**MINING DATA ON GUN VIOLENCE WITH RECONSTRUCTABILITY ANALYSIS**

Taban Yolo

Business Intelligence and Analytics Program

Portland State University

Portland, OR

[tabanyolo@yahoo.com](mailto:tabanyolo@yahoo.com)

***Abstract****—Reconstructability Analysis (RA) is a data mining method that searches for relations in data, especially non-linear and higher order relations. This study shows that RA can provide useful predictions gun violence in different regions of the country.*

***Keywords****—Reconstructability Analysis, Occam, predictive analytics, healthcare, risk prediction, total knee replacement.*

**INTRODUCTION**

Every day, 100 Americans are killed with guns and hundreds more are shot and injured. The effects of gun violence extend far beyond these casualties—gun violence shapes the lives of millions of Americans who witness it, know someone who was shot, or live in fear of the next shooting. In order to illustrate the magnitude of everyday gun violence, Everytown has gathered the most comprehensive, publicly available data. Still, significant data gaps remain—a result of underfunded, incomplete data collection at the state and federal level. Filling these gaps is necessary to truly understand the full impact of gun violence in the United States.

The purpose of Reconstructability Analysis is to provide systems investigators with useful methodological tools for dealing with the various questions regarding the relationship between dependent variables and their various independent variables. The term “system” is viewed in reconstructability analysis as a characterization of certain type of fuzzy measures by which the constraint among variables of interest is described.

Two complementary problems are involved in reconstructability analysis: (i) given an overall system, determine which sets of subsystems can be used to reconstruct it adequately (reconstruction problem); (ii) given a set of systems characterized by the same kind of measure, derive from it as much knowledge as possible regarding the unknown overall system (identification problem).

This paper reports the use of RA applied to the secondary analysis of Gun Violence data. The aim is to discover unexpected relationships in the data.

**DATA**

The data analyzed here is from the Gun Violence Archive which is a not for profit corporation formed in 2013 to provide free online public access to accurate information about gun-related violence in the United States. GVA will collect and check for accuracy, comprehensive information about gun-related violence in the U.S. and then post and disseminate it online.

There are 14 variables in this study with 12 being the independent variables and 1 being the dependent variable. The independent variables are labelled as A, B, C, D, E , F, G, H, I, J, K, L where A = Incident ID of the crime, B = Date which the crime occurred (The format was MM/DD/YYYY), C = The region in which the crime occurred (It should be noted in order to refine the results to carry out RA, the states in the country were split into three regions namely West & Midwest(2), South(3) and Northeast(1)), D = Number of males that were involved in the incident, E = Number of females involved in the incident, F = Number of guns involved during the incident, G = Number of guns involved which the status was unknown, H = Number of guns involved which the status was stolen, I = Number of guns involved which the status was not stolen, J = Number of Adults involved in the incident, K = Number of teen involved in the incident, L = Number of children involved in the incident, M = Number of people injured during the incident. The independent variable is Z, where Z = Number of people killed during the incident.

**BINNING THE DATA**

The Binning Process included setting 1 Dependent variable Z, and 13 independent variables as A to M. In terms of Cardinality, the biggest cardinality set was 8 on IV’s: F and G. The rest of the variables including the DV were set at a cardinality of 4. Also, it should be noted for Binning purposes, IV’s: A and C were set to a bin type of “Do not Bin” simply for the fact that C had already been binned into 4 categories mentioned in the previous section and A showed little or no correlation on the data in general.

Moving on, after Binning the data through the Bin Software, the output file created gives us the different states for each of the variables. Since we decided not to bin A and B, they will be excluded from this analyses below.



To begin with, IV B (Date) was binned into 4 bins with Bin 1 being from 1/1/13 to 5/1/14, Bin 2 being from 5/2/14 to 8/1/15, Bin 3 being from 8/2/15 to 12/1/16 and Bin 4 being from 12/2/16 to 3/1/18. Next, IV D (Number of males involved) was binned into 3 bins with Bin 1 being o males involved, Bin 2 being 1 male involved and Bin 3 being more than 1 male involved. Next, IV E (Number of females involved in the incident) was binned into 2 bins with Bin 1 being 0 females involved and Bin 2 being 1 or more females involved. Next, IV F (Number of guns involved) was binned into 2 bins with Bin 1 being 0 guns involved and Bin 2 being 1 or more guns involved. Next, IV G (From the total number of guns, how many guns were Unknown) was binned into 2 bins with Bin 1 being 0 unknown guns involved and Bin 2 being 1 or more unknown guns involved.

