Explaining Worldwide Unemployment Rates
A Spatial Generalized Additive Model Analysis

Kaylee Hodgson

STAT 637 - Winter 2019

1   Introduction

Political leaders worldwide are concerned with their countries' unemployment rates, and actively look for ways to decrease unemployment. Increased globalization has especially forced political leaders to compare their economic status to that of other countries. Comparing these rates with other countries likely leads politicians to wonder what factors drive the variation in unemployment.

In response, the purpose of this study is to build a well-fit model to explain the variation in unemployment rates, and to find country-level variables that significantly describe these unemployment rates. Because the response values are proportions, different generalized linear models are explored and compared to analyze the data. Unemployment rates appear to be spatially correlated (see Figure 1), so generalized additive models are also explored, which implement the spatial component in the analysis. Finally, variable selection is performed to find a more parsimonious model that only includes the variables that significantly describe unemployment rates. The results of the final model are reported and discussed.

2   The Data

The data on unemployment rates for countries used in this study are from The World Bank, and measure the unemployment as a percentage of the total labor force for each country. The data, although originally measured as percentages, are transformed to proportions by dividing each value by 100.

The distribution of unemployment rates by country is shown in the map in Figure 1, where darker colors indicate higher unemployment. Some interesting patterns appear in this map. The distribution of unemployment rates does not appear to be directly tied to the development level of a country. The four countries in this map with the lowest unemployment rates are Qatar, Cambodia, Niger, and Belarus, and the countries with the highest are Swaziland, South Africa, Palestine, and Lesotho. The region that appears, from the map, to have the most countries with lower unemployment rates is Southeast Asia.

The response variables are from a variety of sources, and attempt to cover the spectrum of macro-level indicators that were predicted to significantly impact unemployment rates, including the security and stability of a country, education levels, health, demographics, economic freedom and status, treatment of women and minorities, and the level of violence. The explanatory variables are listed below with their source and description:

- **State Monopoly on the Use of Force** (The Quality of Government Institute): Scale from 1-10 that measures the extent of the country's territory where the state has a monopoly on the use of force. A score of 1 indicates the state has no monopoly, and 10 indicates that the state has a complete monopoly.
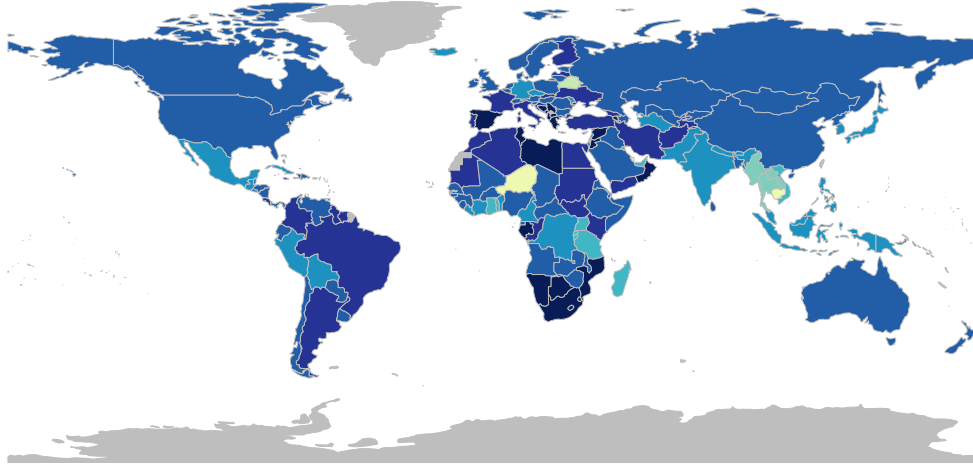
Figure 1: Worldwide Unemployment Rates (darker colors indicate higher unemployment)

