Proc Ling Soc Amer Volume 1, Article 29:1–15
doi:http://dx.doi.org/10.3765/plsa.v1i0.3734

Language Variation Suite: A theoretical and methodological contribution for linguistic data analysis

Olga Scrivner and Manuel Díaz-Campos*

Abstract. In recent years there has been growing interest in quantitative methods for analyzing linguistic data. Advanced multifactorial statistical analyses, such as inferential trees and mixed-effects logistic regression models, have become more accessible for linguistic research as a result of the availability of an open source programming environment provided by the statistical software R. In the present paper, we introduce a novel toolkit, Language Variation Suite, a software program that offers a friendly environment for conducting quantitative analyses. We demonstrate how theory built on traditional monofactorial analysis can be extended to macro and micro multifactorial approaches allowing for a deeper understanding of language variation. The focus of the analysis is based on intervocalic /d/ deletion in Spanish from the Diachronic Study of the Speech of Caracas 1987 and 2004-2010. In contrast to traditional methodological approaches we have treated intervocalic /d/ as a continuous dependent variable according to the intensity ratio measurements obtained. Furthermore, we have integrated various syntactic, phonetic and sociolinguistic factors. Non-parametric and fixed-effects regression models revealed that overall age (younger speakers), sex (male speakers), phonetic context (low vowels), token frequency and morphosyntactic category (past participles) have a significant effect on the lenition of intervocalic /d/. In contrast, the mixed-effects model selected only phonetic context, frequency and category, showing that individual speaker variation is higher than group variation.

Keywords. sociolinguistics; phonetics; variation; Spanish; statistics

1. Introduction. Nearly a century ago, Edward Sapir noted that "language is variable" (Sapir 1921:147). Subsequent research in sociolinguistics has shown that variation is "a universal and functional design feature of language" (Foulkes 2006). Moreover, this variation is structural, systematic, and predictable (Labov 1969). In this view, linguistic variation has been conceptualized as a categorical phenomenon. As a result, sociophonetic variation has been traditionally analyzed as the alternation between two discrete auditory categories, e.g. deletion versus retention. However, it has been recently pointed out that acoustic analysis "reveals important variation that is difficult to detect or analyze auditorily" (Thomas 2013:114). That is, the quality of discrete auditory analysis is affected by the researcher's experience, ear and perception, which are based on his or her native language. As Figueroa (2014) has rightfully noted, "There is hardly anything to hear, but we are hearing it nonetheless". In addition, auditory analysis cannot reveal some important acoustic variations, e.g. vowel quality and consonant intensity.

Similarly, traditional sociolinguistic tools, namely VARBRUL and GoldVarb X (Sankoff et al. 2005), are based on a categorical conception of independent variables (Johnson 2009, Tagliamonte 2011, Díaz-Campos and Dickinson 2017). Furthermore, most of traditional tests (e.g. fixed logistic regression in GoldVarb, T-test, ANOVA and Chi-square) assume independence of observation. With recent advances in statistical programming, the shortcomings of such

* Authors: Olga Scrivner, Indiana University (obscrivn@indiana.edu) and Manuel Díaz-Campos, Indiana University (mdiazcam@indiana.edu).

techniques applied to sociolinguistics have become obvious. First, the nature of sociolinguistic observation deviates from the assumption of normally distributed, balanced and independent data. As Díaz-Campos and Dickinson state, "sociolinguistic studies are based on correlated data". Furthermore, traditional sociolinguistic tools[1] are unable to handle continuous or multinomial variables. Finally, traditional tests do not capture individual-level or word-level variation (Johnson 2016). In fact, these statistical practices have already been abandoned in scientific fields in favor of new statistical methods, such as mixed regression models, conditional trees, and random tree analysis (Kidhardt 2015). Not only can these models measure individual and lexical variability, but they can also handle skewed and small-size corpora, which is often the case with linguistic data. While new methods have recently gained a lot of attention in sociolinguistic literature, their use remains limited, as they require some programming skills, which often presents technological challenges for researchers. There is also a need to test new tools "for their comparability and reliability in the study of language, variation and change" (Díaz-Campos and Dickinson 2017).

In this paper, we propose to address these issues by introducing a user-friendly application—Language Variation Suite—that implements state-of-the-art statistical methods. In addition to mixed effects models and regression tree analysis, this toolkit allows researchers to incorporate continuous and multinomial variables. Furthermore, we examine the weakening of intervocalic /d/ in Spanish, a gradable phonological phenomenon that has been traditionally treated in sociolinguistic studies as a categorical variable. We show that the conceptualization of sociophonological variables as continuous provides greater precision and better understanding of sound change, as it incorporates accurate acoustical criteria that take into account the gradient nature of phonological variables.

The remainder of this paper is organized as follows: Section 2 reviews sociophonetic variables and discusses traditional and current practices of sociolinguistic data analysis. Section 3 describes our corpus and methodology. Section 4 introduces a novel toolkit for sociolinguistic data analysis, Language Variation Suite. In section 5 we present results and discussion. Section 6 draws conclusions and provides future directions for our research.

2. Sociophonetic variable.

2.1. INTERVOCALIC /D/. Lenition of intervocalic /d/ is one of the most studied phenomena in the dialectological and sociolinguistic literature dedicated to Spanish. The realization of intervocalic /d/ as approximant [ð], e.g. *lado* [laðo] 'side', is a systematic articulatory reductive process (Navarro-Tomás 1999 [1918]). On the other hand, deletion of intervocalic /d/ is one of the most extreme manifestations of reduction, as in *cantado ~ cantao* 'sung' (Hualde et al. 2011). In fact, cases of /d/ deletion have been documented since the 17th century (Zamora 1970; Lapesa 1981) and have been found abundantly in many varieties of Spanish (Spain: Navarro Tomás 1999 [1918]; Latin America: Henriquez Ureña 1921; Venezuela: Lipski 1994). In dialectological and sociolinguistic studies, this phenomenon has been traditionally conceptualized as a discrete binary phenomenon based on auditory analysis, namely the presence and absence of /d/. Subsequent quantitative studies have shown that the realization of intervocalic /d/ is influenced by linguistic and extra-linguistic factors. That is, the choice between d-deletion and d-retention is systematic and sociolinguistically predictable. For example, Cedergren (1973) found that in the Spanish of Panama d-deletion is favored in informal styles by women, older speakers and lower socioeconomic participants from rural areas, while Padilla (1996) showed that in Las Palmas de

[1] Rbrul, a relatively new sociolinguistic toolkit, allows for continuous variables (Johnson 2009).

