GAMs: Textual results and residuals plots

1. Overview

This document provides the textual results (`summary()`, `gam.check()`, `AIC()`) and residuals plots of the GAMs discussed in the paper in §3.3. The combined sets of four GAM residuals plots were produced with the package `mgcViz` for R (Fasiolo et al. 2018). The models focus on Northern Sub-Saharan Africa as the area for which have data on the lexical frequency of LV stops. The data points fall in the range of the longitude interval [-18°, 36°] and latitude interval [-9°, 16°].

The GAMs and their visualizations were produced with the `mgcv` package for R (Wood 2006, 2019). The GAM visualizations are contour plots representing the regression surface of the lexical LV frequencies (F_{LV}) as a function of the combination of longitude and latitude using thin-plate regression splines with the heat map color scheme. Lighter shades correspond to higher F_{LV} values. Contour lines are isopleths that mark deviations from the mean in terms of standard deviation. The parameter `too.far` controlling the size of the area to be plotted around each data point was set to 0.05 (half of its default value of 0.1) to strike a balance between the accurateness in the representation of the spatial continuity between the data points and the ease of perception of the visualization as a whole, avoiding too much patchiness in the contour plot.

We used the `gam()` function with the default Gaussian distribution as we could not identify any other distribution that would match our data better and would produce significantly better results in terms of deviance explained, the number of the basis functions used and the distribution of residuals.[1] The smoothing parameter estimation method was set to `REML` (restricted maximum likelihood) with extra penalization (`select = TRUE`) to reduce the risk of undersmoothing (overfitting) and of variability in estimates of the smoothing parameter (cf. Miller 2017). The number of basis functions (the parameter `k`) of GAMs was selected with the `gam.check()` function

[1] We also tried using the heavy-tailed scaled-T distribution which was particularly successful in singling out the Cameroon Gap but overall did not score as well as the default Gaussian option. A reviewer also suggested using a binomial distribution or a beta distribution. However, a binomial distribution implies a binary dependent variable (absence or presence of LV stops) rather than a gradient one (lexical frequency of LV stops), which defies the purpose of our research. A beta distribution accepts a gradient dependent variable but requires rescaling of our frequency data to the interval (0,1) and also excludes the data points without LV stops ($F_{LV} = 0$). Under these conditions, the beta distribution did not score better than the Gaussian option. To be able to take into account the large number of zero values in our data while using a beta distribution, we tried to use zero-inflated beta distributions in the `gamlss` package for R (Rigby & Stasinopoulos 2005, Stasinopoulos et al. 2019) but the resulting GAMs produced visualizations that were much less clear and were marred with visualization artifacts that did not correspond to anything in the data.

by choosing between the values of `k-index` closest to 1 (with the p-value above 0.1; `edf` was always significantly lower than `k'`) the one that produced the model with the best AIC value.

2. GAM in Figure 6

The F_{LV} frequencies are in percentages and include 0% for languages without LV stops.

2.1. Textual results

```
> summary(LVall.NSSA.k13.REML.gam)

Family: gaussian
Link function: identity

Formula:
LVFreq ~ te(long, lat, bs = "tp", k = 13)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.9508     0.5536   36.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
               edf Ref.df     F p-value
te(long,lat) 70.39    167 13.12  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.752   Deviance explained = 77.6%
-REML = 3091.3  Scale est. = 222.23    n = 725


> gam.check(LVall.NSSA.k13.REML.gam)

Method: REML   Optimizer: outer newton
full convergence after 9 iterations.
Gradient range [-1.349752e-05,4.360728e-06]
(score 3091.277 & scale 222.2299).
Hessian positive definite, eigenvalue range [0.4779333,363.716].
Model rank =  169 / 169

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                k'   edf k-index p-value
te(long,lat) 168.0  70.4    0.99    0.29


> AIC(LVall.NSSA.k13.REML.gam)
[1] 6047.775
```

