Urbanization and the Association between Northern Cardinal Observation in United States Counties

By Ian Brandenburg

1. Introduction

The Northern Cardinal (Cardinalis cardinalis) is one of the most observed bird species in the U.S., commonly found in backyards and well-recognized nationwide. It is referred to as a 'backyard' bird species, which means it is a species that is very commonly found in backyards or urbanized habitats. The Northern Cardinal was selected as a representative backyard bird species. This study investigates how human populations and urban development are associated with Northern Cardinal observation frequencies in U.S. counties. Given their adaptation to certain human-developed areas for food sourcing, this research examines the impact of urban development and population changes on these birds' observation levels. Focusing on 2022 data, the study analyses County-Level Rural-Urban Continuum Codes, Urban Influence Codes, Economic Typology, Natural Rate of Change, and Immigration Rates. The research questions this project aims to tackle is: What are the influences of urbanization on backyard bird species in United States counties? It hypothesizes that more developed counties will record higher Northern Cardinal observations, suggesting a growing environmental responsibility for humans to maintain urban habitats conducive to certain bird species adapting to these environments.

2. Datasets

This project merged two datasets. The first, from Cornell University's Ornithology Lab's eBird project, providing Northern Cardinal observations in U.S. counties. Access was granted upon request, and data from January, April, July, and October 2022 were chosen to manage large file sizes and minimize seasonal biases, totalling 982,667 observations. Data cleaning involved aggregating data at the county level, retaining total observation counts, average observation duration, and effort distance. These variables, showing significant skewness, underwent log-transformation, including the key observation count variable for regression analysis. Appendix includes histograms of these variables before and after their log transformation (*Plots 1-3*). Nine states with minimal Northern Cardinal observations and counties with fewer than 10 observations in 2022 were dropped, resulting in 2,288 aggregated counties in the cleaned dataset.

The second dataset, from the United States Department of Agriculture (USDA), contained county population and urbanization, including data on 3,143 counties. It required minimal cleaning, with no need for row eliminations due to missing values. The cleaning process primarily involved narrowing down to key columns: Rural Urban Continuum Code, Urban Influence, Economic Typology Code, Natural Rate of Change, and Net Immigration Rate.

The Rural-Urban Continuum Codes classify counties on a 1-9 scale, with Code 1 representing highly urbanized areas (over 1 million population) and Code 9 indicating rural areas (under 2,500 population, not metro-adjacent). Urban Influence Codes extend this categorization with a 1-12 range, where 1 is the most urban and 12 the most rural. More details of both codes will be found in *Table 1* and *Table 2.* Economic Typology Codes identify the primary economy of a county: 0 for non-specialized, 1 for Farm, 2 for Mining, 3 for Manufacturing, 4 for Federal/State Government, and 5 for Recreation Dependent. The net immigration rate measures population movement in and out of counties, while the natural rate of change measures the rate of births versus deaths. These rates, showing normal distribution, along with categorical development codes being converted to dummy variables, were used in regression models analysing Northern Cardinal observation counts. After merging the datasets, the study included 2,190 observations. The appendix provides detailed code descriptions distribution charts for the categorical county code variables and continuous variables (*Plots 4-8*). Note that plots are located near the end of the appendix, as the regression models are considered of higher importance.

3. Regression Models

There were four sets of regression models run to test the association between urban development and populations on the log of Northern Cardinal observations. As the first three variables were categorical variables, they were run separately in groups of binary variables to avoid multicollinearity and overfitting by too many variables.

3.1 OLS Regression Models of Urban Influence and Rural-Urban Continuum Codes on Log of Observation Count

In this research, two distinct sets of Ordinary Least Squares (OLS) models were executed to examine the association of Rural-Urban Continuum Codes and Urban Influence Codes on the log of Northern Cardinal observations. The selection of the OLS methodology was based on the hypothesized linear association between the urban development status and the log of Northern Cardinal observations. Given that both Rural-Urban Continuum and Urban Influence Codes provide similar insights regarding a county's characteristics, they were analysed independently to enhance the robustness of the findings. In each set of models, the initial code (Code 1 for both sets) was the reference category and excluded from the regression analysis. Subsequently, a series of OLS regressions were conducted for each development code set, progressively incorporating additional development code variables as explanatory factors (Rural-Urban Continuum Codes 2 through 9 and Urban Influence Codes 2 through 12, respectively). To account for potential heteroscedasticity, all models were estimated using heteroscedasticity-consistent standard errors of type HC1. These models can be found in the appendix (*Regression Models 1 & 2*).

The analysis of both model sets yielded similar outcomes. In each case, the comprehensive model, including all codes except the benchmark (code 1), demonstrated superior model efficacy, as evidenced by the highest R-Squared value. These models were selected for the analysis. The R-Squared for the Urban Influence Code OLS regression was determined to be 32.5%, while 34.1% for the Rural-Urban Continuum Code OLS regression. This indicates that approximately 33% of the variation in log observation counts is explainable by each model.

A consistent pattern emerged across both the Rural-Urban Continuum Codes and Urban Influence Codes. All codes exhibited statistically significant correlation coefficients at the 1% significance level. This uniform significance is likely due to the substantial data volume. Notably, in each model, the correlation coefficients increasingly deviated from zero in a negative trajectory. This trend suggests a positive relationship between the urbanization codes and the log of observation counts, which is observed in both the Rural-Urban Continuum Codes (Codes 2-9) and Urban Influence Codes (Codes 2-12). Therefore, the data provides evidence in support of a positive association between both Rural-Urban Continuum Codes and Urban Influence Codes, and the log of Northern Cardinal observations in US counties.

3.2 OLS Regression Models of County Economic Typology on Log Counts of Northern Cardinal Observations

OLS models were executed to examine the association between County Typology Codes and the log of Northern Cardinal observations. The OLS model was selected due to the hypothesized linear association between the urban development status and the log of Northern Cardinal observation frequencies. Economic typology represents the urban development status of a county from a different. In the model, the code 0 (non-specialized counties) was designated as the reference category and excluded from the regression analysis. A series of OLS regressions were conducted for each typology code, progressively incorporating additional code variables as explanatory factors (Code 1: Economically Farming Dependent, etc.). To account for potential heteroscedasticity, all models were estimated using heteroscedasticity-consistent standard errors of type HC1. These models can be found in the appendix (*Regression Models 3*).

The analysis revealed that the regression model with all economic types was the best fit, with an R-Squared explaining 11.8% of the variance in log observation counts. Farming, mining, and manufacturing dependent resulted in statistically significant negative coefficients at the 1% threshold, while government dependent and recreation dependent were not statistically significant. This would suggest that farming, mining, or manufacturing dependent counties are associated with less Northern Cardinal observation counts.

3.3 OLS Regression Models for Natural Change Rate, Net Immigration Rate, Observation Duration, and Effort Distance on Log Observation Counts

The final OLS models examined the associations between population changes and observation methods, with Northern Cardinal observation counts. These models incorporated natural change and net immigration rates per county to explore how population dynamics relate to observation counts. Additionally, the log-transformed variables for duration and distance of observations were included to examine potential associations with observation counts. The chosen model accounted for 20.2% of the variance in observation counts. Both natural change and net migration rates exhibited small, yet significant, positive associations with observation counts at the 1% significance level. For observation methods, the model included variables such as the log duration and log distance of observation. The log duration of observation displayed a positive association at the 1% significance threshold, indicating an association between longer observation times and higher counts. Conversely, the log distance travelled for observation had a negative association at the same significance threshold, indicating an association between greater travel distances for observations and lower counts. These regression models can be found in the appendix (*Regression Models 4*).