Gran Canaria d-deletion occurs mostly in the past participle with *-ado* and is favored by male speakers. Similarly, D'Introno and Sosa (1986) found that male speakers favor deletion in the Venezuelan variety of Spanish, whereas d-retention is favored by high and middle socio-economic groups. Furthermore, Díaz-Campos and Gradoville (2011) revealed the frequency effect on d-deletion in the same variety. Their results show that high lexical frequency and type frequency predict higher deletion rates.

In contrast to the traditional binary approach, acoustic studies have shown considerable variation in the realization of intervocalic /d/. In this view, the degree of /d/ lenition can vary from very close (consonant-like) to very open (vowel-like) realizations of the approximant [ð] (Carrasco 2008, Hualde et al. 2011). This degree is commonly measured by means of *relative intensity*: i) intensity difference—the difference between the lowest intensity point of the approximant and the highest intensity point of the following vowel (Eddington 2011, Simonet et al. 2012) or ii) intensity ratio—the ratio between the lowest intensity point of the approximant and the highest intensity point of the following vowel (Carasco et al. 2012).[2] As a result, lenition is conceptualized as a continuous variable. According to recent acoustic studies, more lenited or vowel-like realizations of /d/ occur in the following contexts: i) before a stressed vowel (Colantoni and Marinescu 2010), ii) in word-medial position (Eddingon 2011), iii) with higher frequency words (Eddington 2011) This continuous scale for intervocalic /d/ allows for greater precision, as it is based on more accurate acoustic measurements.

It should be noted that the previous accounts of /d/ lenition have the following limitations: i) many acoustic studies rely on a monofactorial analysis of the relation between the realization of intervocalic /d/ and one predictor, e.g. duration or vowel context, prosodic context, stress (Simonet et al. 2012, Limanni 2009, Torreira and Ernestus 2011, etc), ii) most studies examine apparent time variation, namely the comparison between speakers of different age groups during the same chronological time period and iii) most sociolinguistic multifactorial studies are based on auditory discrete analysis, which is affected by researchers' perception.

2.2. TRADITIONAL AND NOVEL PRACTICES IN SOCIOLINGUISTICS. The foundation of traditional variable rule practices in the sociolinguistic field was introduced in Labov's classic study on copula deletion (1969). Labov observed that language variation is *inherent* and *systemati,c* in contrast to previous views on language variation that treated it as *optional* or *free* (Cedergren 1974:333). In this approach, language variation, denoted as a linguistic variable, represents "two or more ways of saying the same thing" (Labov 1972:271). Furthermore, each context is independent from other contexts and has a fixed effect, which is based on the presence or absence of a given feature (Cedergren and Sankoff 1974:335). This *Variable Rule* model enables researchers to incorporate the combination of sociolinguistic and linguistic environments in which the linguistic variable occurs (Labov 1969). This model was implemented in the first sociolinguistic statistical tool *VARBRUL*, which was replaced by an improved version, *GoldVarb* (Sankoff et al. 2005). For several decades, the variable rule program has been successfully employed in many sociolinguistic studies, allowing researchers to identify which sociolinguistic factors influence phonological variation. While this program helps analyze the multifactorial interplay of social and linguistic factors, the categories of the *Variable Rule* model must be discrete, and factor groups with 100% or 0% must be excluded. As Díaz-Campos and Dickinson (2017) point out, this design was a product of "linguistic theories at the time where linguistic features were

[2] For additional methods that measure the degree of lenition, such as spectral tilt, velocity curve and EPG, see Carasco et al. (2012) and Hualde et al. (2011).

conceived as [+/-]". Furthermore, the underlying assumption of logistic regression implemented in *GoldVarb* is independence of observation. Recently, it has been argued that linguistic variables are rarely independent and that "many potential predictors are in a nesting relationship with speaker or word" (Johnson 2016). Thus, to improve traditional variable rule analysis and allow for non-discrete continuous predictors, the mixed-effects model has been introduced into the sociolinguistic field. It has been shown that this new model "returns more accurate *p*-values compared to a fixed-effects model that ignores nesting" (Gorman and Johnson 2013:223). This model is available in many types of statistical software, such as PROC GENMOD in SAS, the *glm* package in R and Stata (Agresti 2007:67), and has also been implemented in a new sociolinguistic toolkit, *Rbrul* (Johnson 2009). Finally, there has also been growing interest in using visual statistical methods such as random forests and conditional inference trees to enhance the *Variable Rule* model and improve its limitations (Tagliamonte and Baayen 2012). Random forest and tree-based methods are referred to as non-parametric regression tests. Conditional inference trees (*partykit* package) estimate the distribution of a response (aka a dependent variable) by means of recursive partitioning (Hothorn and Zeileis 2015). In this approach, "the feature space is recursively split into regions containing observations with similar response values" (Strobl et al. 2009:324). This tree-based analysis has been successfully used for multivariate data exploration in many scientific fields. While such a non-parametric approach is relatively new in sociolinguistics, recent studies have shown that random forests "provide the closest fits to the data" (Tagliamonte and Baayen 2012:32) and that conditional trees help "visualize different combinations of factors (independent variables or fixed effects) and their significance" (Díaz-Campos and Dickinson 2017:4). These advanced practices enable researchers to handle imbalanced data, measure individual variation and rank variables according to their significance (Strobl et al. 2009); their implementation, however, requires some programming skills (e.g. R programming language) or access to statistical tools that are not always freely available. In addition, given the vast number of available statistical tests, the question has been raised as to how these current practices affect sociolinguistic studies and concerning their advantages and disadvantages for studies of language variation. In answering these questions, it is necessary to compare and contrast both approaches, traditional variable rule model and innovative models (Johnson 2009, 2016; Eddington 2010; Tagliamonte 2011, 2012; Díaz-Campos and Dickinson 2016).