- **Societal Violence Scale** (The Political Terror Scale): Scale from 1-5 that measures the scope and severity of violence in the society, where higher numbers indicate higher levels of violence.
- **Regime Types** (The Quality of Government Institute): Ordinal scale that measures the regime type with the following levels: 1) "Pure Autocracy," 2) "Inclusive Autocracy," 3) "Liberal Autocracy," and 4) "Minimal Democracy."
- **Private Property Rights** (Social Progress Index): Scale from 0-100 that measures the "extent to which a country's legal framework allows individuals to freely accumulate private property," where higher scores indicate more legal protection.
- **Social Safety Nets** (The Quality of Government Institute): Scale from 1-10 that measures the extent to which a country's government and institutions provide "welfare, unemployment benefits, universal healthcare, free education, workers compensation and so forth which cushion individuals from falling into poverty beyond a certain level."
- **Youth Risk Factor** (Bricker & Foley, 2013): Number of 17-26 year olds as a ratio of the country's labor force.
- **Average Years of Schooling** (UNDP Human Development Reports): Average number of years of education for people 25 years and older.
- **Life Expectancy** (World Health Organization): Average number of years that a newborn is expected to live.
- **Discrimination and Violence against Minorities** (Social Progress Index): Scale from 0-10 that measures the level of "discrimination, powerlessness, ethnic violence, communal violence, sectarian violence, and religious violence."
- **Government Framework for Gender Equality** (The WomanStats Project): Scale from 0-7 that measures the "legislative, practical and international determinants of feminist government policies."
- **GDP per Capita (based on PPP in current prices)** (International Monetary Fund's World Economic Outlook Database): The GDP is "converted to international dollars using purchasing power parity rates and divided by total population."

Hypothesis: These macro-level indicators significantly describe the variation in countries' unemployment rates.

3   Methods

In this section, multiple models are proposed and compared to find the best fitting model for the data. The beta and the binomial generalized linear models are compared first (with and without weights), then the model chosen is compared to that same model with a spatial component added, and finally, variable selection is performed with the final spatial model.

3.1 Beta vs. Binomial Model

The beta regression model assumes that $\mathbf{y}$ has a beta distribution, but the density function is reparametrized to improve the ease of interpretability for the model. The parameterization results in the following density function:

$$f(y_i|\mu_i, \phi_i) = \frac{\Gamma(\phi_i)}{\Gamma(\mu_i\phi_i)\Gamma((1-\mu_i)\phi_i)}y^{\mu_i\phi_i-1}(1-y_i)^{(1-\mu_i)\phi_i-1}, \, 0 < y_i < 1 \tag{1}$$

$$\text{where } \mu_i = \frac{p_i}{(p_i + q_i)} \quad \text{and} \quad \phi_i = p_i + q_i$$

The expected value and variance then become: $\text{E}(y_i) = \mu_i$ and $\text{Var}(y_i) = \frac{\mu_i(1-\mu_i)}{1+\phi_i}$, making the regression model more interpretable because one of the new parameters is the expected value of $\mathbf{y}$, and making the variance of $\mathbf{y}$ naturally heteroskedastic.

The beta regression model with the logit link is then written as:

$$y_i \sim \text{Beta}(\mu_i, \phi_i)$$
$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \mathbf{x}_i'\boldsymbol{\beta} \tag{2}$$

The binomial regression model assumes that $\mathbf{y}$ has a binomial distribution, with the original parameterization:

$$f(y_i|\pi_i, m_i) = \binom{m_i}{y_i}p_i^{y_i}(1-p_i)^{m_i-y_i}, \, y_i \in \{0, 1, ..., m_i\}, \tag{3}$$

where y_i is the number of success for each country. The binomial regression model with the logit link is then written as:

$$y_i \sim \text{Binomial}(\pi_i, m_i)$$
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i'\boldsymbol{\beta} \tag{4}$$

In the beta model, the response is the log odds of μ_i, or the mean value of y_i, where y_i is the unemployment rate. In the binomial model, the response is the log odds of π_i, or the proportion of "successes" (y_i/m_i), which is the unemployment rate.

The response data are proportions without the total number and the number of "successes" for each observation, so the values for the m_i's in the binomial model are explored. First, all $m_i = 100$ to give equal weight to each observation in the model. Then, the model is weighted by population by setting $m_i = \text{Population}_i$. The beta model is also fit with and without the population rates. The results from these comparisons are given in Figure 2 and Table 1.