4. Generalization and external validity

Predictive models were applied to the OLS models' variables to validate the associations with log Northern Cardinal observation counts. These models further validated the negative association between urban influence, rural-urban continuum codes, and observation counts, suggesting lower counts in more rural counties (*Predictive Models 1 & 2*). Economic typology modelling indicated non-specialized counties and recreation dependent counties have the highest observation counts, with government, mining, and manufacturing dependent counties showing lower counts (*Predictive Model 3*). Predictive models for natural change rate, net immigration rate, observation duration, and effort distance yielded results similar to the OLS models in their association with observation counts, with negligible associations of natural change and net immigration rates, a positive association with longer observation durations, and a negative association with greater observation distances (*Predictive Models 4-7*). These findings suggest that observation counts of backyard bird species are positively related to urbanization levels, with recreational areas possibly observing more species and higher levels of urbanization having higher amounts of observations. Observation duration positively correlates with counts, whereas wider observation effort distances may lead to fewer observations. This underscores the potential influence of observation methods and urban development on bird observation frequencies.

5. Causal Interpretation

The Urban Influence and Rural-Urban Continuum code models indicate a positive association between Northern Cardinal observations and urban development at the 1% significance threshold. This may reflect an attraction of these bird species to urbanized habitats, possibly due to food access and predator protection, despite initial concerns about observer bias in more populated areas. Economic typology models suggest a varying likelihood of observing backyard bird species across different urban habitats. Notably, recreation dependent counties, while not statistically significant, appear to host more of these species, contrasting with lower observation counts in government-dependent counties. This could be attributed to over-urbanization limiting natural habitats, as seen in densely urbanized areas like Washington D.C., signifying the importance of environmental responsibility to preserve spaces for wildlife.

Regarding population dynamics, the small but statistically significant positive correlations of net immigration and natural change rates with bird observations suggest a potential link between higher human populations and more bird sightings. This might be due to easier food access for birds in human-dominated environments and increased likelihood of observation in more populated areas. Observation statistics further support these findings by showing longer observation durations positively correlate with higher bird counts, while greater travel distances during observations negatively impact counts. This aligns with the backyard habitat tendencies of Northern Cardinals, indicating that stationary observation in a single location, like a backyard, is more likely to yield higher observation counts.

6. Conclusion

The results of this study on Northern Cardinals suggests that backyard birds in the U.S. are more likely to have higher frequencies in more urbanized counties that have spaces for birds to coexist amongst humans. These spaces can also be defined as green spaces that are well-maintained. This implies the importance of integrating green spaces into urban development, not only for species such as the Northern Cardinal, but for the overall biodiversity of urban environments. These findings suggest that policymakers and urban planners should prioritize the creation and upkeep of green spaces within urban counties, which include allocating budgets and resources for new parks, while maintaining existing ones. Although the results of this project cannot be generalized across all bird species, it does begin to uncover the results of urbanization on wildlife populations. Some species may adapt, while others may not adapt and suffer greater consequences of the actions of humans. Further research will need to be done on the association between urban development and non-adapted bird species to analyze the impact on non-backyard bird species.

7. Appendix

7.1 Variable Description Tables

Table 1: Rural-Urban Continuum Codes for Counties Descriptions

Code	Description
1	Metro areas with over 1M population
2	Metro areas with 250K-1M population.
3	Metro areas with under 250K population.
4	20K+ urban, metro-adjacent.
5	20K+ urban, non-metro-adjacent.
6	2.5K-20K urban, metro-adjacent.
7	2.5K-20K urban, non-metro-adjacent.
8	Rural or <2.5K urban, metro-adjacent.
9	Rural or <2.5K urban, non-metro-adjacent.

Table 2: Urban Influence Codes for Counties Descriptions

Code	Description
1	Large metro area, 1M+ residents.
2	Small metro area, under 1M residents.
3	Micropolitan, adjacent to large metro.
4	Noncore, adjacent to large metro.
5	Micropolitan, adjacent to small metro.
6	Noncore, adjacent to small metro, town ≥2.5K residents.
7	Noncore, adjacent to small metro, no town ≥2.5K.
8	Micropolitan, not metro-adjacent.
9	Noncore, adjacent to micro area, town ≥2.5K.
10	Noncore, adjacent to micro area, no town ≥2.5K.
11	Noncore, not adjacent to metro/micro, town ≥2.5K.
12	Noncore, not adjacent to metro/micro, no town ≥2.5K.

7.2 Regression Models

Regression Model 1: OLS Regression Models of Urban Influence Codes on Log of Observation Count

											Dependent variable: ln_OBSERVATION_COUNT
	Reg 1	Reg 2	Reg 3	Reg 4	Reg 5	Reg 6	Reg 7	Reg 8	Reg 9	Reg 10	Reg 11
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Urban Influence Code 2	1.089***	1.140***	1.112***	1.159***	1.008***	0.918***	0.706***	0.405***	0.082	-0.207*	-0.638***
	(0.081)	(0.082)	(0.083)	(0.086)	(0.093)	(0.096)	(0.105)	(0.112)	(0.116)	(0.117)	(0.110)
Urban Influence Code 3		0.725***	0.698***	0.745***	0.593***	0.503***	0.292**	-0.009	-0.332**	-0.621***	-1.052***
		(0.132)	(0.133)	(0.135)	(0.139)	(0.141)	(0.148)	(0.153)	(0.155)	(0.157)	(0.151)
Urban Influence Code 4			-0.309**	-0.262**	-0.414***	-0.503***	-0.715***	-1.016***	-1.339***	-1.628***	-2.059***
			(0.128)	(0.130)	(0.135)	(0.137)	(0.143)	(0.149)	(0.151)	(0.152)	(0.147)
Urban Influence Code 5				0.348***	0.196*	0.106	-0.106	-0.407***	-0.730***	-1.019***	-1.450***
				(0.108)	(0.114)	(0.116)	(0.124)	(0.130)	(0.133)	(0.134)	(0.128)
Urban Influence Code 6					-0.663***	-0.753***	-0.965***	-1.266***	-1.589***	-1.878***	-2.309***
					(0.101)	(0.104)	(0.112)	(0.119)	(0.122)	(0.123)	(0.117)
Urban Influence Code 7						-0.914***	-1.125***	-1.426***	-1.750***	-2.038***	-2.469***
						(0.153)	(0.159)	(0.164)	(0.166)	(0.167)	(0.163)
Urban Influence Code 8							-0.965***	-1.266***	-1.590***	-1.878***	-2.309***
							(0.130)	(0.136)	(0.139)	(0.140)	(0.135)
Urban Influence Code 9								-1.631***	-1.955***	-2.244***	-2.675***
								(0.144)	(0.147)	(0.148)	(0.142)
Urban Influence Code 10									-2.209***	-2.498***	-2.929***
									(0.160)	(0.161)	(0.157)
Urban Influence Code 11										-2.358***	-2.789***
										(0.190)	(0.186)
Urban Influence Code 12											-3.144***
											(0.181)
Constant	5.422***	5.371***	5.398***	5.351***	5.503***	5.593***	5.804***	6.105***	6.429***	6.718***	7.149***
	(0.042)	(0.044)	(0.047)	(0.052)	(0.063)	(0.067)	(0.079)	(0.089)	(0.093)	(0.095)	(0.086)
Observations	2190	2190	2190	2190	2190	2190	2190	2190	2190	2190	2190
R^2	0.074	0.082	0.084	0.087	0.101	0.111	0.130	0.169	0.216	0.256	0.325
Adjusted R^2	0.074	0.082	0.083	0.085	0.099	0.108	0.128	0.166	0.213	0.253	0.321
Residual Std. Error	1.688 (df=2188)	1.680 (df=2187)	1.679 (df=2186)	1.677 (df=2185)	1.665 (df=2184)	1.656 (df=2183)	1.638 (df=2182)	1.602 (df=2181)	1.555 (df=2180)	1.516 (df=2179)	1.445 (df=2178)
F Statistic	181.874*** (df=1; 2188)	101.647*** (df=2; 2187)	71.985*** (df=3; 2186)	54.725*** (df=4; 2185)	60.415*** (df=5; 2184)	57.775*** (df=6; 2183)	58.291*** (df=7; 2182)	66.057*** (df=8; 2181)	75.328*** (df=9; 2180)	79.048*** (df=10; 2179)	98.223*** (df=11; 2178)
Note:											$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Regression Model 2: OLS Regression Models of Rural-Urban Continuum Codes on Log of Observation Count

