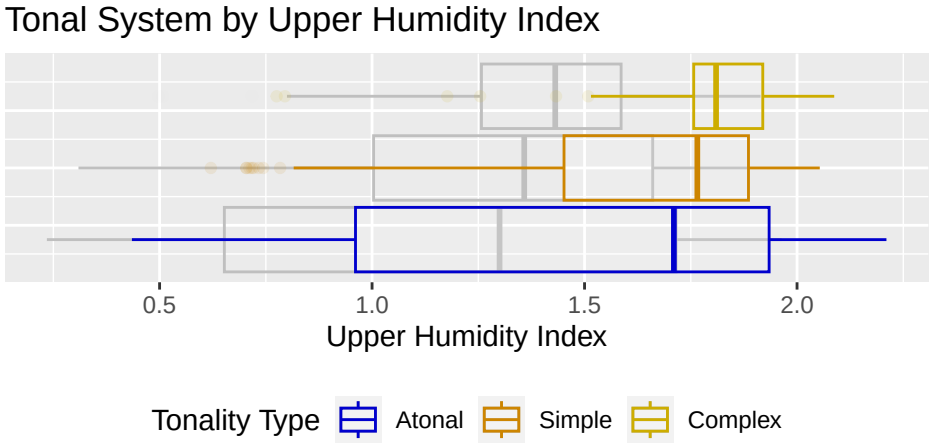
To gloss over it briefly, we have two constructs. The first statistic uses the humidity average over a range restricted by nearby quantiles $\pm\delta\,\%$, where the parameter is the central quantile q and δ is a small fixed percentage, telling us the "window" size. The second statistic is a humidity score, where the parameter k is the multiples of σ_H we add to our mean μ_H. Each case has its own benefits and drawbacks, which we go over in more detail in the discussion section.

1. Quantile Window Statistic

Let h_q^* denote the q^{th} humidity quantile for a fixed geographic point (observed over the entire time range of data). Then, we define a **quantile window statistic** with a fixed δ (percentage) and centered about a quantile q (taken to be the parameter) which we denote $\mu_q^{(\delta)}$ to be the value:

$$\mu_q^{(\delta)} = \mathbb{E}[\mathbb{H}_q] \text{ where } \mathbb{H}_q = \{h \mid h_{q-\delta}^* < h < h_{q+\delta}^*\}$$

Essentially, we restrict the humidity data centered about q to a window of size δ (a fixed percentage) before taking the expectation. The idea of using this is to potentially obtain better separation using the upper quantiles than if we just used the MH. To demonstrate this, $\text{UHI} = \mu_{90}^{(10\%)}$ is shown below to have slightly improved separation over just using MH (shown in light grey). It is precisely this notion of "better" separation that we seek to formally assess and quantify using logistic GLMs.



2. k-SD Score Statistic

For a fixed geographic point, an alternative, more simple statistic we define is the k-SD **humidity score statistic**, which we define as:

$$\widetilde{H}_k = \mu_H + k\sigma_H \text{ where } k \in \mathbb{R}^+$$

Parameterized by the sole (positive) real parameter k, this statistic is essentially a standardized humidity score, where k is a score "slider". Here, μ_H denotes the regular MH, averaging the humidity data over all our months and years, and σ_H is derived analogously. Note that k is strictly positive since we are specifically interested in higher quantiles.

With these two statistics established, we fit our logistic models, varying these parameters, and observing the behavior of our model. That being said, we move on to the last section, which is creating a "null" survival model.

Null Survival Model

Lastly, we construct a "null survival" model to alternatively capture the author's hypothesis in a direct and visual way. It is important to note this is a divergence or extrapolation of survival models in the normal sense, since our model is not temporal, but rather, observed through the axis of "aridity". Our goal is to show that complex tone does not "survive" in arid conditions. To further clarify, we are not working with any predictors, so we call this a "null survival model" because we are interested in estimating the *baseline hazard rate* over the aridity axis. As such, we later define some *modular* notion of aridity d_θ (derived from some humidity statistic θ) from which we estimate the baseline hazard rate

$$\lambda(d_\theta) \sim \mathbf{0}_{\text{Complex}}$$

through the indicator step function where failures are observed non-complex languages (i.e. observed "losses" or "failures" to obtain of complex tonality). In the end, we end up with a interpretable model, with direct allusion into the authors' hypothesis, and numerical estimates we could work interpret.