3. Methodology.

3.1. CORPUS. The data used in this project comes from a diachronic corpus, *Corpus histórico del habla caraqueña 1987 y 2004–2010 (CHHC'87/04–10)* `Diachronic Study of the Speech of Caracas 1987 and 2004-2010' (Bentivoglio and Sedano 1993, Bentivoglio and Malaver 2006). This corpus consists of one hundred sixty half-hour sociolinguistic interviews with audio recordings and transcripts, conducted with native speakers of Caracas. The current research focuses on a subset of thirty-two speakers who are equally divided among three age groups (20–34, 35–54, 55 and older), both genders and three socioeconomic groups (upper, middle, lower). For this study, we included only word-internal instances of intervocalic /d/ (e.g. *ocupado* 'busy', *vida* 'life'). In addition, we included cases where the preceding or following vowel was a diphthong (e.g. *cambiado* 'changed', *fastidiar* 'to annoy'). The total of 1031 tokens containing intervocalic /d/ was collected from this corpus.

3.2. ACOUSTIC ANALYSIS. Acoustic analysis was performed using PRAAT (Boersma 2001). We manually segmented sections of sound waves corresponding to intervocalic /d/ and its preceding and following vowels. The acoustic measurements for intervocalic /d/ were obtained by using the *relative intensity ratio* method described in Carrasco et al. (2012). This method requires two measurements from the intensity curve: the lowest intensity point of /d/ and the highest intensity point of a vowel. The intensity ratio is calculated by dividing the lowest intensity point of /d/ by the vowel's highest intensity point. PRAAT scripts are developed to extract the highest and the lowest points as well as to calculate intensity ratio formulas.[3] A sample script is illustrated in Figure 1.

```
label$ = Get label of point... tier point

if label$ <> ""
    if label$=="V"
        n=n+2
            k=k+2
        # calculates the onset and offset First interval
           onset1 = Get starting point... tierVowels k
           offset1 = Get end point... tierVowels k
        labelWord$ = Get label of interval... tierWord n
        vpoint = Get time of point... tier point
        select Intensity 'soundname$'
        v1 = Get value at time... vpoint Cubic
        # get maximum intensity and time
        max_int1 = Get maximum... onset1 offset1 Parabolic
        max_time1 = Get time of maximum... onset1 offset1 Parabolic
        diff=v1/max_int1
        resultline$ = "'labelWord$"'tab$"'max_int1"'tab$"'v1"'tab$"'diff1"'newline$'
        fileappend "'resultfile$"' 'resultline$'
        k=k+1
    endif
```

Figure 1. Sample of PRAAT script for automatic acoustic measurement

The obtained ratio provides a value between 1, a more vowel-like production, and 0, a more stop-like production. For example, Figure 2 demonstrates an instance of a more lenited /d/ in *comunicado* 'informed' (ratio = 0.98), and Figure 3 exhibits a less lenited /d/ in *regadera* 'shower' (ratio = 0.89).[4]

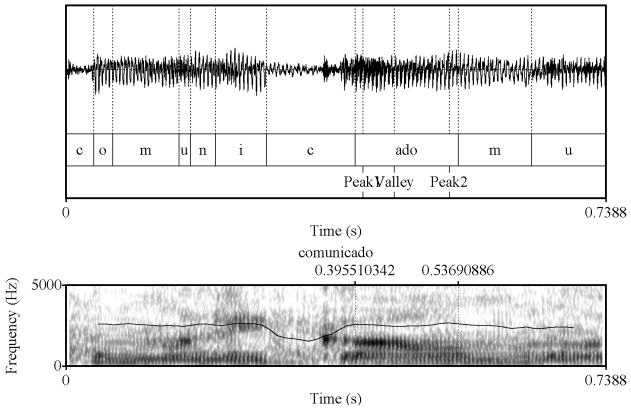


Figure 2. Sound wave, spectrogram and intensity contour of a more lenited intervocalic /d/ in *comunicado* 'informed'

[3] In contrast to the previous relative intensity methods, we measured the intensity of a preceding vowel.

[4] *Valley* designates the lowest point of intervocalic /d/, and *Peak* indicates the highest point of a vowel.

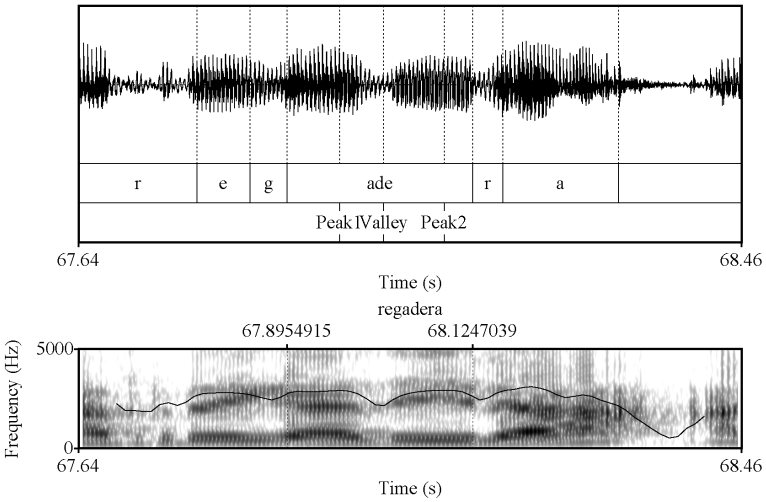


Figure 3. Sound wave, spectrogram and intensity contour of a less lenited intervocalic /d/ in *regadera* 'shower'