								Dependent variable: ln_OBSERVATION_COUNT
	Reg 12	Reg 13	Reg 14	Reg 15	Reg 16	Reg 17	Reg 18	Reg 19
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Rural-Urban Continuum Code 2	1.130***	1.257***	1.361***	1.363***	1.218***	0.847***	0.519***	-0.469***
	(0.106)	(0.108)	(0.109)	(0.110)	(0.118)	(0.127)	(0.134)	(0.131)
Rural-Urban Continuum Code 3		0.898***	1.002***	1.004***	0.859***	0.488***	0.160	-0.829***
		(0.102)	(0.104)	(0.104)	(0.113)	(0.122)	(0.130)	(0.127)
Rural-Urban Continuum Code 4			0.958***	0.961***	0.815***	0.444***	0.116	-0.872***
			(0.098)	(0.099)	(0.108)	(0.117)	(0.125)	(0.122)
Rural-Urban Continuum Code 5				0.051	-0.095	-0.465**	-0.793***	-1.782***
				(0.174)	(0.179)	(0.185)	(0.191)	(0.189)
Rural-Urban Continuum Code 6					-0.419***	-0.789***	-1.117***	-2.106***
					(0.088)	(0.100)	(0.109)	(0.105)
Rural-Urban Continuum Code 7						-1.295***	-1.623***	-2.611***
						(0.112)	(0.121)	(0.118)
Rural-Urban Continuum Code 8							-1.528***	-2.517***
							(0.143)	(0.140)
Rural-Urban Continuum Code 9								-3.025***
								(0.125)
Constant	5.549***	5.422***	5.318***	5.316***	5.462***	5.832***	6.160***	7.149***
	(0.039)	(0.042)	(0.046)	(0.047)	(0.064)	(0.079)	(0.091)	(0.086)
Observations	2190	2190	2190	2190	2190	2190	2190	2190
R^2	0.049	0.077	0.098	0.098	0.106	0.152	0.190	0.341
Adjusted R^2	0.049	0.076	0.097	0.097	0.104	0.150	0.187	0.339
Residual Std. Error	1.710 (df=2188)	1.685 (df=2187)	1.666 (df=2186)	1.667 (df=2185)	1.659 (df=2184)	1.617 (df=2183)	1.581 (df=2182)	1.426 (df=2181)
F Statistic	112.799*** (df=1; 2188)	92.499*** (df=2; 2187)	83.933*** (df=3; 2186)	63.098*** (df=4; 2185)	64.214*** (df=5; 2184)	85.025*** (df=6; 2183)	87.607*** (df=7; 2182)	145.267*** (df=8; 2181)
Note:								$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Regression Model 3: OLS Regression Models of Economic Typology on Log of Observation Count

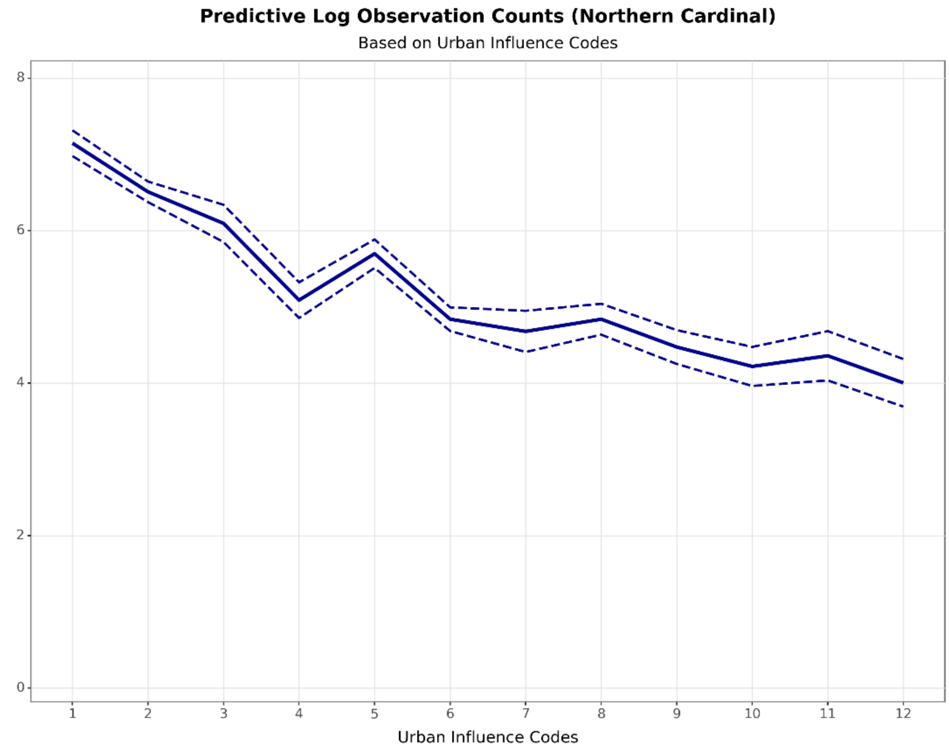
	Dependent variable: ln_OBSERVATION_COUNT				
	Reg 20	Reg 21	Reg 22	Reg 23	Reg 24
	(1)	(2)	(3)	(4)	(5)
Economically Farming Dependent	-1.662***	-1.729***	-1.905***	-1.894***	-1.864***
	(0.106)	(0.107)	(0.109)	(0.112)	(0.115)
Economically Mining Dependent		-1.074***	-1.250***	-1.239***	-1.209***
		(0.141)	(0.143)	(0.145)	(0.147)
Economically Manufacturing Dependent			-0.754***	-0.742***	-0.713***
			(0.083)	(0.086)	(0.090)
Economically Federal/State Government Dependent				0.058	0.088
				(0.114)	(0.117)
Economically Recreation Dependent					0.166
					(0.127)
Constant	5.843***	5.910***	6.086***	6.075***	6.045***
	(0.038)	(0.039)	(0.046)	(0.052)	(0.058)
Observations	2190	2190	2190	2190	2190
R^2	0.068	0.089	0.117	0.117	0.118
Adjusted R^2	0.068	0.088	0.116	0.116	0.116
Residual Std. Error	1.693 (df=2188)	1.675 (df=2187)	1.649 (df=2186)	1.649 (df=2185)	1.649 (df=2184)
F Statistic	243.984*** (df=1; 2188)	148.919*** (df=2; 2187)	120.819*** (df=3; 2186)	90.798*** (df=4; 2185)	73.382*** (df=5; 2184)
Note:					*p<0.1; **p<0.05; ***p<0.01

Regression Model 4: OLS Regression Models for Natural Change Rate, Net Immigration Rate, Observation Duration, and Effort Distance on Log Observation Counts