Discussion & Results

Logistic GLMs

Overview

The models below were constructed to comparatively investigate the explanatory power of our upper-quantile humidity statistics towards the response variable `complex_tonal`. For purposes of continuity and simplicity, we always use a `probit` link, citing the reasons discussed earlier. As mentioned before, our outcome of interest was to establish that upper-quantile-based statistics have comparable if not greater explanatory power in predicting complex tonality than the mean humidity (and its lower quantiles). To reiterate, our logistic GLMs are not to be used for prediction or classification, but rather, inference regarding the information contained in the upper quantiles.

While perfect separation by humidity is not possible, and actually not expected, the logistic GLM remains a powerful tool to assess the power of continuous predictors in explaining dichotomous or polytomous response variables. That is, despite the significant overlap between the two tonal categories, the model still gives us a gateway to *comparatively* assess predictor performance by seeing which ones best minimize loss over all of our observations in aggregate, despite the significant overlap. As such, we should not be surprised to find that the fitted values in our models rarely exceeded $p = 0.5$, a minimum fitted value one would hope for if they were constructing a classifier. So, this would create an estimator with terrible specificity / classification rate, because our data is not structured in a way pliable to class separability. All of this is to say, logistic GLM models still provide a versatile and powerful tool to comparatively assess continuous predictors in explaining indicator response variables, irrespective of class separability. In our case, we leverage this tool to *comparatively* assess different upper-quantile humidity statistics in their ability to explain complex tonality.

That being said, we begin with the discussion of the k-SD score models. After that, we cover the quantile window models. In the end of our discussion of logistic GLMs, we talk about dispersion and the family predictors for all of our models overall, and their inferential implications. Afterwards, we talk about our 'null survival' models, and move on to our conclusion.

k-SD Score Model

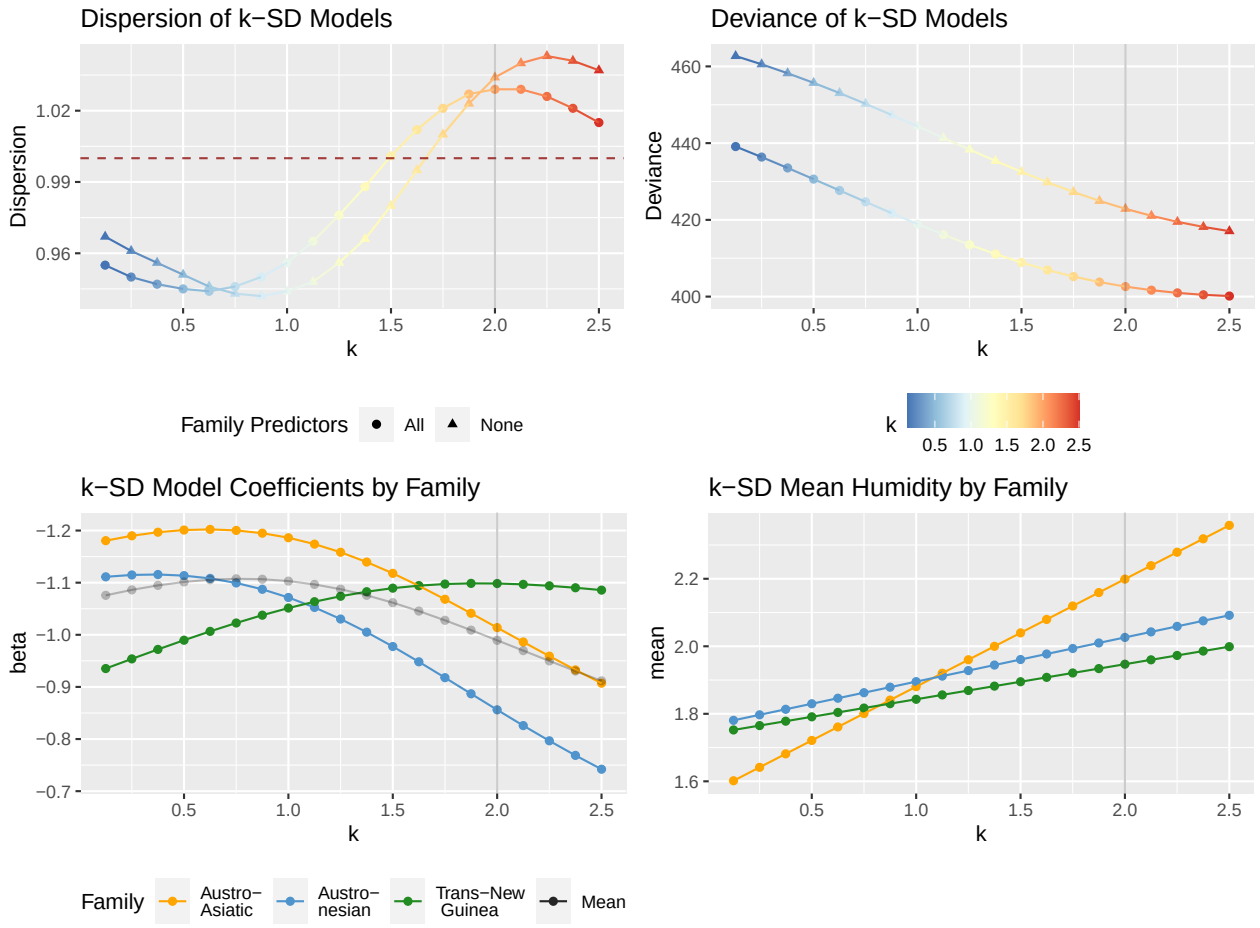
Below, we observe the models of `complex_tonal` based on values of the humidity k-SD score statistic $\widetilde{H_k}$ where $k \in (0,\ 2.5)$, spanned with 20 increments of $0.125\sigma_H$. To summarize, the results below are obtained from fitting binomial GLMs with the formula:

$$\mathbf{1}_{\text{Complex}} \sim (\mu_H + k\sigma_H)$$

Plots Description

Below, we find our plots summarizing our parameterized model results. On the top row, we observe our primary model performance heuristics: dispersion and residual deviance. On the x-axis, we can see how our "extremity" parameter k varies from 0 to 2.5, which respectively represents going from our baseline, $\widetilde{H_0} = \mu_H$, to the upper extreme $\widetilde{H_{2.5}} = (\mu_H + 2.5\sigma_H)$, which well overshoots the maximum humidity actually observed for many data points in our dataset. Speaking of which, this is why we include a reference line at $k = 2$, since roughly speaking, $\widetilde{H_2} = \mu_H + 2\sigma_H \approx \max_H$ best represents the maximum for most of our data points, representing a pivotal value to focus on in terms of our analysis. Here, we mean $k \approx 2$ minimizes $\mathbb{E}(|\widetilde{H_k} - \max_H|)$ over all our observations.

Additionally, the shape aesthetic is used to delineate the inclusion/exclusion of our family predictors. Also, a dashed red line is added to reference the theoretical binomial dispersion of $\phi = 1$. With regards to the bottom row, we observe two plots regarding the family indicator predictors. On the left, we observe the estimates of the coefficients, which are placed on an inverted axis to emphasize their magnitudes or effect sizes. In black, we have a faint line showing the average size of all these coefficients, giving a rough sense in which how much the family predictors weigh in the model overall. On the right, we simply have the predictor values used for each family, placed for reference.



Results Overview: Score Model

The first immediate takeaway from the score models is the suggestion that **upper quantiles of the data do indeed have increasingly stronger explanatory power** in explaining complex tonality! We can see that as k increases, there is a continuous and undeniable pattern showing a reduction in deviance. For reference, this model has a null deviance of $\chi^2_* = 475$. In the case of $k = 0$ (the equivalent of just using MH) we observe a baseline residual deviance of $\chi^2 = 465$. Then, the model deviance drops at a steady rate, eventually tapering off near $\chi^2 = 405$ around our critical point $k = 2$. Any marginal reduction in deviation afterwards is negligible, which is expected granted points extrapolating beyond $k = 2$ start to lose credibility as mean estimates start to exceed actual observed maxima. This is one weakness of this model, which actually is a motivation for the quantile statistic we discuss next. While it is not shown, extrapolating further beyond $k = 3$ would eventually show that the deviance starts to increase, since the score statistic starts to lose its basis and grip on the actual data, significantly bloating the predictor beyond reality and causing poorer fits. This also cements a trend we see later, which is that the right-tails (most extreme quantiles) tend to have worse information.