The residual plots in Figure 2 show that the beta model without the population weights has the smallest residuals. In Table 1, the AIC, BIC, and residual deviance are reported for reference, but are not useful for comparing across these completely different models. However, the in-sample MSE is compared across the four models and both of the beta models have lower MSE's than the binomial models, and the beta model without the weights has the lowest. The residual plots and the MSE indicate that the beta regression model has the best fit, so that model is explored further with an added spatial component in the next section.
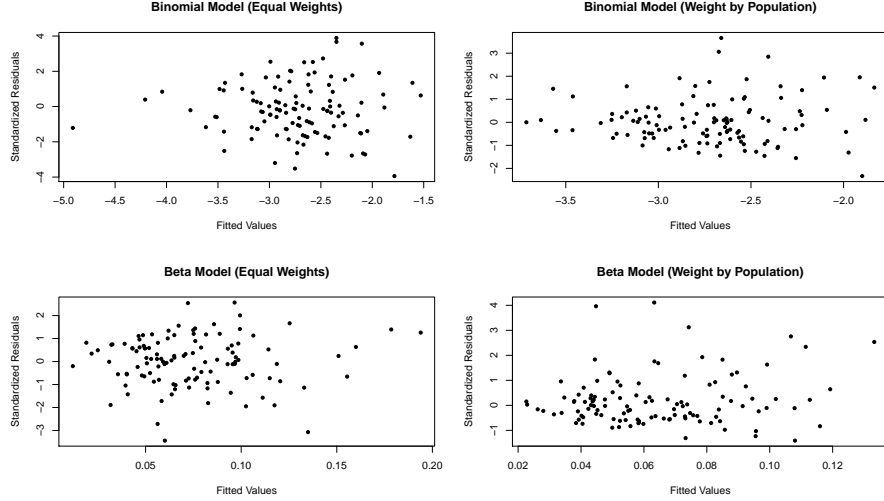
Figure 2: Residual plots for beta and binomial models with and without the population weights.

	Binomial	Binomial (Weighted)	Beta	Beta (Weighted)
AIC	704.70	66848430.91	-399.50	-28598994453.31
BIC	751.37	66848477.58	-350.10	-28598994403.91
Deviance	274.50	66846636.67	132.53	8231111993.63
MSE	0.339	0.490	0.002	0.003

Table 1: Model diagnostics for beta and binomial models with and without population weights.

3.2 Spatial Generalized Additive Model

A generalized additive model (GAM), a semi-parametric model, is used to add the spatial component to the beta regression model. Like the generalized linear model (GLM), the GAM estimates linear coefficients, but also adds a smoothing component to the model which essentially functions as a random effect, so that the structure becomes:

$$g(y_i) = \mathbf{x}_i'\boldsymbol{\beta} + f(\mathbf{z}_i), \tag{5}$$

where $\mathbf{z}_i = (\text{latitude}_i, \text{longitude}_i)$. This particular model uses Gaussian Process smoothing with a power exponential correlation function. The smoothing function is then structured as follows:

$$f(\mathbf{z}_i) \sim \text{ GP}(\text{m}(\mathbf{z}_i), \text{cov}(\mathbf{z}_i, \mathbf{z}_j)), \tag{6}$$

with the following correlation function:

$$\text{R}(\tau) = \exp(-\tau), \text{ where } \tau = ||\mathbf{z}_i - \mathbf{z}_j||. \tag{7}$$