2.2. Residuals plots

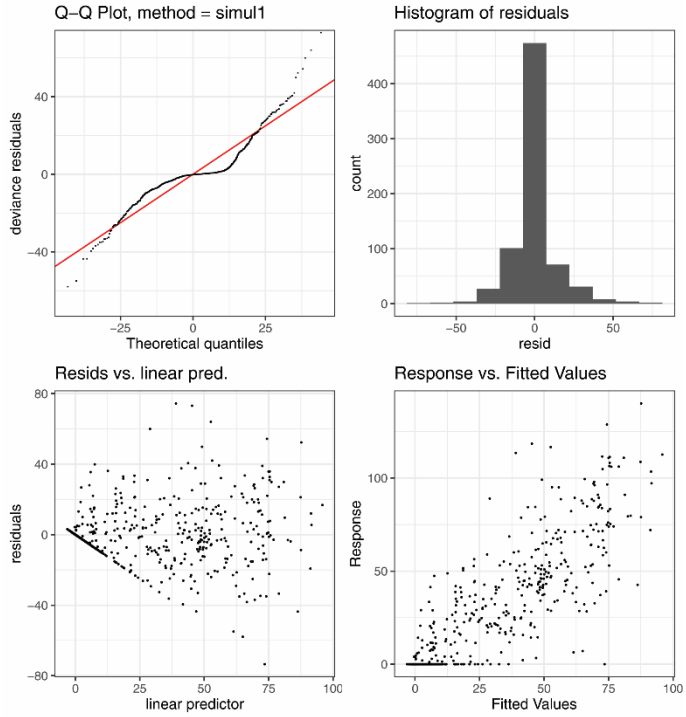

FIGURE S.1. The four residuals plots for the GAM of the dataset in percentages, including the languages without LV stops as 0%, as visualized in Figure 6 (§3.3).

3. GAM in Figure 7

The F_{LV} frequencies are in percentages with all languages without LV stops (0%) removed.

3.1. Textual results

```
> summary(LVFreq.k15.REML.gam)

Family: gaussian
Link function: identity

Formula:
LVFreq ~ te(long, lat, bs = "tp", k = 15)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   45.919      1.152   39.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
               edf Ref.df     F p-value
te(long,lat) 29.83    190 1.779  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.518   Deviance explained = 56.4%
-REML = 1435.6  Scale est. = 417.71    n = 315
```

```
> gam.check(LVFreq.k15.REML.gam)

Method: REML   Optimizer: outer newton
full convergence after 6 iterations.
Gradient range [-4.393715e-05,3.93072e-05]
(score 1435.614 & scale 417.7074).
Hessian positive definite, eigenvalue range [1.101378e-05,157.7352].
Model rank =  225 / 225

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                 k'   edf k-index p-value
te(long,lat) 224.0  29.8    0.99    0.28

> AIC(LVFreq.k15.REML.gam)
[1] 2830.881
```

3.2. Residuals plots

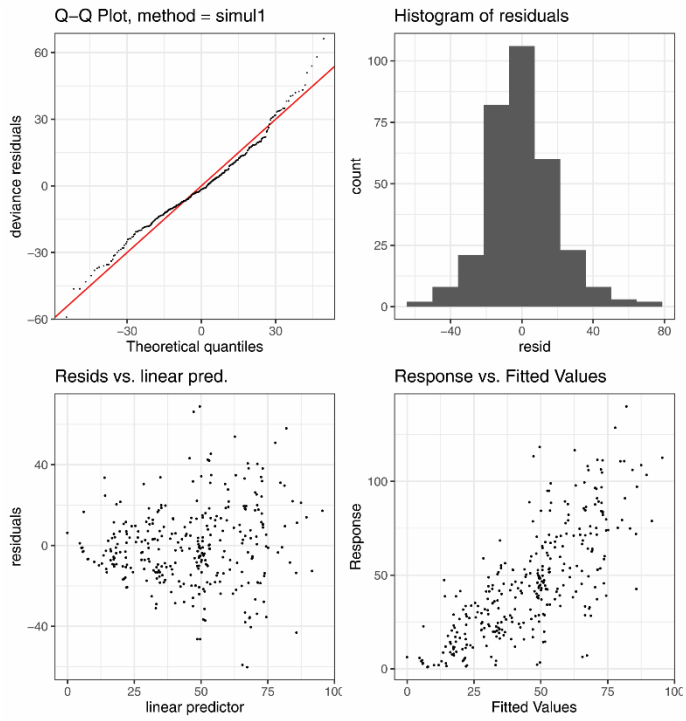

FIGURE S.2. The four residuals plots for the GAM of the dataset in percentages with all languages without LV stops removed, as visualized in Figure 7 (§3.3).

4. GAM in Figure 8a

The F_{LV} frequencies in percentages are from the subset of the full dataset that includes only the 178 languages with LV stops for which our lexical source in RefLex has at least 400 entries (cf. §2.3) plus languages without LV stops with F_{LV} of 0%.