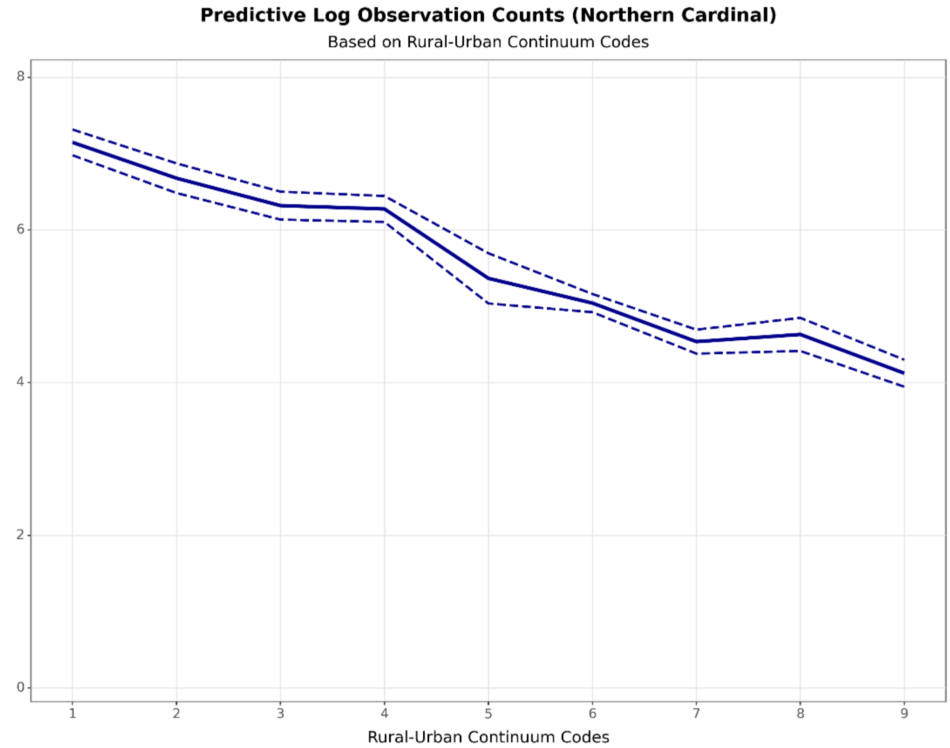
	Dependent variable: ln_OBSERVATION_COUNT				
	Reg 25	Reg 26	Reg 27	Reg 28	Reg 29
	(1)	(2)	(3)	(4)	(5)
Natural Change Rate 2022	0.135***	0.141***		0.137***	0.132***
	(0.009)	(0.009)		(0.009)	(0.009)
Net Migration Rate 2022		0.023***		0.021***	0.021***
		(0.003)		(0.003)	(0.003)
Log Duration of Observation (minutes)			1.205***	0.859***	1.062***
			(0.091)	(0.079)	(0.086)
Log Distance Traveled for Observation (km)			-0.571***		-0.455***
			(0.072)		(0.069)
Constant	6.137***	6.014***	1.282***	2.407***	2.050***
	(0.050)	(0.053)	(0.355)	(0.331)	(0.341)
Observations	2182	2182	2186	2182	2178
R^2	0.107	0.136	0.088	0.183	0.202
Adjusted R^2	0.107	0.135	0.087	0.182	0.201
Residual Std. Error	1.653 (df=2180)	1.627 (df=2179)	1.674 (df=2183)	1.582 (df=2178)	1.562 (df=2173)
F Statistic	220.247*** (df=1; 2180)	165.735*** (df=2; 2179)	94.159*** (df=2; 2183)	151.045*** (df=3; 2178)	125.233*** (df=4; 2173)
Note:					*p<0.1; **p<0.05; ***p<0.01

7.3 Predictive Models

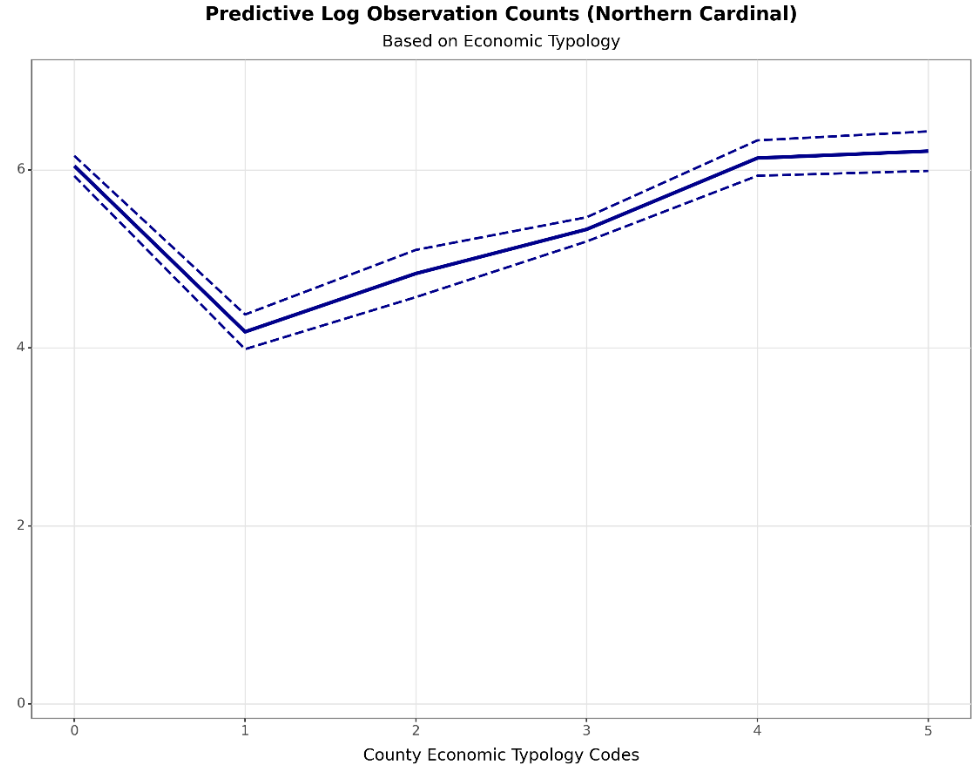
Predictive Model 1: Urban Influence Codes on Log Observation Counts



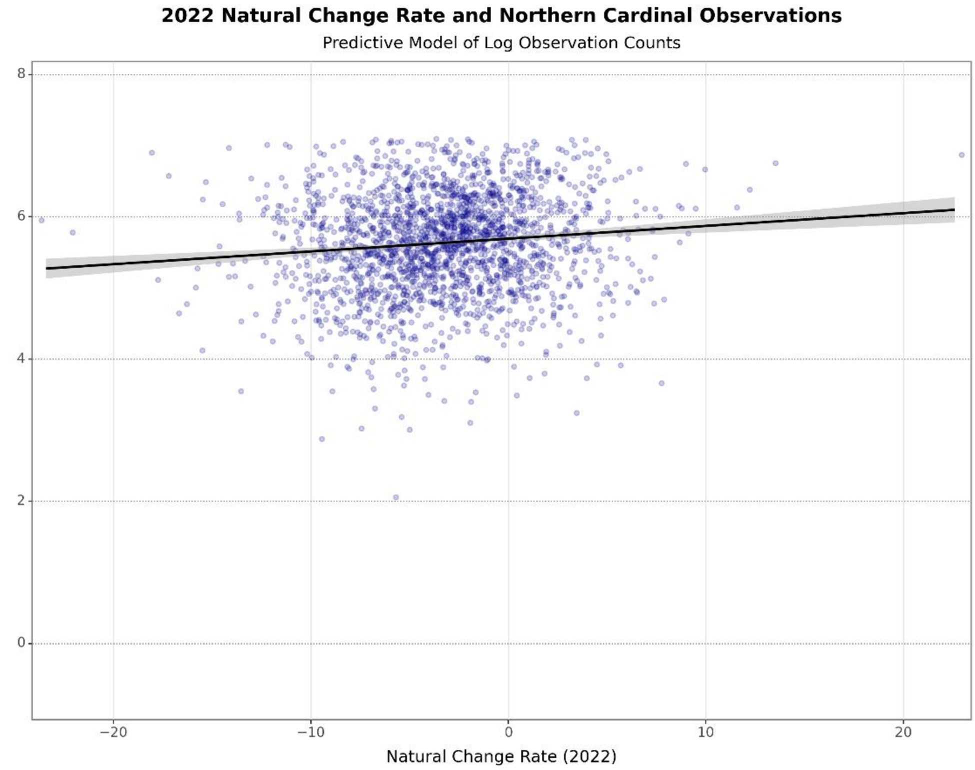
Predictive Model 2: Rural-Urban Continuum Codes on Log Observation Counts