Furthermore, IV H (From the total number of guns, how many guns were Stolen) was binned into 2 bins with Bin 1 being 0 guns stolen involved and Bin 2 being 1 or more guns stolen involved. Next, IV I (From the total number of guns, how many guns were Not Stolen) was binned into 2 bins with Bin 1 being 0 guns not stolen involved and Bin 2 being 1 or more guns not stolen involved. Next, IV J (Number of Adults involved) was binned into 4 bins with Bin 1 being 0 Adults involved and Bin 2 being 1 Adult involved, Bin 3 being 2 Adults involved and Bin 4 being 3 or more Adults involved. Next, IV K (Number of Teens involved) was binned into 2 bins with Bin 1 being 0 teens involved and Bin 2 being 1 or more teens involved. Next, IV L (Number of Children involved) was binned into 2 bins with Bin 1 being 0 children involved and Bin 2 being 1 or more children involved. Next, IV M (Number of people injured during the incident) was binned into 2 bins with Bin 1 being 0 people injured during the incident and Bin 2 being more than 1 person injured during incident. Finally, DV Z (Number of people killed during the incident) was binned into 2 bins with Bin 1 being 0 people killed during the incident and Bin 2 being more than 1 person killed during incident.

**RESULTS**

This paper reports results of a Loopless Search, Loopless fit, All Search, and an All fit for the DV (Number of People killed).

1. **Loopless Search (Directed System Output)**

Table I shows a summary of the results obtained from the Loopless Search performed. Looking at the summary of the search, the dBIC model predicts that D (Number of males involved), J (Number of Adults) and N (Number of people injured during the incident) best predict the Independent variable Z (Number of people dead during an incident). The %C(Data) which shows the performance of each model on the given data is 76.03%. This is a decent number for the model. We can also note the %cover which gives a percentage of data covered, and at 100%, we can be confident with this number. It should be noted that the dBIC model is less complex than other models and tends to under fit making it less predictive than what is statistically warranted.

However, one interesting thing to note is that the dAIC model predicts that D (Number of males involved), J (Number of Adults), G (Number of Unknown guns involved) and N (Number of people injured during the incident) best predict the Independent variable Z (Number of people dead during an incident). Looking at the %cover of the model at 97.916%, we can see it is less than that of the dBIC model. It should also be noted that dAIC models are more complex than other models and tends to over fit making them more predictive than what is statistically warranted.





Table I: Loopless Search Output

Therefore, taking all this into consideration, the Loopless Search gives us a best model of IV:DJMZ (using the dBIC model). We will further do more exploratory analysis by fitting this Loopless model and see what states of the variables warrant more research.

1. **Fit of the Loopless Search**

Table II shows a summary of the results obtained from the Loopless fit performed. We can look at the Z=1 and Z = 2 state percentages and do further analysis of the data set. For each of the states we are going to look at the maximum percentage achieved for one state and simultaneously looking at the minimum state achieved for the opposite state. After determining the states, we will look at the p-margins and states of the dependent variables to see any relations created.

For the maximum percentage of Z = 1 (98.137%) we have the corresponding minimum percentage of Z = 2 (1.863%). The dependent variables D, J and M are all in state 1 which means that there is a 98.137% chance of 0 people being killed if 0 males, 0 adults and 0 people are injured. This information is not useful since it is obvious if the IV’s are held at 0, we would automatically assume nobody would be killed.

If we further explore, the reverse where the maximum percentage of Z = 2 (46.652%) we have a corresponding minimum percentage of Z = 1 (53.348%). The dependent variable D is in state 3, J is in state 3 and M is in state 1 which means that there is a 46.652% chance of more than 1 person being killed if 2 adults, at least one of the 2 adults is a male, 2 adults and 0 people are injured during the incident. More so, looking at the p-margin (0), the model is statistically significant since the value is smaller than alpha of 0.05.