The beta model with the spatial component is compared to the original beta model in the previous section in Figure 3 and Table 2. The residual plots in Figure 3 for the two models are very similar and, while the AIC indicates that the spatial beta model may fit better, the BIC indicates that the beta model without the spatial aspect may fit better. However, because the residual deviance is clearly smaller with the spatial component than without, and because there seems to be some spatial correlation from Figure 1, the spatial beta model is used instead of the original beta model.
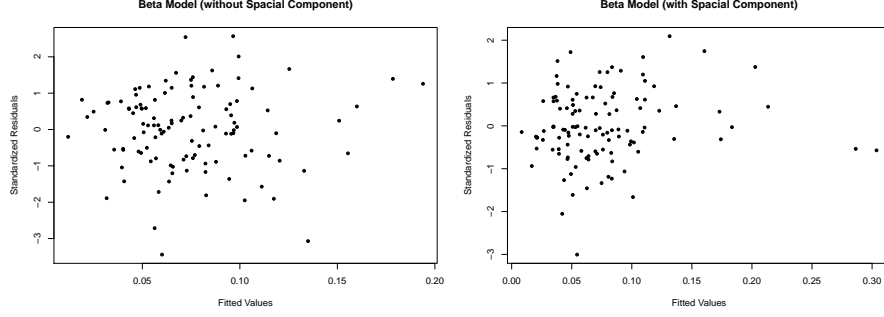
Figure 3: Residual plots for the beta model with and without the spatial component.

	Beta	Spatial Beta
AIC	-399.50	-446.01
BIC	-350.10	-341.00
Deviance	132.53	80.51

Table 2: Model diagnostics to compare the beta model with and without the spatial component.

3.3 Variable Selection

Backward selection is used for the variable selection, with the spatial beta GAM. The backward selection is performed by deleting the least significant variable at each step until all variables remaining in the model are significant. The diagnostics for the model at each step of the backward selection are displayed in Figure 4. Both the AIC and BIC decrease (improve) fairly steadily at each step of the backward selection, and the residual deviance unsurprisingly increases, but not significantly, at each step. The adjusted R^2 actually increases at each step of the backward selection. This is likely because the variables have different missing values, so as each variable is deleted, the number of observations included in the model increases.

The final variables included in the model after backward selection are the Societal Violence Scale, Regime Type, GDP per Capita, and Social Safety Nets.

3.4 Density Comparison

Although model diagnostics are compared across models in previous sections, the distributional assumptions of the models are also compared to the density of the data here. Figure 5 plots the density of the observed data against the density of the fitted values from 1) the original beta model, 2) the beta model with the added spatial component, and 3) the beta model with the added spatial component after variable selection. The density of the fitted values from the third, or final, model much more closely matches the density of the observed data than the other two models.

This diagnostic again confirms that the final model chosen is the best fit model, and is therefore used in the next section to model the variation in unemployment rates.

4 Results and Discussion

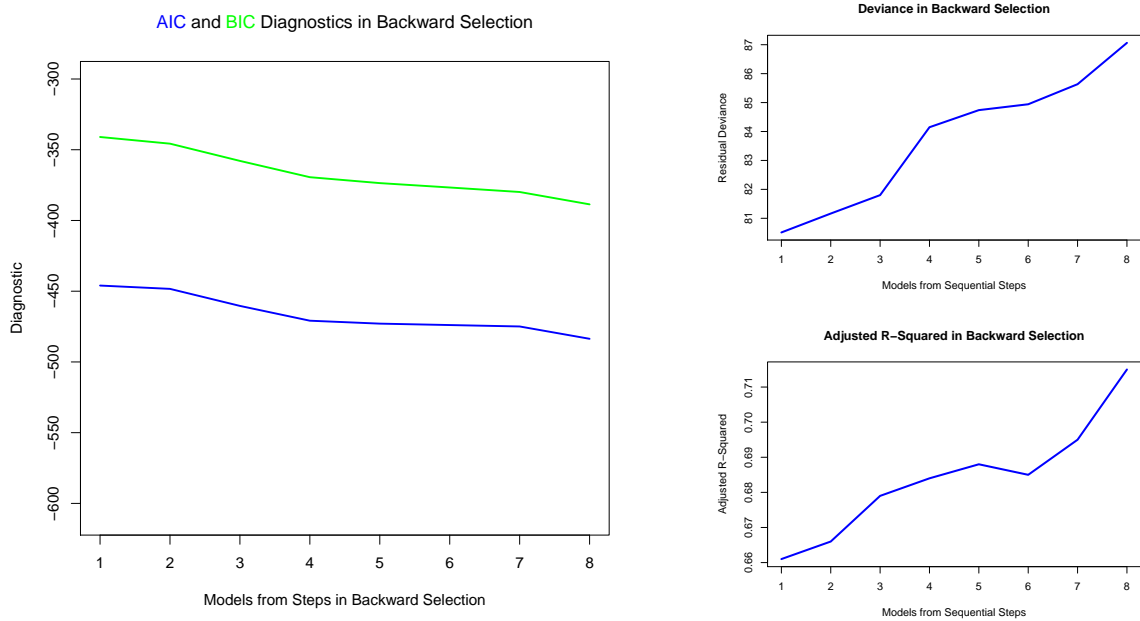The final beta GAM after variable selection is:

Figure 4: Diagnostics for the model at each step of backward selection: AIC/BIC, Adjusted R^2, and Residual Deviance.
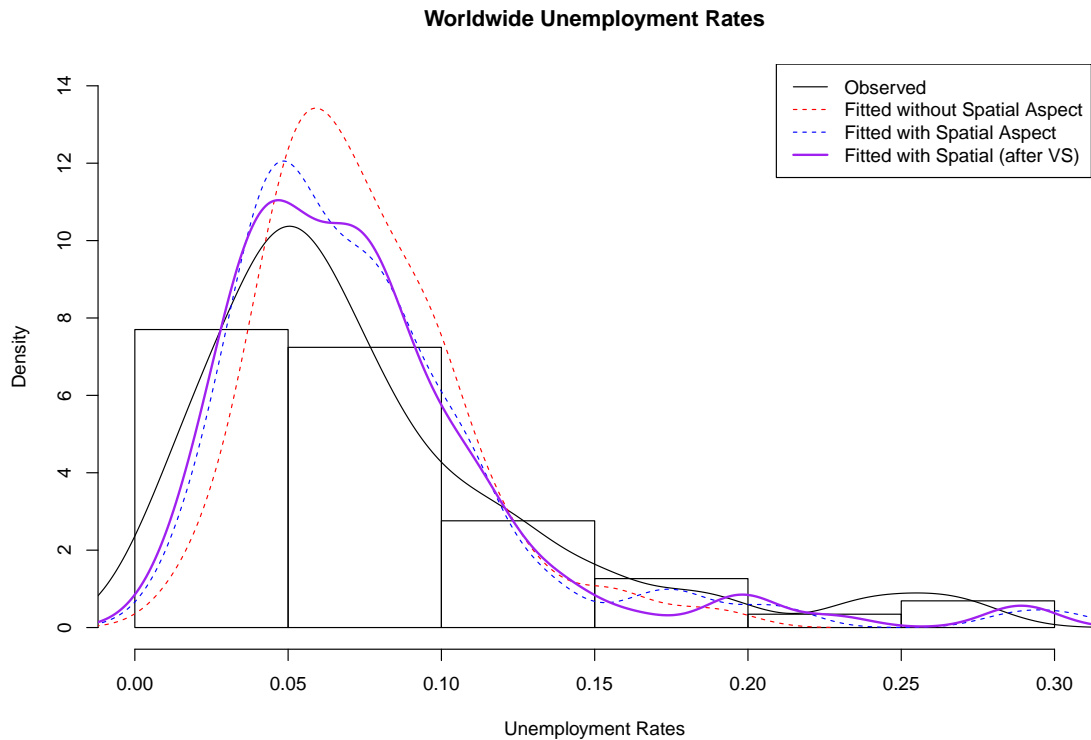

Figure 5: Comparison of the observed unemployment rate density to 1) the original beta model, 2) the beta model with the added spatial component, and 3) the beta model with the added spatial component after variable selection (VS).

$$y_i \sim Beta(\mu_i, \phi_i)$$

$$g(y_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) = \mathbf{x}'_i\boldsymbol{\beta} + f(\mathbf{z}_i) \qquad (8)$$

$$f(\mathbf{z}_i) \sim \; \mathrm{GP}(\mathrm{m}(\mathbf{z}_i), \mathrm{cov}(\mathbf{z}_i, \mathbf{z}_j)).$$

The adjusted R^2 for the final model is 0.715, and the deviance explained is 77.1%. The estimates from the model are reported in Table 3, where the baseline for the Social Violence Scale is 1, or the lowest level of violence, and the baseline for the Regime Type is "Pure Autocracy."