3.3. SOCIOLINGUISTIC VARIABLES. For the purpose of our investigation, the intensity ratio serves as the dependent continuous variable. The summary of our independent variables is illustrated in Table 1:

Factors	Values	Type
Preceding vowel	low, mid, high	categorical
Following vowel	low, mid, high	categorical
Stress	stressed, unstressed	binary
Grammatical Category	adjective, participle, noun, verb, adverb, pronoun	categorical
Token Frequency[5]	log	continuous
Lexical Frequency[6]	log	continuous
Age	20–34, 35–54, 55+	categorical
Gender	male, female	categorical
Socio-economic level	high, low	categorical
Period	1987 and 2004/2010	categorical

Table 1: Summary of dependent and independent variables

Individual participants and word tokens are also included as variables for mixed-effect regression analysis to measure variability between speakers and word-specific effects. The codified data is stored in CSV format (comma separated values), which makes it easy to manage and analyze data. In the next section, we will describe our new statistical tool for data analysis.

[5] AntConC was used to calculate token and lexical frequencies (Anthony 2010).

[6] The frequency of lexema (root) in the corpus, e.g. cansado, cansada (`tired' m, f).

4. Language Variation Suite. Previous sociolinguistic tools, such as GoldVarb and Rbrul, were designed to run on personal computers. As a result, they require installation and computer memory usage. That is, a particularly large dataset may need to run for several hours to perform analysis, depending on the user's hardware. Furthermore, not all tests are available in such applications. For instance, while Rbrul carries out a mixed-effect regression analysis, it does not include conditional tree and random forest analyses. Recently, a new programming environment, R, has received attention in the sociolinguistic literature. As Tagliamonte (2011) points, "R is exponentially more powerful tool for statistical analysis than Goldvarb or Rbrul" (2011:168). R has already been used in psycholinguistics, and it has started gaining popularity for the analysis of linguistic data (Jenset 2010). However, it involves a steep learning curve and has no user-friendly interface (Tagliamonte 2011).

We propose a new tool, *Language Variation Suite*, created with the powerful statistical R package and designed with a user-friendly interface (see Figure 4). In addition, our program runs online and does not require installation or memory usage. Furthermore, this application carries out state-of-the-art statistical tests, e.g. conferential trees, cluster analysis and random forest, as well as graphical data visualization.[7] As a result, *Language Variation Suite* makes advanced statistical methods accessible to a broader audience, as its use does not require programming skills.

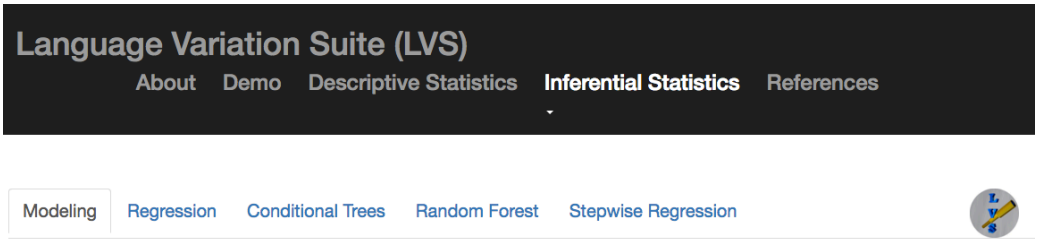


Figure 4. Language Variation Suite: On-line interface

Various statistical *R* packages are used in this program, e.g. *mlogit*, *lme4*, *randomForest*, *wordcloud*, *ca*, *stats*. The architecture of this tool consists of two components: i) a server script and ii) a user-interface definition. The server script includes codes for various functions and expressions, e.g. *renderPlot* or *renderTable*. The user-interface definition controls the html output of these functions and defines which functions require user input (interaction) and which functions return output. To illustrate this program, we provide samples of an *R* script and its output on the interface. Figure 5 demonstrates a function for selecting a statistical model. The user has to select the type of model, fixed or mixed, and the type of dependent variable, binary or continuous. On the left, we present a code for this function, and on the right, there is an actual html output on the interface.

[7] Currently, the application (v.1.) is hosted at https://languagevariationsuite.shinyapps.io/Pages.

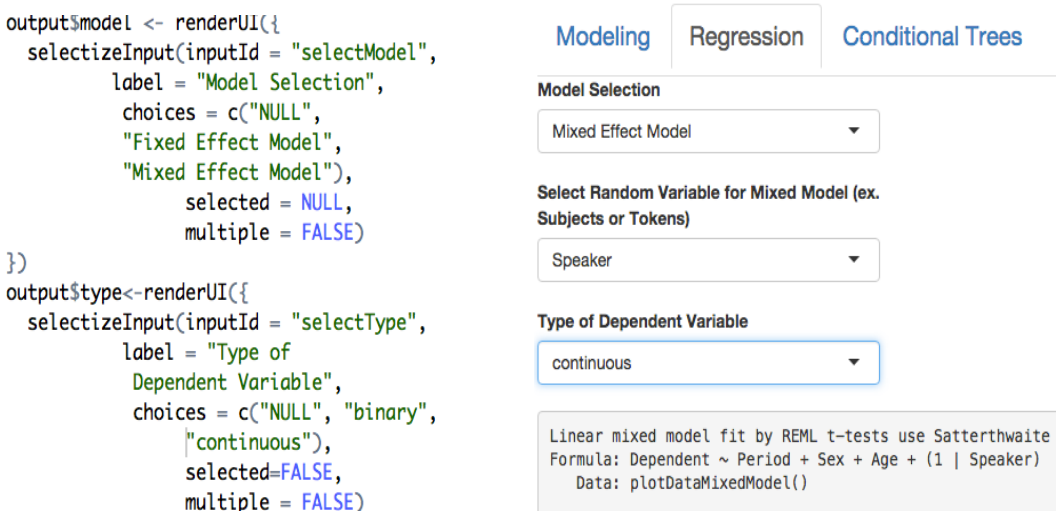


Figure 5. A sample script written in Rstudio (left) for a statistical model's selection and its output as a ShinyApp (right).