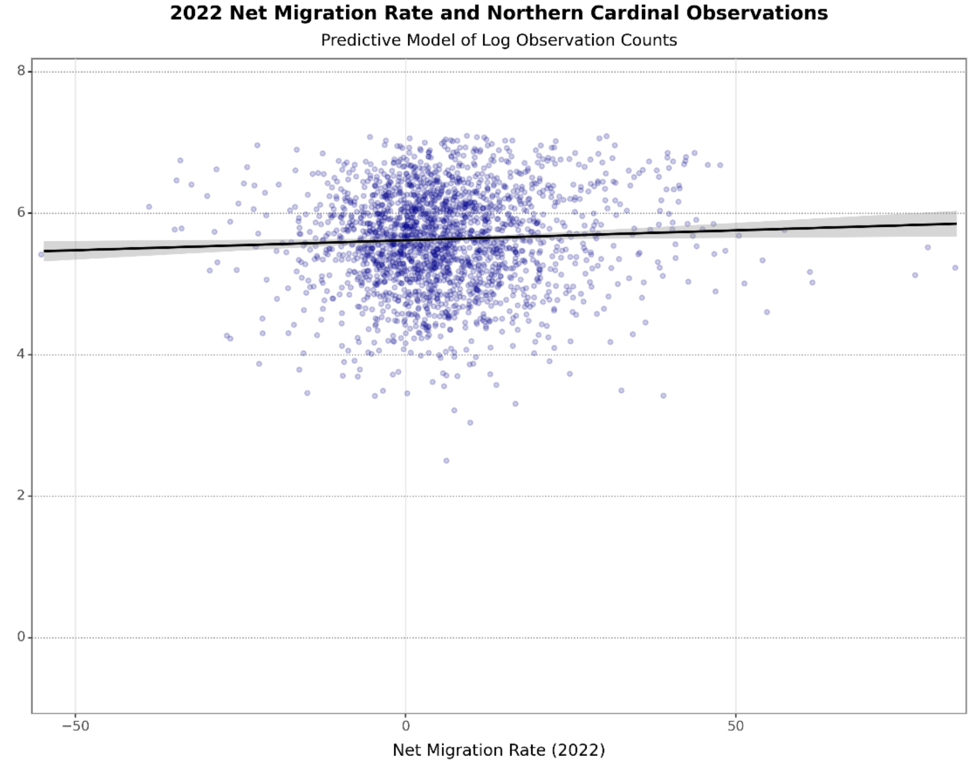
Predictive Model 3: County Economic Typology Codes on Log Observation Counts



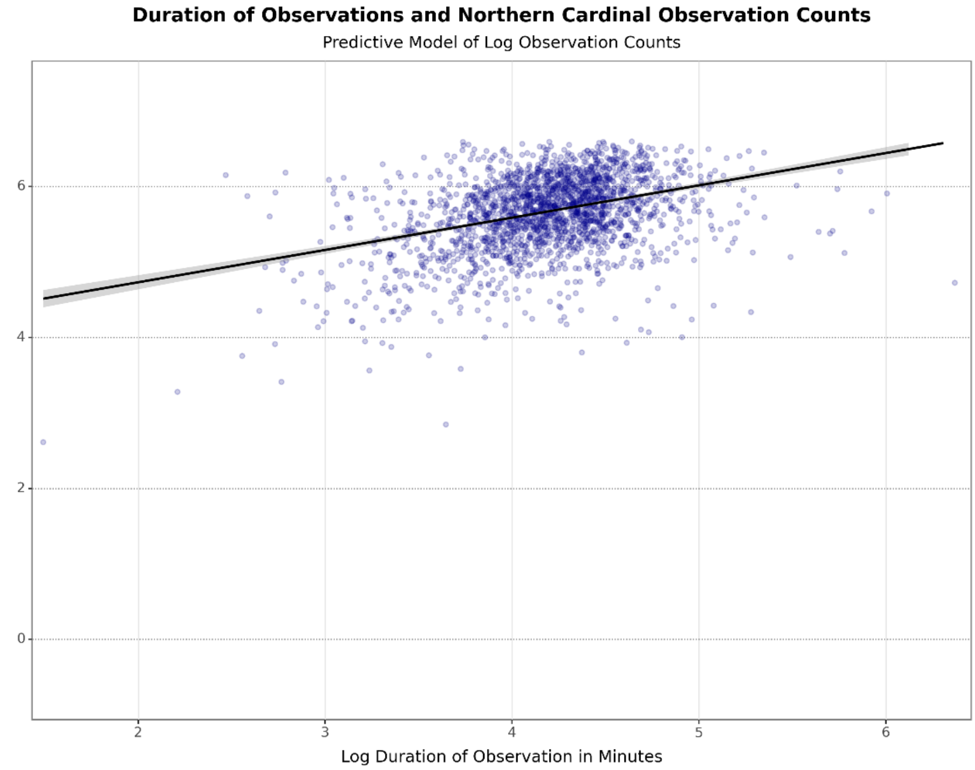
Predictive Model 4: Natural Change Rate on Log Observation Counts



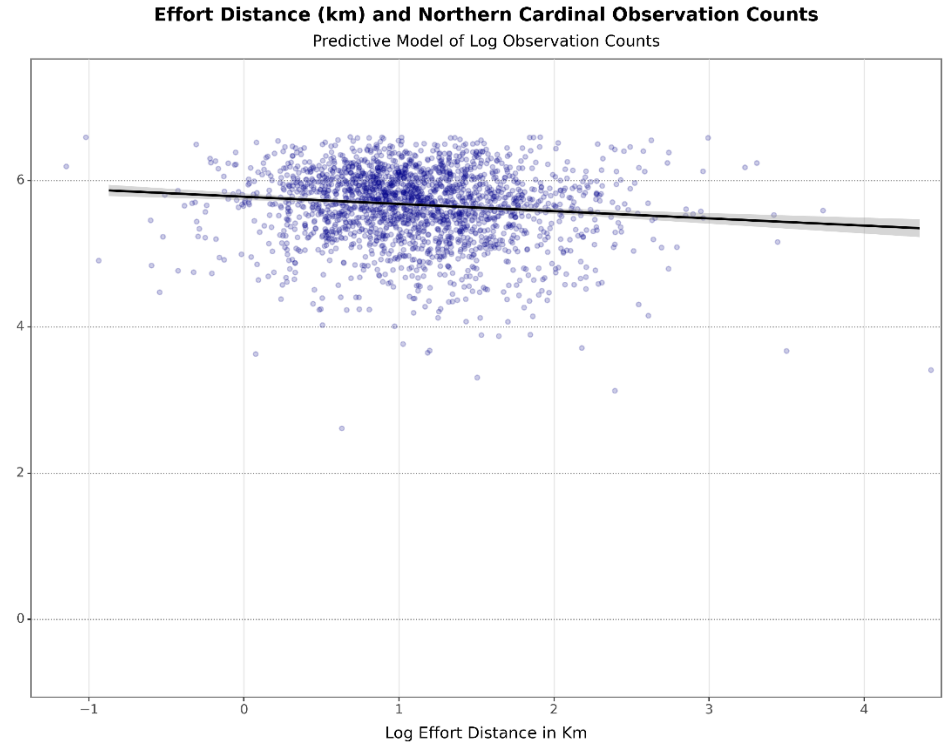
Predictive Model 5: Net Immigration Rates on Log Observation Counts



Predictive Model 6: Duration of Observations on Log Observation Counts



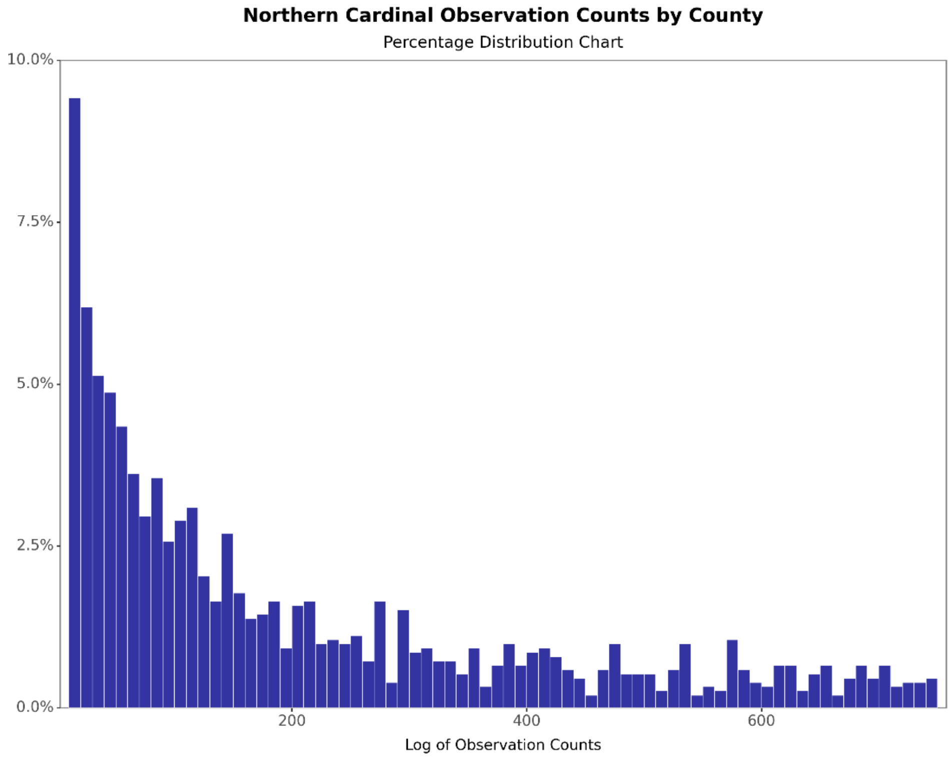
Predictive Model 7: Effort Distance on Log Observation Counts



