Analysis of Predictive Models Applied to Public safety: a case study

Alexsandra Lima[1], *Raydonal Ospina*[1,2], *Cristiano Ferraz*[2]
[1] Federal University of Pernambuco
[2] Federal University of Bahia

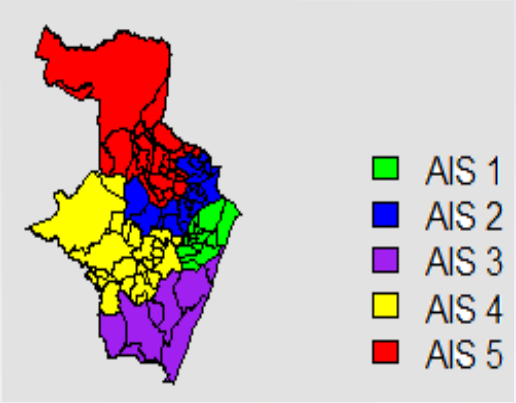
Abbreviated abstract: Using data from 2009 to 2011 on the occurrence of crimes, spatial and socioeconomic information from the neighborhoods of the city of Recife involving classifications of the Integrated Security Areas, we evaluated four predictive models, are they: linear regression model, generalized linear regression model, regression model based on principal components and generalized linear regression model based on information similarities. The modeling results were evaluated using different performance metrics. The results point to an increase in the predictive performance of models that use information from principal components and similarity predictors.

4th Conference on
Statistics and
Data Science
Salvador, Brazil (online)
December 1-3, 2022

Problem, Data, Metrics

• Objective: To analyze the performance of prediction models, using criminal data and socioeconomic and spatial variables in Recife, including integrated security areas;

• Integrated Security Areas



• Data:

➢ occurrence of crimes: CVLI, theft, rape, aggression and drug trafficking: private domain base of the Secretary of Social Defense of Pernambuco;

➢ Latitude, longitude, distances to the airport, beach, park and subway, provided by Recife City Hall;

➢ Socioeconomic characteristics of individuals and households, collected from the 2010 Demographic Census;

• Metrics:

➢ Mean Squared Error (MSE)

➢ Root Mean Squared Error (RMSE)

➢ Mean Absolute Percentage Error (MAPE)

➢ Pseudo R^2

➢ AIC

➢ BIC

Methods

• Models:

➢ OLS: verifies the existence of a relationship between a dependent variable and one or more independent variables; Suppose there is a linear relationship between the response variable, over a set of explanatory variables;

➢ PCR: dimensionality reduction and multicollinearity removal. 1. Decompose the independent variables matrix, determining the linear combination of the variables, using the Principal Components Analysis method 2. Use regression to establish the relationship between the dependent variables and the new matrix of uncorrelated variables (main components). The first three CP's were responsible for more than 89% of the total variation on the CVLI rate in the city of Recife;

➢ MLG: it makes it possible to use other distributions for the errors and a link function relating the mean of the response variable to the linear combination of the explanatory variables; It assumes that the response variable has a distribution belonging to an exponential family distribution;

➢ DB-MLG: forecasting tool that can also be applied to qualitative explanatory variables, maintaining compatibility with the OLS method; The method is based on distance analysis and consists of two steps: 1. Distances are calculated from actually observed predictors by means of an adequate dissimilarity function. 2. from the distances between the observations we obtain latent variables which, in turn, are the model's regressors;

Results and Conclusions

Metrics Results

			MSE	$RMSE$	$MAPE$	p_R^2	AIC	BIC
OLS	data	*training*	0.011	0.107	0.148	0.625	11.987	90.209
		test	0.271	0.520	1.700	0.029	71.925	134.538
	data + AIS	*training*	0.004	0.067	0.065	0.719	64.458	46.435
		test	0.728	0.853	0.413	0.020	53.875	121.818
MLG	data	*training*	0.128	0.358	0.074	0.672	40.949	61.246
		test	3.071	1.752	1.707	0.034	139.525	202.138
	data + AIS	*treino*	0.311	0.558	0.098	0.513	22.340	133.234
		teste	6.186	2.487	0.559	0.015	172.394	240.336
PCR	data	*training*	0.115	0.339	0.232	0.405	53.908	64.780
		test	0.130	0.361	0.237	0.265	52.495	59.156
	data + AIS	*training*	0.119	0.345	0.207	0.272	56.360	67.232
		test	0.143	0.378	0.215	0.396	53.551	60.212
$DB-GLM$	data	*training*	0.010	0.102	0.138	0.618	13.105	91.264
		test	0.089	0.299	0.243	0.315	22.279	20.947
	data + AIS	*training*	0.012	0.111	0.030	0.724	2.173	2.197
		test	0.112	0.335	0.209	0.482	17.570	16.238

- Improved accuracy in DB-GLM and PCR models;
- Include new classifications of integrated security areas;
- Test other models, e.g., GAMLSS;