Although it is not shown, it is worth at least mentioning that all the coefficients are statistically significant, especially the humidity statistic, which we are most concerned about. The language predictors fluctuate more, but with few exceptions, generally remain below $p < 0.05$. All said and done, it is a remarkable result to see the score statistic, which is a simple linear combination of two static predictors (μ_H and σ_H) give us a simple parameterized model that shows an undeniable, *continuous* pattern of deviation reduction. As expected, it also tapers off near the $k = 2$ region, as nearly no more "real" information exists beyond $k \gg 2$ (taking real to mean grounded in the true distribution of the data overall). While this method has its issues, this does serve as a simple and quick method to firmly establish that upper quantile data *does* contain increasingly more information in explaining complex tonality. In the next section, we move on the second class of models, and afterwards ending with a holistic discussion on dispersion and family predictors.

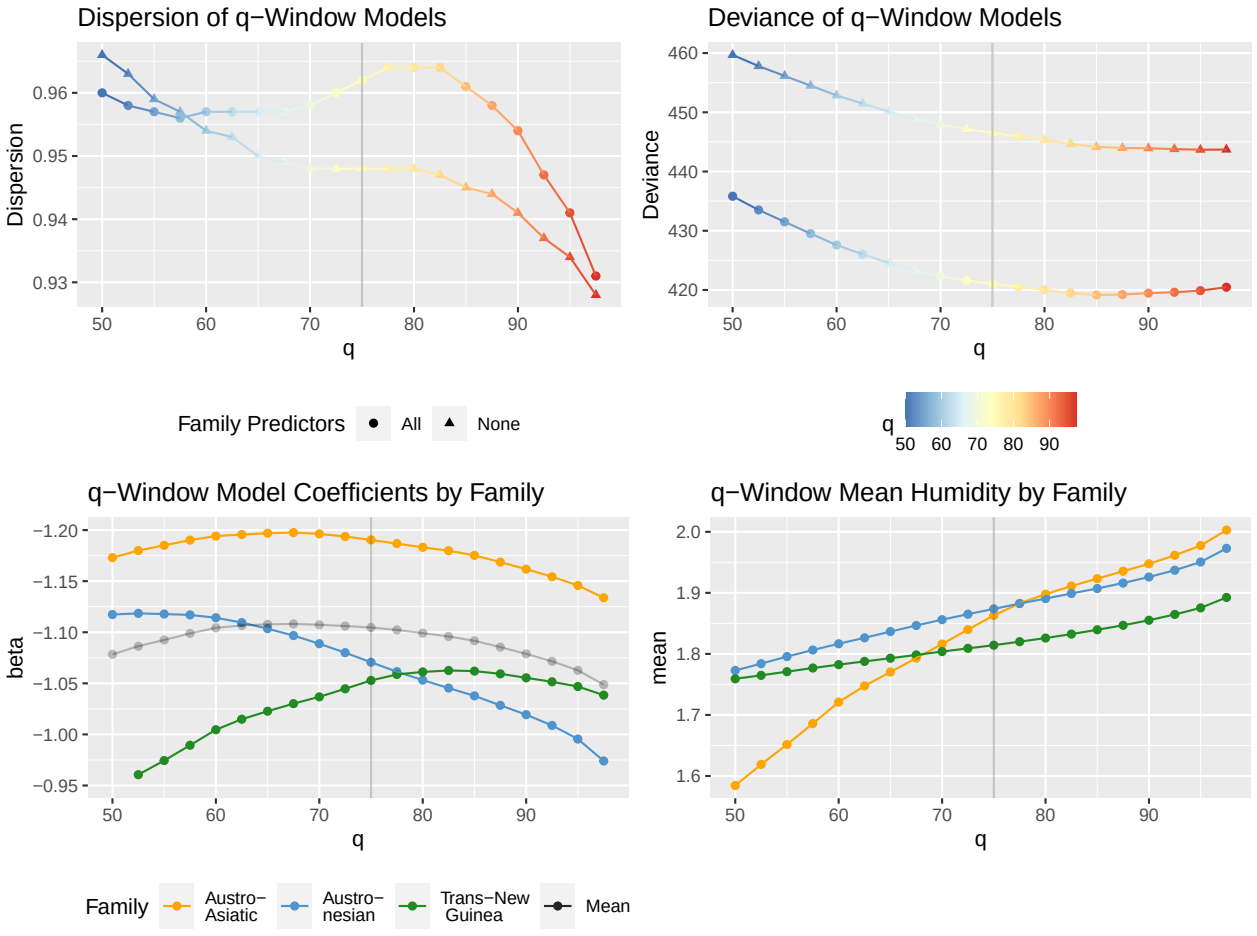
Window Quantile Model

Below, we observe the models of `complex_tonal` based on values of $\mu_q^{(\delta)}$ with $\delta = 2.5\%$ where $q \in (50, 97.5)$, spanned with 20 increments of δ. To summarize, the results below are obtained from fitting binomial GLMs with the formula:

$$\mathbf{1}_{\text{Complex}} \sim \mu_q^{(2.5\%)}$$

Plots Description

Below, we find our plots summarizing our parameterized model results, exactly in the same format as the previous set of models. The primary difference is in the axis, which has our "extremity" parameter q, representing the center quantile of our averaged range. A grey vertical bar is placed for reference for $\mu_{75}^{(2.5\%)}$ for no specific reason other than reference; however, it is worth noting that $q = 75$ is approximately the quantile parameter where our deviance reduction effect begins to taper off. Additionally, unlike the previous model, no parameter achieves a dispersion of $\phi = 1$, so the dashed red line is omitted.



Results Overview: Quantile Model

Akin to the first set of models, again, is the suggestion that upper quantiles of the data do indeed have **increasingly stronger explanatory power** in explaining complex tonality. We can see that as q increases, there is a similar continuous pattern showing a reduction in deviance, which like before, **wanes on the right-tails**. Just like before, this model has a null deviance of $\chi^2_* = 475$. In the case of our starting point $q = 50$, where we use $\mu_{50}^{(2.5\%)}$, we observe a slightly lower baseline residual deviance of $\chi^2 = 460$ compared to the score models.

Dropping at a steady rate, the model deviance drops and eventually tapers off near $\chi^2 = 420$ around our critical area near $q = 75$, showing diminishing marginal returns in deviance reduction afterwards. With regards to the limiting behavior, we can actually observe the deviance start to increase for very extreme quantiles that are close to the maximum. Although it was not displayed in the previous plots, this resembles the last model, where the deviance started to pick up again after $k \gg 2.5$. Additionally, the models appear to be increasingly underdispersed as q grows beyond $q = 80$, a critical phenomenon to be discussed next.

Dispersion & Family Predictors

Finally, we comprehensively close out our discussion on logistic GLMs. At this point, we definitively conclude from this section that **the usage of upper quantiles, far enough from the maximum quantiles, gives us increasingly more information in explaining complex tonality.** The primary engine for this inference is the observation of *continuous* deviance reduction as our parameters k and q grew. That being said, it is important to assess other factors, such as the residuals' behavior, how the fitted values are distributed, and accordingly, dispersion. As we dive more into more detail, we argue that the score models are actually not sound to base our inferences, in contrast to our quantile models which remain so. In the end, this serves a cautionary tale to look beyond summary statistics such as deviance, as models are always more complex than single numbers.

Speaking of numbers, to preempt our discussion, we briefly recount how dispersion is estimated in Logistic (Bernoulli) GLMs. Recall that the sample dispersion estimate is $\phi = \frac{1}{\nu}\sum r_P^2$ where r_P is the Pearson residual. For the Bernoulli distribution, the squared Pearson residual can be further simplified by case. Namely, $r_P^2 = \frac{\hat{p}}{1-\hat{p}}\,(y = 0)$ and $r_P^2 = \frac{1-\hat{p}}{\hat{p}}\,(y = 1)$. That being said, we contextualize everything by noting that almost all the time, our fitted values observed a range of $\hat{p} \in (0,\, 0.4)$. With that being said, we begin with the discussion of the score models.

As we see in the score model, the model appears to become overdispersed as k grows. We find that this is the result of smaller fit values for $(y = 1)$, the complex tonal observations. We briefly mentioned how $k \approx 2$ is a critical value since it approximates the maximum humidity overall. If we condition by group, a disparity appears, where $k \approx 2.3$ for $y = 1$ and $k \approx 1.8$ for $y = 0$. This disparity is partially explainable due to the negative skew of larger MH values (where the mean is to the left of the median). Nonetheless, this implies that our $y = 1$ predictors are disadvantaged, leading to smaller fits near 0.2, which greatly increase dispersion as $\hat{p} = 0.2$ would translate to $r_P^2 = 4$ here. As such, the benefits of better fit for $y = 0$ do not pan out. In contrast, something else happens in the quantile model, where the model becomes greatly underdispersed after $q = 80$. This is caused by the fact that there is very little variation in predictors near the maximum quantile, as all the values hit a "wall". As such, this induces worse "separability", as both groups start to have similar predictors, which causes the quantile model to be underdispersed and eventually perform worse. Also worth noting is that maximum quantiles are less stable than smaller ones, especially in positively skewed distributions. As such, this indicates a "golden region" of $q \in (70,\, 80)$, where a lot of information is contained. This also explains the high fluctuation, as skewed distributions start to overlap in complex and unique ways. Lastly, we conclude our discussion on the language predictors. // Addition of language predictor does disperse the model in the right direction. // Our language families of Austronesian, Austro-Asiatic, and Papuan are isolated languages, which is worth noting. Given that they are genealogically connected, they also form a cluster. However, we still need to provide evidence this is not just reverse engineering the fact that these families demonstrate non-complex tonal, humid regions. Some explanations include genealogy (clustering) since languages inherit tonal systems. Also, as mentioned, isolation from other languages could put them in a separate class. Lastly, linguistics phenomena such as population size could showcase confounding variables that have already been linked to linguistic features, such as complex morphology and large population size. // In the end, the inclusion of the predictors as well-specified remains an open question.