5. Results.

5.1. DESCRIPTIVE STATISTICS. The overall distribution of intervocalic /d/ is illustrated in the density plot (see Figure 6). This plot shows a unimodal distribution, with its peak at 0.956.[8] These results suggest that deletion is not the norm and that lenited variants are common in this speech community.

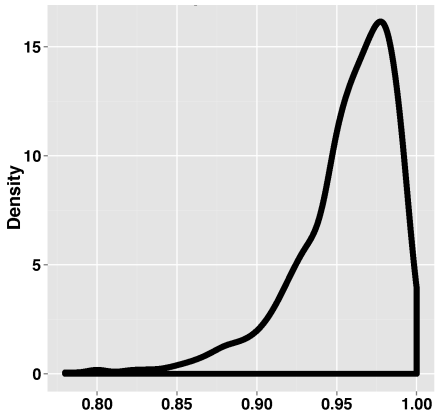


Figure 6. Kernel density plot for intervocalic /d/ distribution (intensity ratio)

Looking at two chronological datasets separately, it is noticeable that there is a sharp peak in the 1987 dataset (see Figure 7), whereas the curve becomes more evenly distributed around its peak in the 2004/2010 dataset, as shown in Figure 8.

[8] Since intervocalic /d/ tends to be produced as an approximant in intervocalic position, a zero value is not expected. The lowest ratio was 0.75.

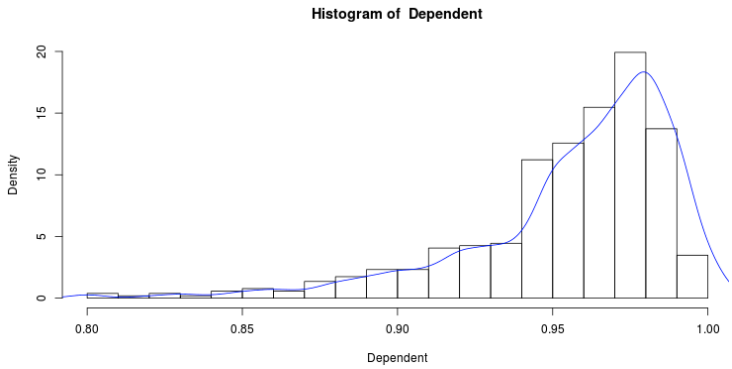


Figure 7. Intensity ratio for the 1987 dataset

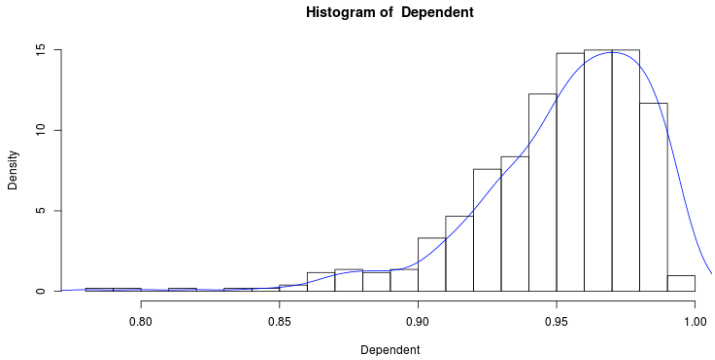
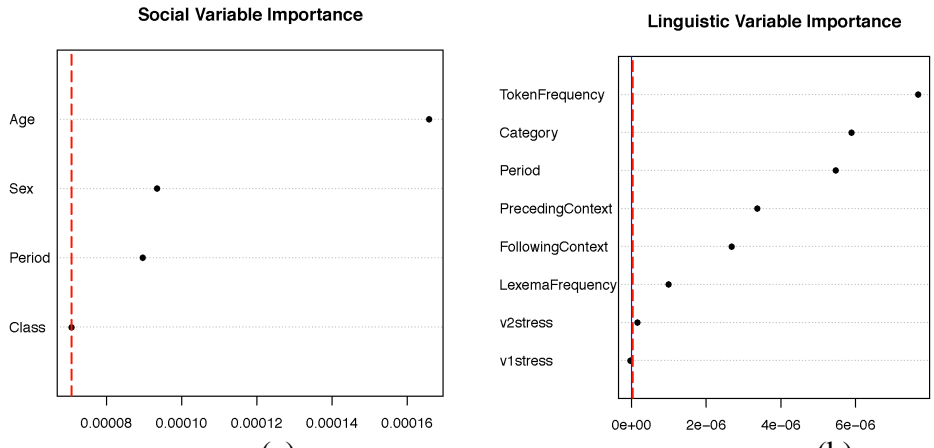


Figure 8. Intensity Ratio for the 2004/2010 dataset

5.2. INFERENTIAL STATISTICS. In this section we turn to the advanced statistical methods available in *Language Variation Suite*, namely random forest, conditional trees and mixed-effects regression. The random forest model determines the relative importance of independent factors with respect to a dependant variable. Figure 9a depicts social factors, and Figure 9b shows linguistic factors for intervocalic /d/ lenition, where independent factors are plotted according to their importance. All factors placed to the right of the dashed vertical line are considered significant.[9]



(a) (b)

Figure 9. Variable importance for intervocalic /d/ lenition

[9] For more information about random forest analysis, see Tagliamonte and Baayen (2012).

According to these results, the most important predictor among social factors is Age, and Token Frequency is by far one of the most important predictors among linguistic factors. Sex, Period, Category, Preceding and Following Contexts also contribute significant effects in predicting intervocalic /d/ lenition. As Tagliamonte and Baayen (2012) point out, random forest allows for collinear variables (highly correlated factors) to be considered jointly. For example, our model includes the following variables: category, phonetic contexts and frequency. While not falling into the same type, these factors are nonetheless highly correlated. It is well known that *-ado* is a preferred context for /d/ deletion: *-ado* is a frequent past participial suffix and at the same time it is a common phonetic context for /d/ deletion. Based on the model ranking, the order of strength is frequency >category >preceding context >following context.