Table II: Loopless Fit Output

We can do further explanatory analysis by performing a Search All output and analyze the results. At this point, the Loopless fit has some interesting information, but we can need to evaluate further models to see if we get different results.

1. **Search – All (Directed System Output)**

For the Search All output results, we have to make adjustments before we get an output file. Since the Search All method gets the best combinations of all variables at each level, we can increase both our Search width and Search levels (We do these to see how far in complexity the models can get). Again, the goal is not to get the most complex model but instead to see if the complex model can be used compared to the model selected from the Loopless Search.

To begin with, we will perform an All Search with a Search Width of 3 and a Search Level of 7. Looking at Table III, we can see that the best dBIC model according to the Search is IV:CZ:DZ:EZ:GZ:JMZ.

The model has a dDF (Measures the level of complexity of a model) of 13 and dDH (Reduction in the level of uncertainty of the DV, thus the higher the better) of 16.1716%. Compared to the Loopless Search model that had dDF of 23 and dDH of 15.938%, we can therefore conclude that this model is less complex and reduces the level of uncertainty of the DV better than the Loopless model. It should also be noted that the %cover drops from 100% to 84%.

Given this, the model is not that too complex however we can notice that the search used all 7 levels and therefore we can assume if we increase the Search levels, maybe we can get a better model.

Therefore, we are going to increase the Search level to 15 with a Search Width of 3 and try and interpret the results. If we get a better complex model than the previous model, then we should use it.





Table III: Search All Output (Width = 3, Levels = 7)

Looking at Table IV, the we can see that the best dBIC model according to the Search is IV:CZ:DEZ:EGZ:EMZ:HZ:JKZ:JMZ which is a little bit more complex than the previous model.





Table IV: Search All Output (Width = 3, Levels = 15)

The model has a dDF (Measures the level of complexity of a model) of 13 and dDH (Reduction in the level of uncertainty of the DV, thus the higher the better) of 17.2649%. Compared to the Loopless Search model that had dDF of 23 and dDH of 15.938%, we can therefore conclude that this model is less complex and reduces the level of uncertainty of the DV better than the Loopless model.

Again, we can notice that the search used all 15 levels for the dAIC and therefore we can assume if we increase the Search levels, maybe we can get a better model.

However, one good observation from this search is that at level 9 on ID \*29, we can see the model IV:CZ:DEZ:GJZ:JMZ that is far less complex than the dBIC model while sacrificing only a fraction of the uncertainty. The model has a dDF of 18 and dDH of 16.764%. This is tolerable in my opinion given the simplicity it offers.

As stated earlier, we are going to increase the Search level to 20 with a Search Width of 4 and try and interpret the results. If we get a better complex model than the previous model, then we should use it.

Looking at Table V, we can see that the best dBIC model according to the Search is IV:CZ:DEZ:EGZ:EMZ:HZ:JKZ:JMZ.

The model has a dDF (Measures the level of complexity of a model) of 13 and dDH (Reduction in the level of uncertainty of the DV, thus the higher the better) of 17.2649%. Compared to the Loopless Search model that had dDF of 23 and dDH of 15.938%, we can therefore conclude that this model is less complex and reduces the level of uncertainty of the DV better than the Loopless model. It should also be noted that the %cover drops from 100% to 33%.

At search level 20, we can conclude that, however higher we go, the more complex the models are going to get. Remember our goal is not to get a very complex model but one complex enough that will not be difficult to interpret.

Another, great observation from this search can be noticed which is at level 9 on ID \*39, the model IV:CZ:DEZ:GJZ:JMZ is the same model that was predicted when we had a Search Width of 3 and Search Level of 15.

With that said, all the 3 dBIC models predicted by the Search All are a bit too complex as compared to the IV:CZ:DEZ:GJZ:JMZ model seen above. Normally, we would fit the best dBIC model, but for the sake of understandability and complexity, we are going to the IV:CZ:DEZ:GJZ:JMZ model and see what results we get from the fit. Also, the IV:CZ:DEZ:GJZ:JMZ model includes variables C and D which are noteworthy to be explored.