Null Survival Models

In this final section, we explore the usage of survival models in an unconventional way to model and verify the authors' hypothesis that complex tonality cannot exist in arid climates. Here, we use the term "survival model" in a loose and more semantic than the traditional mathematical sense, since we are modifying the time axis to our case, an "aridity axis". To come up with such axis, we construct a *modular* notion of dryness, which can correspond to any **fixed** humidity statistic discussed above, call it θ. As such, we can formally create a corresponding dryness/aridity index, to be called d_θ. Once the statistic is defined and unambiguous, we drop the θ. It is important to note that somewhat like survival models, there is *in practice* a bounded range in the axis of interest. In this model, this is always the case, since aridity is define through the arithmetic inverse of humidity, which itself is bounded.

Additionally, we do not use other predictors, which generally tends to be the case with survival models, so we call this a *null* survival model, as we are merely trying to model the baseline hazard rate $\lambda(d_\theta)$, which tells us the "risk" of losing tonality as aridity d increases. As such, we set our status/death variable not as the indicator variable `complex_tonal` that we have been regularly using, but rather, its complement `is_not_complex_tonal`. As such, a status of 1 in our survival model indicates a language that has "lost" its complex tonality. More formally, it has lost the ability to develop complex phonemic tone due to the lack of necessary humidity conditions. While the fussy semantics of this construction seem ancillary, it is important to be precise in the claims being made implicitly through our language, especially if we are diverging from convention. At any rate, we now arrive at our model formula:

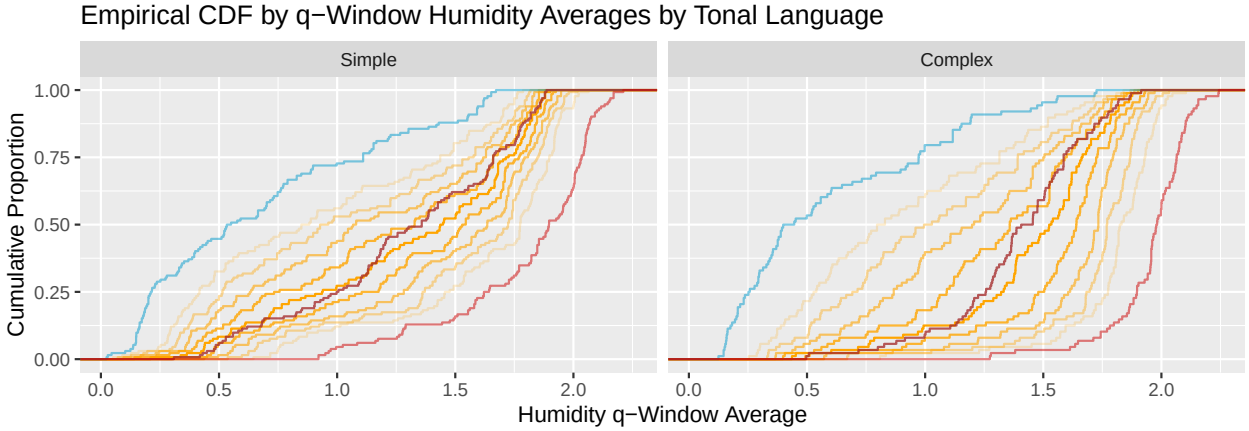
$$\lambda(d) \sim \mathbf{0}_{\text{Complex}}$$

To achieve fitting the goal of fitting our baseline hazard, we use the Kaplan–Meier curve. To avoid needlessly complex terminology, this essentially just reduces to the complement of the empirical CDF distribution function, since no truncation or censoring occurs in our case. As mentioned already, this is non-parametric, and only involves using the measure-theoretic indicator step function at the observation of a "death". Without further ado, we observe our CDFs using class of θ: the quantile window statistics.