The second non-parametric method, namely conditional tree, is a single representation of recursive partitioning. While it is inferior to random forest ranking,[10] the single tree makes it possible to visualize the partitioning of a dependent variable by independent factors. Following the methodology of Tagliamonte (2012:153) and to avoid complex trees, we will look at social and linguistic factors separately. Social factors are shown in Figure 10, and linguistic factors are illustrated in Figure 11. It should be noted that factor groups are represented in a hierarchical order from top to bottom. In this model, the node numbers simply show the sequential labels from left to right, terminal nodes represent relative frequency of response, and *p*-values indicate the level of factor significance (Strobl et al. 2009).

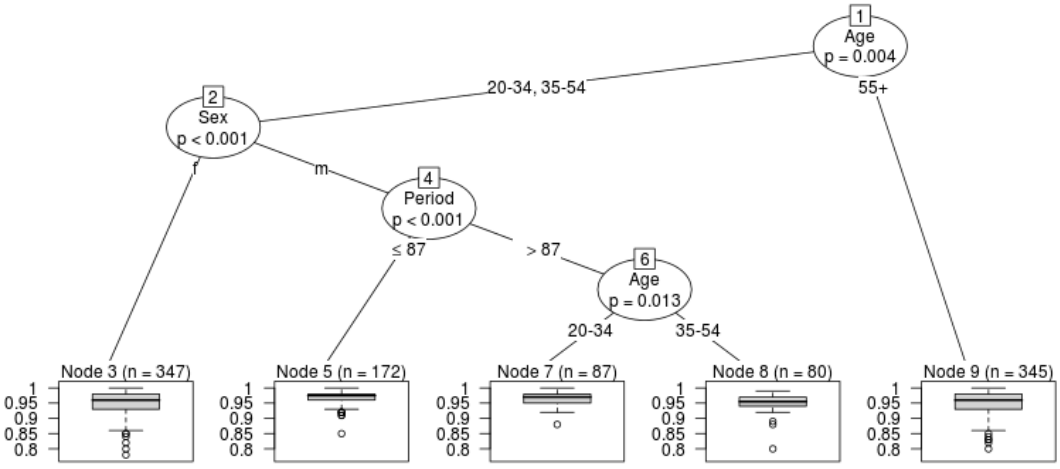


Figure 10. Conditional inference tree with social factors

[10] The splitting criterion is sensitive to small corpus size and outliers.

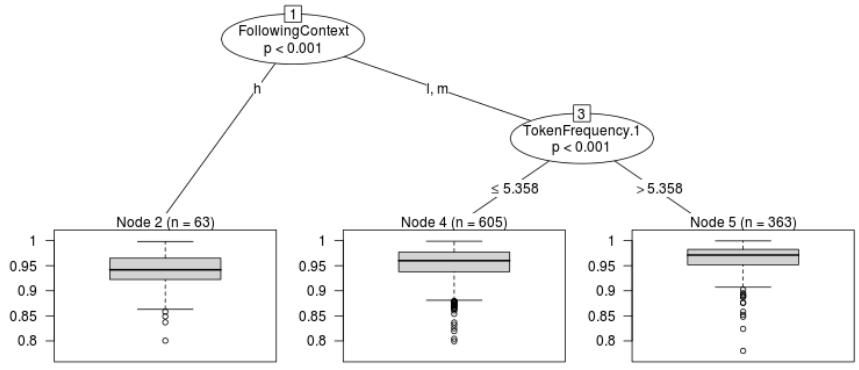


Figure 11. Conditional inference tree with linguistic factors

According to our tree model, **Age** is the most important social factor, splitting speakers into two groups: 20–54 and 55+. Recall that our dependent variable is continuous and therefore, terminal nodes are represented by box-plots with mean value (dark solid line). The 20–54 old group is further differentiated by sex with male speakers, especially in 1987, using more lenited variants of intervocalic /d/. In 2004/2010 only younger male speakers (20–34) produce more lenited /d/. Socio-economic class is not selected as significant, which confirms the results from random forest analysis (see Figure 9a). Among our linguistic factors, only preceding context and token frequency are selected as significant. Preceding context is split between high vowels and low/mid vowels. While random forest identifies frequency as the most important predictor (see Figure 9b), the conditional tree suggests that frequency is the most important predictor for low/mid vowels. In addition, we see that more frequent tokens exhibit more lenited variants, which supports previous accounts on /d/ deletion (see Díaz-Campos and Gradoville 2011).

Finally, we will perform a parametric analysis, where we will compare fixed-effects and mixed-effects models. It should be noted that each model has its own advantages and disadvantages. As Johnson (2016) states, fixed-effects models ignore individual variation, which may lead to Type I Errors, where "a chance effect is mistaken for a real difference between the populations". In contrast, mixed-effects models are prone to Type II Errors: "if speaker variation is at a high level, we cannot discern small population effects without a large number of speakers" (Johnson 2016:22-23). In addition, we need to select the best model for each regression analysis. *Language Variation Suite* performs model comparison by using AIC (*Aikake Information Criterion*), BIC (*Bayesian Information Criterion*) and *Anova* with Likelihood Ratio Test.[11] Table 2 illustrates the results for the best fixed-effects model ($p<2.515e-11$) based on Anova and AIC criteria. This model includes the following independent factors: preceding and following contexts, token frequency, sex, age, period, morpho-syntactic category. According to the results, following phonetic context and token frequency exert a highly significant effect on lenited intervocalic /d/. As their coefficient estimates are positive (0.0197 and 0.0012, respectively), low vowels and frequent tokens favor more lenited variants. Other significant factors by order of significance are age group of 20–34 ($p<0.01$), following mid vowel ($p<0.01$), male speakers ($p<0.01$), past participle ($p<0.01$) and preceding low vowel ($p<0.05$).[12]

[11] See Appendix E for more information on AIC and BIC – online: http://onlinelibrary.wiley.com/store/10.1002/9781118856406.app5/asset/app5.pdf

[12] Period and age group of 35–54 are only very marginally significant (0.0498 and 0.0436).