	Estimate	Standard Error	z value	p-value
(Intercept)	-1.763	0.4492	-3.926	0.000
Societal Violence Scale: 2	-1.255	0.4334	-2.895	0.004
Societal Violence Scale: 3	-1.267	0.4326	-2.930	0.003
Societal Violence Scale: 4	-1.500	0.4452	-3.370	0.001
Societal Violence Scale: 5	-1.028	0.4467	-2.301	0.021
Regime Type (Inclusive Autocracy)	0.066	0.2557	0.258	0.797
Regime Type (Liberal Autocracy)	1.161	0.2066	5.621	0.000
Regime Type (Minimal Democracy)	0.160	0.1459	1.098	0.272
GDP per Capita	0.0004	0.0006	-6.155	0.000
Social Safety Nets	0.156	0.0432	3.607	0.000
	Estimate	Degrees of Freedom	Chi-Squared value	p-value
s(Latitude,Longitude)	22.44	26.9	152	0.000

Table 3: Final spatial beta model coefficient estimates after variable selection.

While many of the variables originally included in the model are not significant indicators of unemployment rates, as hypothesized, the four variables remaining are. However, the directionality of the relationships in the model may be surprising. For example, as GDP per capita increases the unemployment rate also increases on average, and societies with higher levels of violence have lower average unemployment rates. Specific interpretations of the coefficient estimates are in terms of the log odds of the average unemployment rates. For example, as the GDP per capita increases by 1, the log odds of the average unemployment rate increases by 0.0004 on average.

The relationships estimated in this model are less surprising considering the unemployment rate distribution displayed in Figure 1, which indicates that many less stable and less developed countries have low unemployment rates. The relationships indicated in Table 3 help to explain variation in unemployment rates, but are not causal, and further theoretical and statistical research should explore the reasons for these relationships, and should explore what countries with lower unemployment rates are doing "right" that is leading to the lower unemployment, despite worse economic, political, and safety status.

5 Conclusion

Unemployment impacts the lives of people across the globe, and policy-makers therefore look for explanations of unemployment rate variations, in order to find ways to improve the economic status of their constituents. The beta GAM explains 77.1% of the deviance of unemployment rates, using a spatial component with Gaussian Process smoothing. Four out of the eleven variables explored (and hypothesized to significantly affect unemployment rates) are significant to the model: Societal Violence Scale, Regime Type, GDP per Capita, and Social Safety Nets. Future research should explore the theoretical reasons for the results in this paper and look for other "positive" determinants of lower unemployment.

References

[1] Bricker, Noah and Mark Foley (2013) "The Effect of Youth Demographics on Violence: The Importance of the Labor Market," *International Journal of Conflict and Violence*, Vol 7: (1) 179-194.

[2] Dahlberg, Stefan, Soren Holmberg, Bo Rothstein, Anna Khomenko & Richard Svensson (2016). *The Quality of Government Basic Dataset*, version Jan16. University of Gothenburg: The Quality of Government Institute, http://www.qog.pol.gu.sedoi:10.18157/QoGBasJan16.

[3] Gibney, Mark, Linda Cornett, Reed Wood, Peter Haschke, Daniel Arnon, and Attilio Pisanò (2018). The Political Terror Scale 1976-2017. http://www.politicalterrorscale.org.

[4] International Monetary Fund (2015). *World Economic Outlook Database* (2015). https://www.imf.org/en/Data.

[5] Social Progress Index (2016). *The Social Progress Imperative.* https://www.socialprogress.org.

[6] United Nations Development Programme (2015). *Human development report 2015.* Retrieved from http://hdr.undp.org/sites/default/files/hdr15-report-en-1.pdf.

[7] The WomanStats Project (2018). *WomanStats Project Database.* http://www.womanstats.org

[8] The World Bank (2018). *World Bank Open Data.* https://data.worldbank.org

[9] World Health Organization (2015). *SDG Health and Health-related Target Indicators.* http://www.who.int/gho/en/.