Continuity: The Quantile Window Statistic

As shown in our logistic GLMs, the k score statistic is less stable and grounded in our data. In contrast, our quantile-based statistic is more stable, and less susceptible to issues such as distribution skew. All that being said, we now observe the CDFs for both types of our tonal languages, with curves corresponding to $\mu_q^{(10\%)}$ for $q \in [10, 90]$ in 10 increments of $\delta = 10\%$. In the plot below, we label the minimum humidity statistic in light blue, the maximum in red, MH in dark gold, and finally, our $\mu_q^{(\delta)}$ curves in gold.

Just like in our logistic GLM models with the deviances, we can see a continuous relationship, where the mean curve appears to be "twisted" in either direction, and approaches the minimum and maximum for lower and higher values of q, respectively. Most critically, we notice that as q increases and the curve tends towards the maximum, it attains a noticeably higher level of convexity, which in survival modeling, corresponds to more information.



Relative Aridity Indices

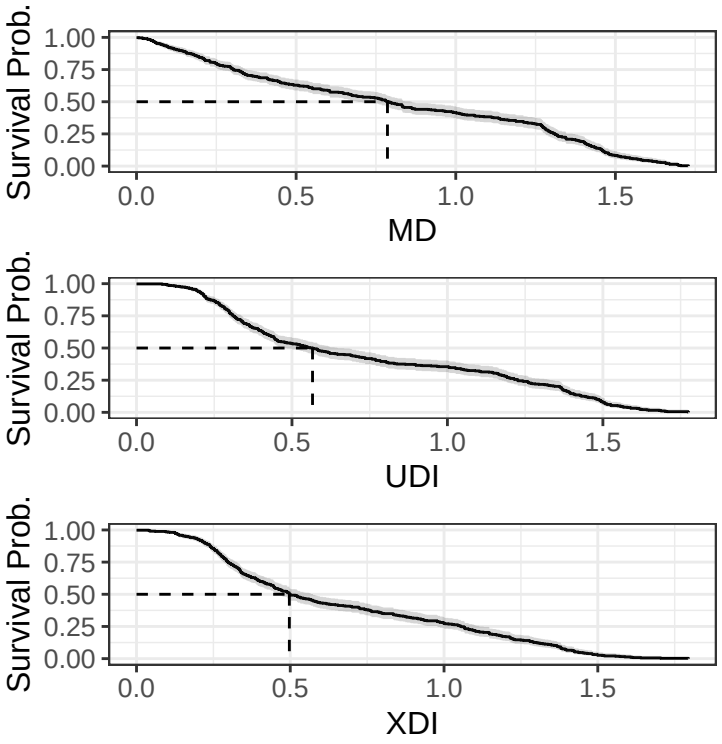
Now, at this point, we have identified the formula for our model's hazard rate, and made the case for the usage of the quantile window statistics. The last thing that remains is the construction of the aridity axis d. As hinted at previously, this will be a *modular* definition, or in other words, a function of some corresponding humidity statistic θ. Without further ado, fix some humidity statistic θ, and let $\theta^* = \max(\theta)$. Then, we define the θ-based *relative aridity index* d_θ to be:

$$d_\theta = (\theta^* - \theta)$$

In other words, this captures a notion of distance from its maximum potential. Since measuring dryness is not intuitive, we could call this θ-based humidity potential, where 0 indicates no remaining potential for increase, and the highest aridity, $\theta_{\min}$, attains the highest possible level of "potential." Note that the translation by the statistic sample maximum θ^* is necessary for one main reason: to have our aridity axis start at zero. However, any constant is plausible in theory, but in practice, the model lacks value if it lacks interpretability or utility, as such, we assume to have a large sample of humidity data at hand.

Finally, to finish up, we define three specific aridity indices, and compare their survival curves. Note that the simple tonality group has less information, as we saw in the CDF plot, so we stick to just the complex models. Let $\bar{d}$ denote the 50% survival quantile, such that $\bar{d} = S^{-1}(50\%)$.