	Estimate	Std Error	t-value	p-value
Intercept	1.3946250	0.2333471	5.977	3.15e-09
Following context=low	0.0196532	0.0044887	4.378	1.32e-05
Token frequency	0.0012424	0.0003757	3.307	0.000976
Age=20-35	0.0078232	0.0023894	3.274	0.001096
Following context=mid	0.0141762	0.0044632	3.176	0.001537
Sex=Male	0.0061610	0.0019732	3.122	0.001845
Category=past participle	0.0091008	0.0031174	2.919	0.003586
Preceding context=low	0.0058361	0.0024404	2.391	0.016962

Table 2: Coefficients of a generalized linear fixed-effects model with an R^2 of 0.07564

Our second model, mixed-effects regression model, examines the effect of individual speaker and token variability. Table 4 presents random effects and Table 4 exhibits fixed effects.

Groups	Variance	Std. Deviation
Token	1.405e-05	0.003748
Speaker	1.174e-04	0.010835
Residual	8.857e-04	0.029761

Table 3. Random Effects: tokens and speakers

	Estimate	Std. Error	t-value	p-value
Intercept	9.299e-01	5.080e-03	183.053	< 2e-16
Following context=low	2.030e-02	4.356e-03	4.660	3.86e-06
Following context=mid	1.548e-02	4.327e-03	3.579	0.000375
Token frequency	1.121e-03	3.806e-04	2.944	0.004358
Category=past participle	8.391e-03	3.088e-03	2.717	0.006896
Preceding context=low	6.020e-03	2.412e-03	2.496	0.013260

Table 4: Fixed effects of a generalized linear mixed-effects model

Overall variance in this model is 0.001 (1.405e-05+1.174e-04+8.857e-04). Tokens represent only 1.4% of variation (1.405e-05/0.001), whereas speakers' variation is 11.5% of the data variation. Significant factors are the following, in order of their significance: following low vowel, following mid vowel, token frequency, past participle and preceding low vowel. Our model did not select any sociolinguistic factors, demonstrating that random effects for speakers are stronger than fixed effects. However, we should keep in mind that the model may not detect small population effects considering the small size of speakers (Johnson 2009, 2016). In contrast, random effects for word variation are less strong (only 1.4%), and the fixed effect for token frequency remains very significant. Similarly, following context remains by far the most significant factor favoring lenited variants ($p<0.000$). Finally, past participles and low preceding vowels also influence /d/ lenition ($p<0.01$ and $p<0.05$, respectively).

6. Discussion. The intensity ratio measurement reveals that intervocalic /d/ deletion is not the norm in the corpus of Caracas and that the lenited realization of intervocalic /d/ is more common in this speech community. In fact, the density distribution maintains its intensity peak at 0.95–0.96 across time from 1987 until 2004/2010. To examine the role of linguistic and extra-linguistic contexts on the lenition, we used *Language Variation Suite*, which implements state-of-the-art statistical methods. Its non-parametric tree-based analysis allowed us to interpret visually the role of independent factors. Furthermore, the comparison between fixed- and mixed-effects models provided a better understanding of group- and individual-level variation. First of all, in parametric and non-parametric tests, we found an effect of token frequency and following phonetic context: more frequent tokens and low vowel /a/ strongly favor more lenited realization of /d/. In addition, the grammatical category, namely past participle, appears to play a role in explaining the lenition process. These findings are consistent with the study by Díaz-Campos and Gradoville (2011), where frequency and *-ado* participles favor /d/ deletion. Concerning sociolinguistic factors, non-parametric tests and the fixed-effects regression model indicate a strong effect of age (younger speakers) and sex (male speakers) on lenited variants. In contrast, the mixed-effects model showed that individual variation in our corpus was higher than group variation (11.5%). As a result, none of social factors were selected.

Taken together, our comparative analyses show that by conceptualizing sociophonetic variable as continuous, we gain a better understanding of this phenomenon. In addition, advanced statistical practices offer a novel way to interpret the results of sociolinguistic multifactorial analysis.

7. Conclusion. This research project contributes to the statistical analysis of socio-phonological variables. Following the methodology from recent acoustic studies, the present investigation uses intensity ratio to measure the degree of lenition. Furthermore, this study addresses questions concerning the statistical analysis of gradient phonological variables by contrasting traditional variable rule analysis with the current practices of using mixed-effects modeling and tree-based analysis.

One of the novel implementations of this project is the creation of an interactive sociolinguistic toolkit that implements state-of-the-art statistical methods—Language Variation Suite.[13] The accessibility of the tool online and its user-friendly interface are two principal components that were missing from the previous sociolinguistic tools. In addition, the deployment of the tool on the Shiny server also increases its computational power: no longer beholden to the memory limitation of personal computers, statistical calculations can now run on a server.

References

Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis*. 2nd edition. John Wiley & Sons. http://dx.doi.org/10.1002/0470114754

Anthony, Laurence. 2010. AntConc (Version 3.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge University Press, Cambridge, U.K.

Bentivoglio, Paula & Irania Malaver. 2006. La lingüística de corpus en Venezuela: un nuevo proyecto. *Lingua Americana* 19. 37–46.