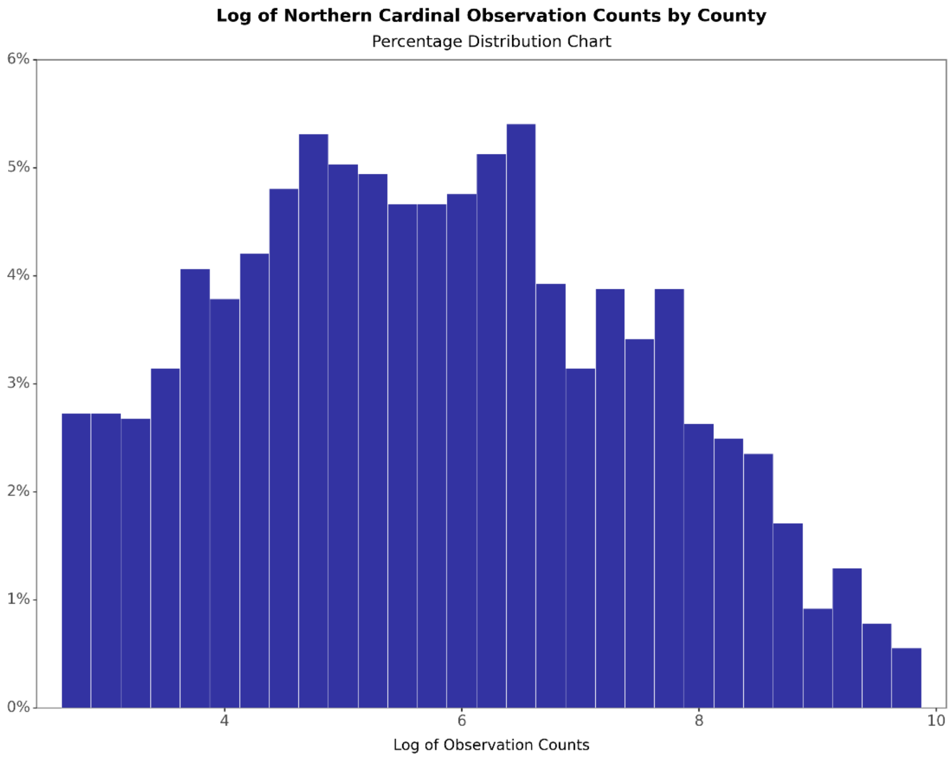
7.4 Distribution Charts

Plot 1(a & b): Distribution Plots of Observation Counts pre/post log-transformation.

a)

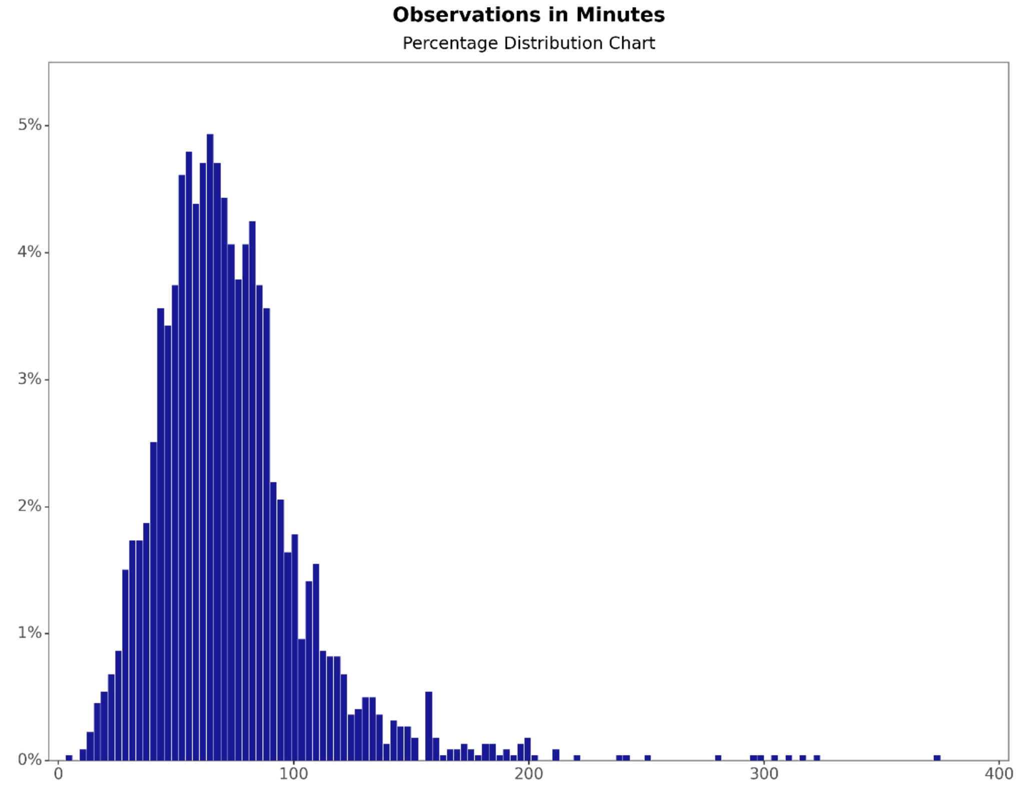


b)

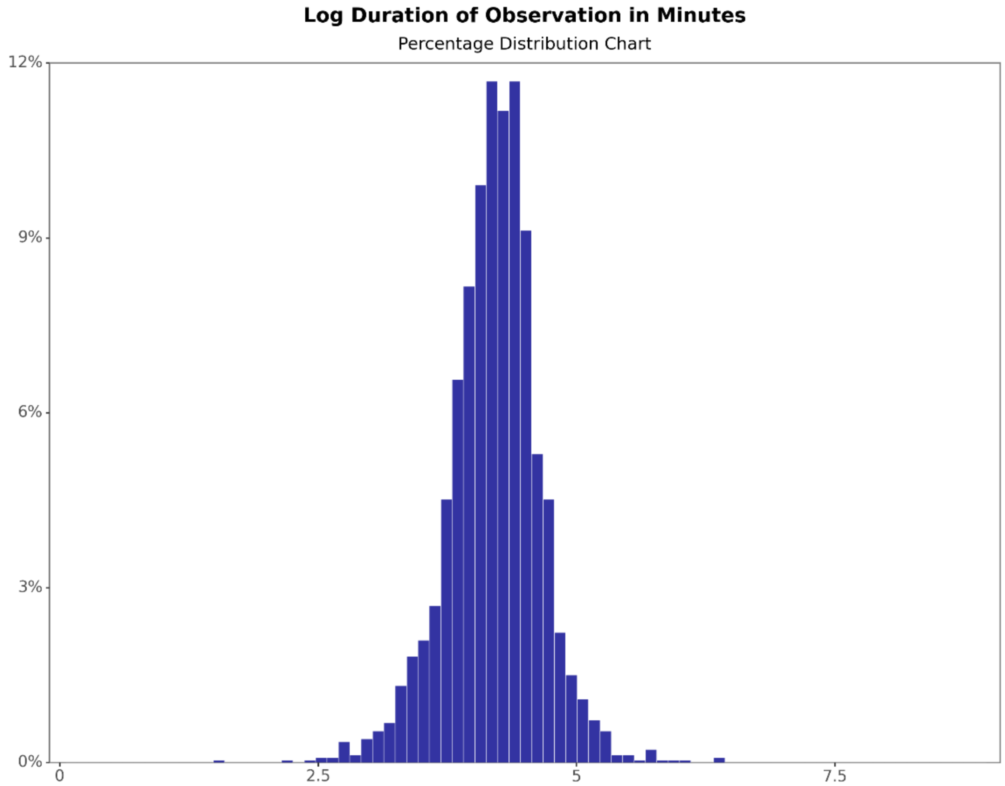


Plot 2 (a & b) Distribution Plots of Observation Duration pre/post log-transformation

a)