- MD : $= d_{\text{MH}}$, where $\theta = \mu_H \Rightarrow \theta^* = 1.96$ and $\bar{d} = 0.75 \iff \theta_{\text{MH}} = 1.20$.
- UDI : $= d_{\text{UHI}}$, where $\theta = \mu_{90}^{(10)} \Rightarrow \theta^* = 2.21$ and $\bar{d} = 0.6 \iff \theta_{\text{UHI}} = 1.61$.
- XDI : $= d_{\max}$, where $\theta = \max(H) \Rightarrow \theta^* = 2.34$ and $\bar{d} = 0.5 \iff \theta_{\max} = 1.84$.



To sum up, we could see how the CDF convexity translates to more information, and earlier observances of the 50% quantile. This is great, because it accounts for non-complex languages even with MH values that are high, alluding to the issue of separability we couldn't avoid with logistic regression GLMs. Nonetheless, they do both account for these observations in similar ways, but it is different since logistic regressions are more "aggregate" in nature of deriving values/estimates, and this is more direct/sequential. Also, to be realistic, no new information is created, as it is worth highlighting that the model has much less convexity/information in larger values ($\theta > 1$), which we expect, since most tonal languages have been accounted for already. Nonetheless, this is a great visual demonstration of the authors' hypothesis, showing increased aridity leads to less complex tonal languages.

Conclusion

In conclusion, we have explored various models and techniques to further explore and support the hypothesis laid out by Everett and Roberts. We used the humidity score and the quantile window average. While they both had their differences, they both showed one there is comparatively more information to explain complex tonality in upper quantiles of the humidity distribution, as long as they are far enough from the maximum. We could understand the use of upper-quantile statistics as an attempt to better separate the data as best as possible, but nonetheless, perfect separation is impossible, and not expected. Furthermore, our discussion about family predictors remained inconclusive, despite their seemingly positive contribution to model performance, as we could not tell if they are merely inflating humidity statistic coefficients.

After discussing some issues regarding the score statistic regarding its grounding on the data and inaccuracy due to its projective nature, we decide to choose the window quantile statistic as the more inferentially sound statistic. We carry this over in our null survival model. Through our survival model, we were able to construct a model that closely demonstrates the authors' hypothesis semantically and mathematically. In contrast to the authors emphasis on using lower quantiles to show an explicit baseline, we leverage the power of statistics to show that the $q \in (70, 80)$ percentile range contains a rich slew of information that pans out in a complex manner to give us demonstrably higher explanatory power. Since humidity is a skewed distribution with deterministic structure, this is especially amplified.

Through the power of statistics, we have demonstrated the following idea – given only a select range of quantiles, we were able to extract information that relies on lower ones. In other words, by knowing the upper quantile statistics, there is a sense that we could encode some baseline minimum of the humidity distributions needed to obtain complex tonality. It does not need to be direct, especially since the "curse" of skewness could be a blessing in this case. We summarize the contrast with the authors by saying that they seek statistical evidence of minimal humidity through sufficiency based on the "comfort of a reliable, observed floor" (lower-bound paradigm), where we instead employ the implicit use of the maximum to mathematically model the notion that we may have "confident aspirations of a minimum sufficient humidity, having previously seen how high the ceiling was" (upper-bound paradigm). When we understand that humidity, as a geographic phenomenon is chaotic and random, but still in a sense has predictable, deterministic structure with finite uncertainty. As such, only through the power of statistics, we uncover the secret beautiful gems behind these deterministic systems, embedding themselves in the most unexpected ways possible, like in our case, complex tonal languages, and the geographic humidity structure!

References

Dryer, M., & Haspelmath, M. (2011). The World Atlas of Language Structures Online. Retrieved from http://wals.info/

Everett, C., Moran, S., & Roberts, S. (2015). Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. Proceedings of the National Academy of Sciences of the United States of America, 112(5), 1322-1327. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4321236/#d35e393

Hombert, J., Ohala, J., & Ewan, W. (1979). Phonetic explanations for the development of tones. Language, 55(1), 37-58.

Kalnay, E., et al. (1996). The NCEP/NCAR 40-Year Reanalysis Project. Bulletin of the American Meteorological Society. Retrieved from http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP-NCAR/.CDAS-1/.MONTHLY/.Diagnostic/.above_ground/

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. PLoS ONE, 5(1), e8559.

Maddieson, I. (2011). Tone. In M. Dryer & M. Haspelmath (Eds.), The World Atlas of Language Structures Online (Feature 13A). Max Planck Digital Library, Munich. Retrieved from http://wals.info/feature/13A