4.1. Textual results

```
> summary(LVall.NSSA.Sw200.k11.REML.gam)

Family: gaussian
Link function: identity

Formula:
LVFreq ~ te(long, lat, bs = "tp", k = 11)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.3998     0.5401   24.81   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
               edf Ref.df     F p-value
te(long,lat) 64.54    119 14.03  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =   0.74   Deviance explained = 76.8%
-REML = 2450.1  Scale est. = 171.5     n = 588


> gam.check(LVall.NSSA.Sw200.k11.REML.gam)

Method: REML   Optimizer: outer newton
full convergence after 5 iterations.
Gradient range [-0.0001840783,0.0001245249]
(score 2450.122 & scale 171.5035).
Hessian positive definite, eigenvalue range [6.656047e-05,295.3576].
Model rank =  121 / 121

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                k'   edf k-index p-value
te(long,lat) 120.0  64.5    1.03    0.74


> AIC(LVall.NSSA.Sw200.k11.REML.gam)
[1] 4760.321
```

4.2. Residuals plots

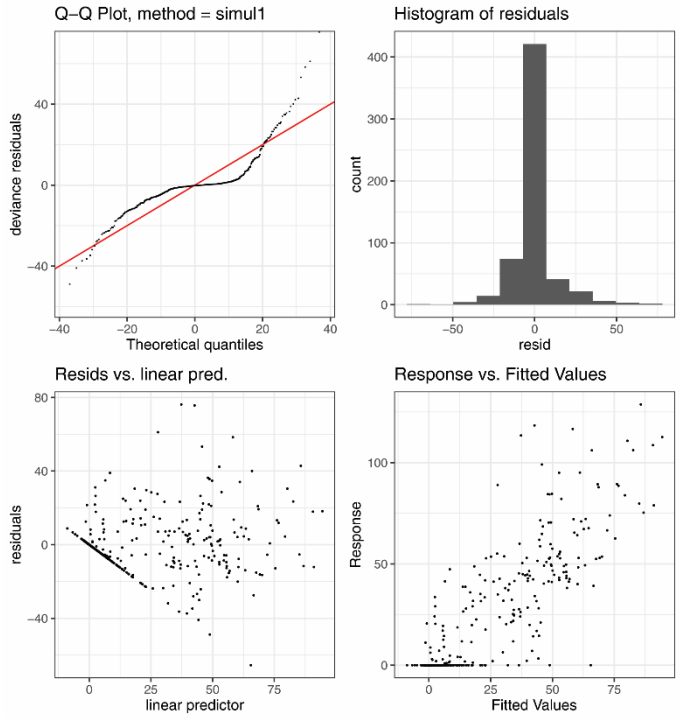

FIGURE S.3. The four residuals plots for the GAM of the F_{LV} frequencies (in percentages) for the subset of the full dataset that includes only the 178 languages with LV stops for which our lexical source in RefLex has at least 400 entries (cf. §2.3) plus languages without LV stops with F_{LV} of 0%. The GAM is visualized in Figure 8a (§3.3).

5. GAM in Figure 8b

The F_{LV} frequencies in percentages are from the quasi Swadesh 200 lists of the subset of the full dataset that includes only the 178 languages with LV stops for which our lexical source in RefLex has at least 400 entries (cf. §2.3) plus languages without LV stops with F_{LV} of 0%.

5.1. Textual results

```
> summary(LVall.NSSA.Sw200.LVFreq200.k11.REML.gam)

Family: gaussian
Link function: identity

Formula:
LVFreq200 ~ te(long, lat, bs = "tp", k = 11)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.7353     0.5731   16.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
              edf Ref.df     F p-value
```

```
te(long,lat) 54.49    120 6.581  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.574   Deviance explained = 61.3%
-REML = 2460.3  Scale est. = 193.09    n = 588
```

```
> gam.check(LVall.NSSA.Sw200.LVFreq200.k11.REML.gam)

Method: REML   Optimizer: outer newton
full convergence after 8 iterations.
Gradient range [-0.001534484,0.0007170387]
(score 2460.309 & scale 193.0923).
Hessian positive definite, eigenvalue range [0.07868284,294.7809].
Model rank =  121 / 121

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                k'   edf k-index p-value
te(long,lat) 120.0  54.5    1.05    0.79
```

```
> AIC(LVall.NSSA.Sw200.LVFreq200.k11.REML.gam)
[1] 4823
```