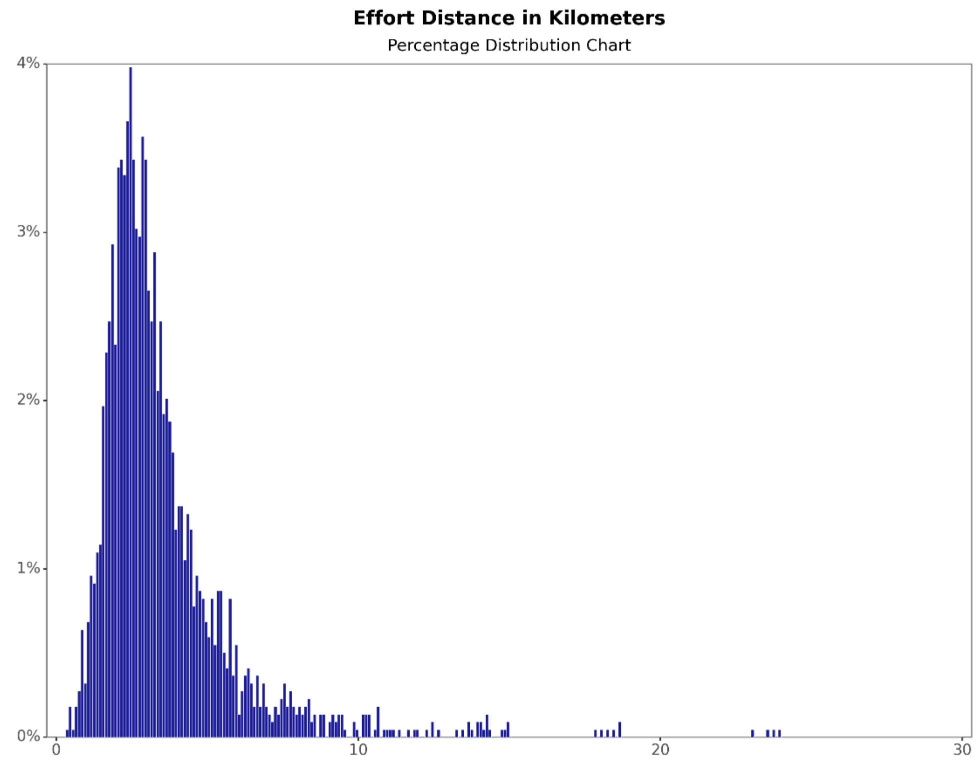


b)

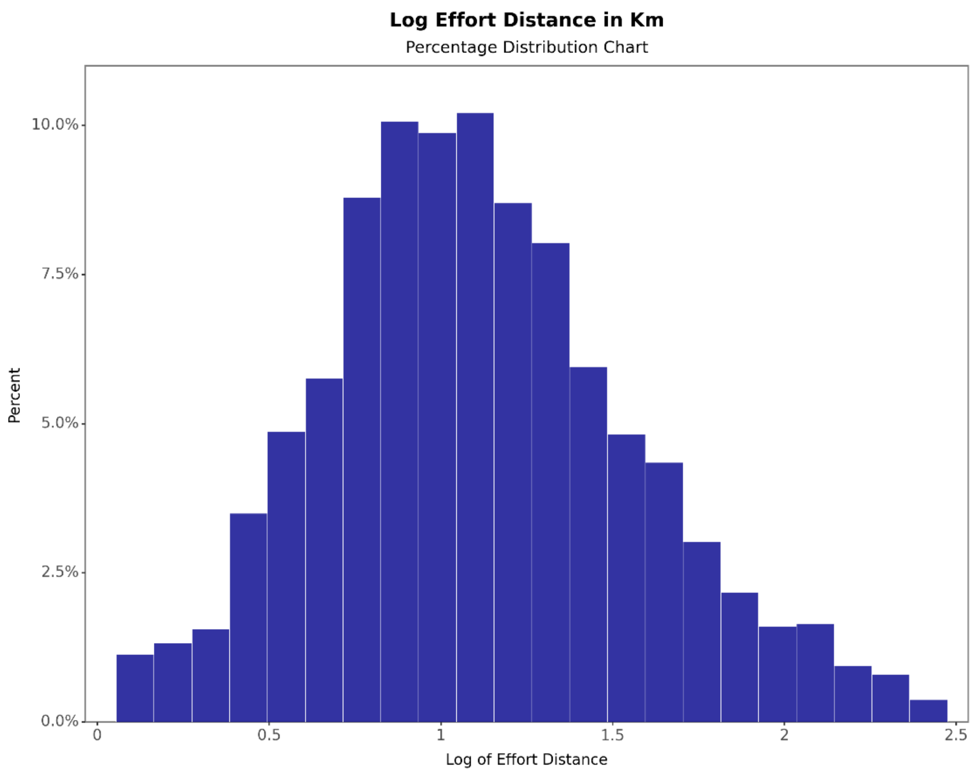


Plot 3(a & b): Distribution Plots of Observation Distance pre/post log-transformation

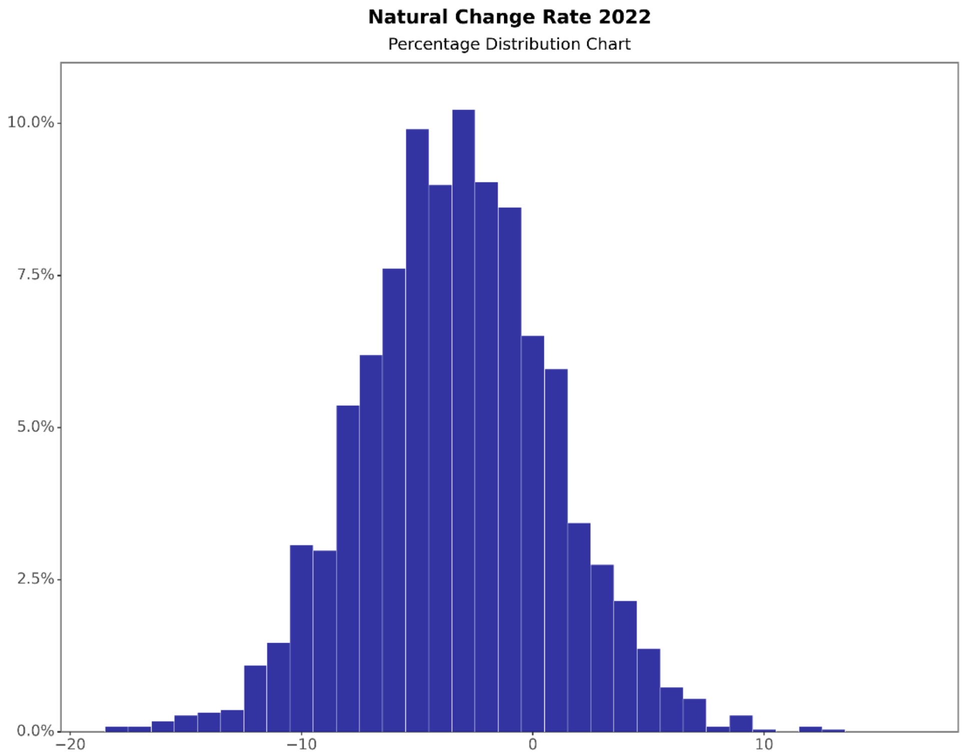
a)