[13] Online: https://languagevariationsuite.wordpress.com/

Bentivoglio, Paula & Mercedes Sedano. 1993. Investigación sociolingüística: sus métodos aplicados a una experiencia venezolana. *Boletín de Lingüística* 8. 3–35.
Bierens, Herman J. 2004. Information Criteria and Model Selection. Lecture notes at the Pennsylvania State. Pennsylvania State University.
Boersma, Paul. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10). 341–345.
Carrasco, Patricio, José Ignacio Hualde & Miquel Simonet. 2012. Dialectal differences in Spanish voiced obstruent allophony: Costa Rican versus Iberian Spanish. *Phonetica* 69. 149–179. http://dx.doi.org/10.1159/000345199
Carrasco, Patricio. 2008. An acoustic study of voiced stop allophony in Costa Rican Spanish. PhD thesis. University of Illinois at Urbana-Champaign.
Cedergren, Henrietta & David Sankoff. 1974. Variable Rules: Performance as a Statistical Reflection of Competence. *Language* 50. 333–355.
Colantoni, Laura & Irina Marinescu. 2010. The scope of stop weakening in Argentine Spanish. In M. Ortega-Llebaria (ed.), *Selected Proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology*. 100-114. Somerville, MA: Cascadilla Proceedings Project.
Díaz-Campos, Manuel & Stephanie Dickinson. 2017. Using Statistics as a Tool in the Analysis of Sociolinguistic Variation. In Gabriel Rei-Doval and Fernando Tejedo-Herrero (eds.), *Lusophone, Galician and Hispanic Linguistics: Bridging Frames and Traditions*. To appear.
Díaz-Campos, Manuel & Michael Gradoville. 2011. An Analysis of Frequency as a Factor Contributing to the Diffusion of Variable Phenomena: Evidence from Spanish Data. In Luis Ortiz-Lopez (ed.), *Selected Proceedings of the 2009 Hispanic Linguistics Symposium*. Somerville, MA
Eddington, David. 2011. What are the contextual phonetic variants of /b d g/ in colloquial Spanish? *Probus* 23. 1–19.
Figueroa, Mauricio. 2014. There is hardly anything to hear, but we are hearing it nonetheless. *Workshop on Ibero-Romance Phonology and Morphology*. London, United Kingdom.
Foulkes, Paul. 2006. Phonological variation: A global perspective. In B. Aarts and A. McMahon (eds.), *Handbook of English Linguistics*. 625–669. Oxford: Blackwell.
Gorman, Kyle & Daniel Ezra Johnson. 2013. Quantitative analysis. *The Oxford handbook of sociolinguistics*. 214–240. Oxford University Press.
Jenset, Gard B. 2010. A corpus-based study on the evolution of There: Statistical analysis and cognitive interpretation. Phd thesis. University of Bergen.
Johnson, Daniel Ezra. 2016. Progress in regression: why sociolinguistic data calls for mixed-effects models. *Language Variation and Change.* To appear.
Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed effects variable rule analysis. *Language and Linguistics Compass* 3. 359–383.
Henríquez Ureña, Pedro. 1921. Observaciones sobre el español de América. *Revista de Filología española* VIII. 357–390.
Hothorn, Torsten & Achim Zeileis. 2015. partykit: A Modular Toolkit for Recursive Partytioning in R. *Journal of Machine Learning Research* 16. 3905–3909.
Hualde, José Ignacio, Ryan Shosted & Daniel Scarpace. 2011. Acoustics and articulation of Spanish /d/ spirantization. *Proceedings of the International Congress of Phonetic Sciences XVII.* 906–909.
Hualde, José Ignacio, Miquel Simonet & Mariana Nadeu. 2011. Consonant lenition and phonological recategorization. *Journal of Laboratory Phonology* 2. 301–329.

Kidhardt, Adrian. 2015. New Directions in Quantitative Hispanic Sociolinguistics. MA thesis. Arizona State University.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45(4). 715–762.
Lapesa, Rafael. 1981. *Historia de la lengua española*. Madrid: Gredos.
Limanni, Anna. 2009. Men, women and lenition: Gender differences in the production of intervocalic voiced stops in Mexican Spanish. *Canadian Acoustics* 37(3). 194–195.
Martínez-Celdrán, Eugenio. 1991. Sobre la naturaleza fonética de los alófonos de /b d g/ en español y sus distintas denominaciones. *Verba* 18. 235–253.
Navarro Tomás, Tomás. 1999 [1918]. *Manual de pronunciación española*. Madrid: Consejo Superior de Investigaciones Científicas.
Oroz, Rodowo. 1966. *La lengua castellana en Chile*. Santiago: Ed. Universitaria.
Quilis, Antonio. 1981. *Fonética acústica de la lengua española*. Gredos: Madrid.
RStudio Team. 2015. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. Available from http://www.rstudio.com/.
Sankoff, David, Sali Tagliamonte & Eric Smith. 2005. Goldvarb X: A variable rule application for Macintosh and Windows. Department of Linguistics, University of Toronto.
Sapir, Edward. 1921. *Language: An introduction to the study of speech*. New York: Brace & World.
Simonet, Miquel, José Ignacio Hualde & Marianna Nadeu. 2012. Lenition of /d/ in spontaneous Spanish and Catalan. In *13th Annual Conference of the International Speech Communication Association 2012* 2. 1414–1417.
Strobl, Carolin, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods* 14. 323–348.
Tagliamonte, Sali & R. Harald Baayen. 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24. 135–178.
Tagliamonte, Sali. 2011. Language in Society: Variationist Sociolinguistics : Change, Observation, Interpretation. Hoboken, NJ: Wiley-Blackwell.
Tagliamonte, Sali. 2006. *Analyzing Sociolinguistic Variation*. Cambridge University Press.
Thomas, Eric R. 2013. Sociophonetics. In J. K. Chambers and N. Schilling-Estes (eds.), *The Handbook of Language Variation and Change*. 108–127. Oxford, UK/ Malden, MA: Wiley-Blackwell.
Torreira, Francisco & Myriam Ernestus. 2011. Realization of voiceless stops and vowels in conversational French and Spanish. *Laboratory Phonology* 2(2). 331–353. http://dx.doi.org/10.1515/labphon.2011.012
Zamora, Vicente Alonso. 1970. *Dialectología española*. Segunda edición, Biblioteca Románica Hispánica. Madrid: Editorial Gredos S.A.