5.2. Residuals plots

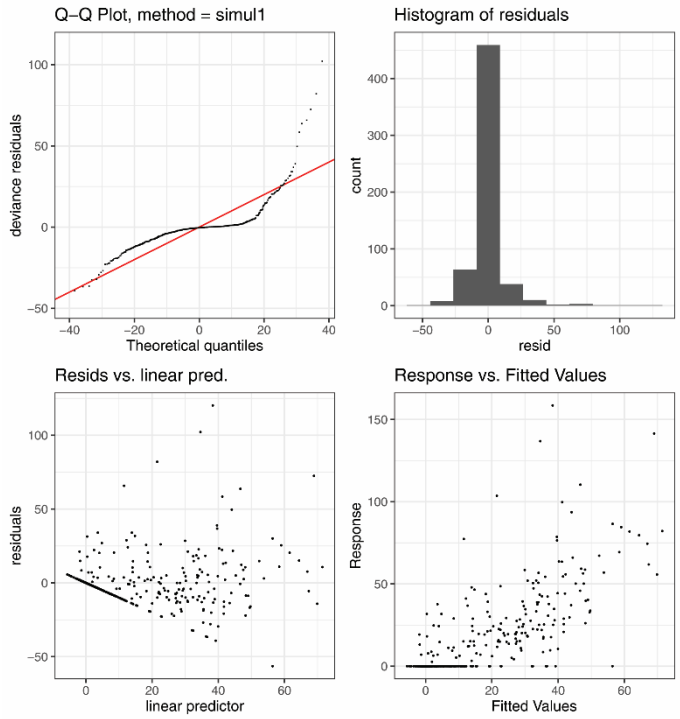

FIGURE S.4. The four residuals plots for the GAM of the F_{LV} frequencies (in percentages) in the quasi Swadesh 200 lists for the subset of the full dataset that includes only the 178 languages with LV stops for which our lexical source in RefLex has at least 400 entries (cf. §2.3) plus languages without LV stops with F_{LV} of 0%. The GAM is visualized in Figure 8b (§3.3).

6. GAM in Figure 9

The F_{LV} frequencies, including 0% for languages without LV stops, have been scaled up by 0.83 (the minimal F_{LV} value different from zero) and log-transformed.

6.1. Textual results

```
> summary(LVall.NSSA.lg.k18.REML.gam)

Family: gaussian
Link function: identity

Formula:
LVFreq ~ te(long, lat, bs = "tp", k = 18)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.54837    0.02821   54.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
               edf Ref.df     F p-value
te(long,lat) 108.1    317 11.41  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.833   Deviance explained = 85.8%
-REML = 983.28  Scale est. = 0.57688   n = 725


> gam.check(LVall.NSSA.lg.k18.REML.gam)

Method: REML   Optimizer: outer newton
full convergence after 5 iterations.
Gradient range [-3.8556e-07,4.125533e-05]
(score 983.2789 & scale 0.5768804).
eigenvalue range [-4.125346e-05,366.2168].
Model rank =  324 / 324

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

              k' edf k-index p-value
te(long,lat) 323 108       1    0.48


> AIC(LVall.NSSA.lg.k18.REML.gam)
[1] 1764.077
```

6.2. Residuals plots

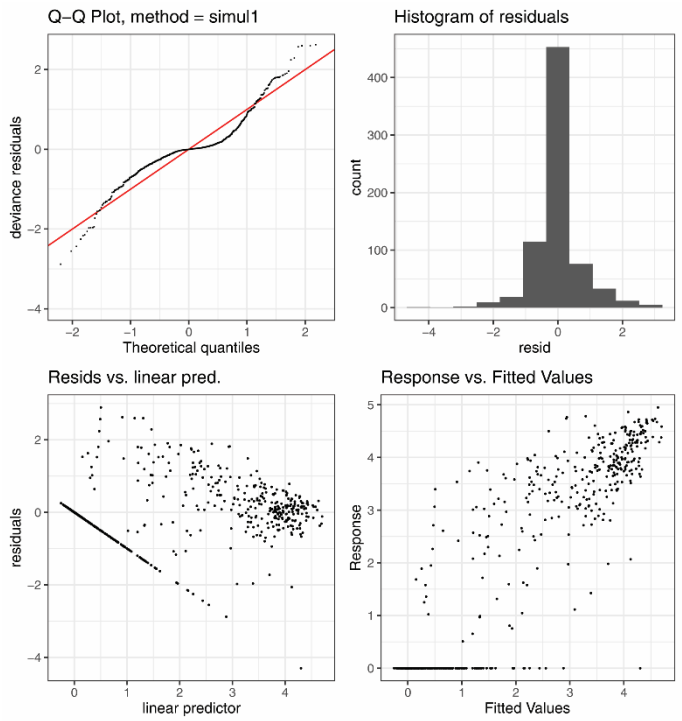

FIGURE S.5. The four residuals plots for the GAM of the log-transformed (after scaling up by 0.83) F_{LV} frequencies (including the languages without LV stops), as visualized in Figure 9 (§3.3).

REFERENCES

MILLER, DAVID L. 2017. Why is the default smoothing method "REML" rather than "GCV.Cp"?. Online: https://github.com/DistanceDevelopment/dsm (Accessed on 28 May, 2020).

RIGBY, ROBERT A., and MIKIS D. STASINOPOULOS. 2005. Generalized additive models for location, scale and shape. *Applied Statistics* 54(3).507–554.

STASINOPOULOS, MIKIS D.; ROBERT A. RIGBY; VLASIOS VOUDOURIS; CALLILOPE AKANTZILIOTOU; MARCO ENEA; and DANIIL KIOSE. 2019. *gamlss: Generalised Additive Models for Location Scale and Shape*. Online: http://CRAN.R-project.org/package=gamlss.