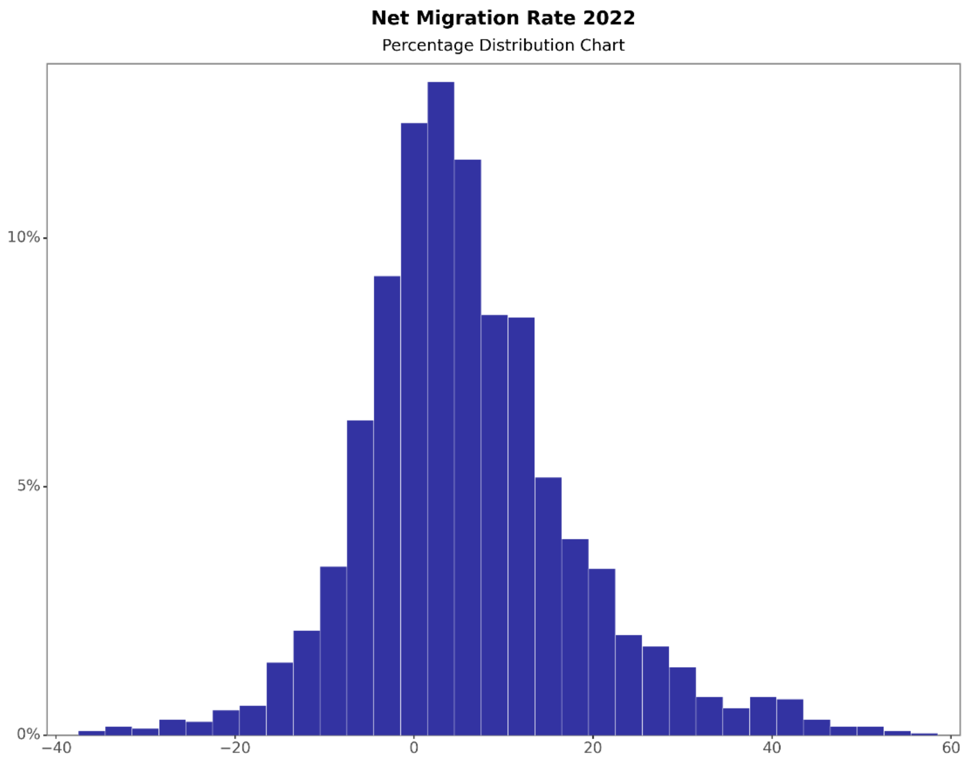
b)



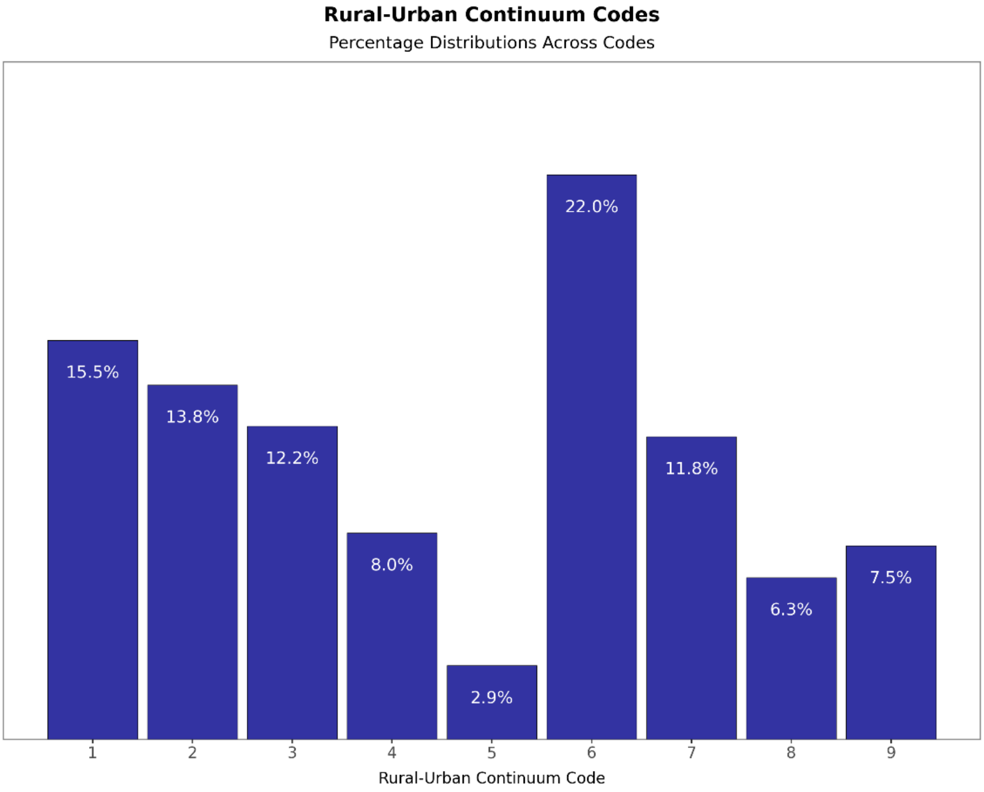
Plot 4: Distribution of Natural Change Rate in 2022



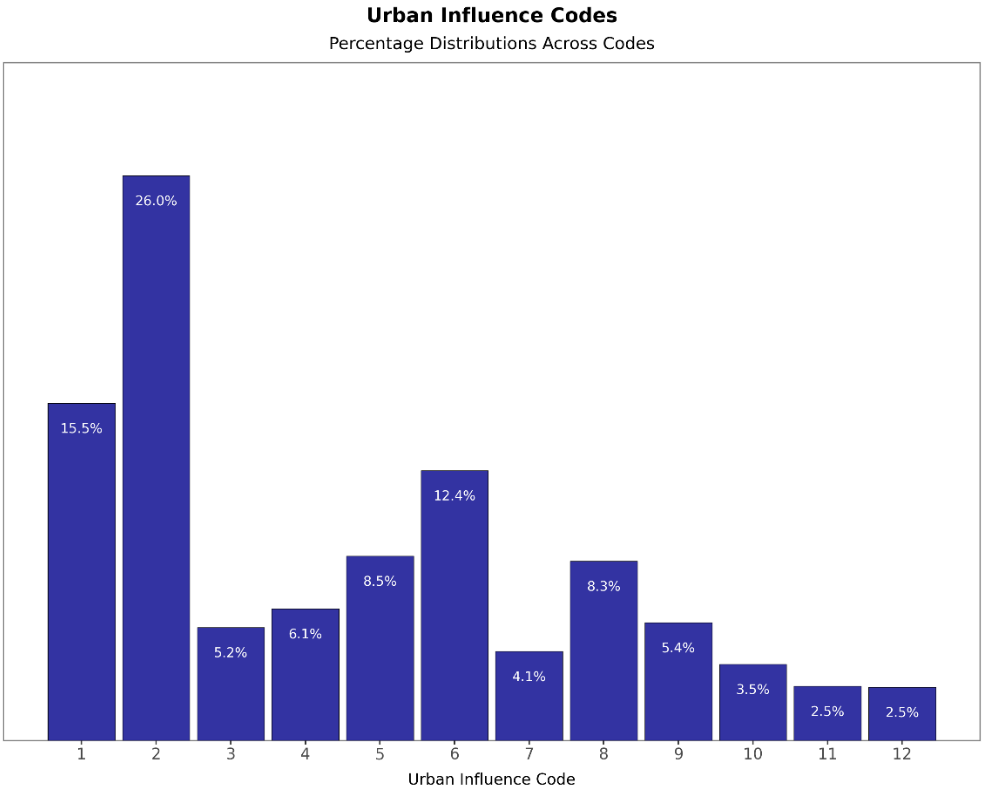
Plot 5: Distribution of Net Migration Rate 2022



Plot 6: Distribution of Rural-Urban Continuum Codes



Plot 7: Distribution of Urban Influence Codes



Plot 8: Distribution of County Economic Typology in 2015