Table V: Search All Output (Width = 4, Levels = 20)

1. **Fit of the Search All model**

For this paper, not all the data on the fit table is displayed in the Table VI below. The reason being any model with a p-margin greater than alpha value of 0.05 will not be statistically significant. Thus, the Table below has all models with a p-margin value less than the alpha of 0.05.

Before, we look at the output for the Fit Search All Table, we will need to look at the individual composite state of the IV’s when we do a fit of the IV:CZ:DEZ:GJZ:JMZ model and thus this will help us better interpret the Fit Search All table.

**Composite State of C in relation to CZ:**

Looking at Table VI, we can interpret what state C is in relation to CZ. Again, we will be looking at the p-margins and the maximum and minimum percentage values of the DV for each state. From the Table, we can see that if an incident occurred in the Northeast region of the country there is a 81.014% chance that 0 people are killed and 18.896% chance that 1 or more people are killed. Looking at the reverse, if an incident occurred in the West & Midwest region of the country there is a 72.945% chance that 0 people are killed and 27.055% chance that 1 or more people are killed. These percentages are not that useful in my opinion as they tend to lean on state 1 of the DV.



Table VI: Composite State of C in relation to CZ

**Composite State of DE in relation to DEZ:**

Looking at Table VII, we can interpret what state DE is in relation to DEZ. Again, we will be looking at the p-margins and the maximum and minimum percentage values of the DV for each state. From the Table, we can see that if no males or females are involved there is a 95.836% chance that 0 people are killed and 4.164% chance that 1 or more people are killed (Interesting). Looking at the reverse, if 1 male and 1 more females are involved there is a 65.308% chance that 0 people are killed and 34.692% chance that 1 or more people are killed.

****

Table VII: Composite State of DE in relation to DEZ

**Composite State of GJ in relation to GJZ:**

Looking at Table VIII, we can interpret what state GJ is in relation to GJZ. Again, we will be looking at the p-margins and the maximum and minimum percentage values of the DV for each state. From the Table, we can see that if 0 unknown guns and 0 adults are involved there is a 94.873% chance that 0 people are killed and 5.127% chance that 1 or more people are killed. Looking at the reverse, if 1 or more unknown guns and 3 or more adults are involved there is a 49.784% chance that 0 people are killed and 50.216% chance that 1 or more people are killed.



Table VIII: Composite State of GJ in relation to GJZ

**Composite State of JM in relation to JMZ:**

Looking at Table IX, we can interpret what state JM is in relation to JMZ. Again, we will be looking at the p-margins and the maximum and minimum percentage values of the DV for each state. From the Table, we can see that if 1 adult and more than 1 person is injured there is a 97.743% chance that 0 people are killed and 2.257% chance that 1 or more people are killed. Looking at the reverse, if 2 adults are involved and no one is injured there is a 59.121% chance that 0 people are killed and 40.879% chance that 1 or more people are killed.



Table IX: Composite State of JM in relation to JMZ



Table VI: Fit Search All

Table VI shows a summary of the results obtained from the Search All fit performed. We can look at the Z=1 and Z = 2 state percentages and do further analysis of the data set. For each of the states we are going to look at the maximum percentage achieved for one state and simultaneously looking at the minimum state achieved for the opposite state. After determining the states, we will look at the p-margins and states of the dependent variables to see any relations created.

For the maximum percentage of Z = 1 (98.137%) we have the corresponding minimum percentage of Z = 2 (1.863%). The dependent variables D, J and M are all in state 1 which means that there is a 98.137% chance of 0 people being killed if 0 males, 0 adults and 0 people are injured. This information is not useful since it is obvious if the IV’s are held at 0, we would automatically assume nobody would be killed.

If we further explore, the reverse where the maximum percentage of Z = 2 (46.652%) we have a corresponding minimum percentage of Z = 1 (53.348%). The dependent variable D is in state 3, J is in state 3 and M is in state 1 which means that there is a 46.652% chance of more than 1 person being killed if 2 adults, at least one of the 2 adults is a male, 2 adults and 0 people are injured during the incident. More so, looking at the p-margin (0), the model is statistically significant since the value is smaller than alpha of 0.05.

<https://everytownresearch.org/gun-violence-america/>

<https://link.springer.com/chapter/10.1007/978-3-642-82144-8_14>

<http://gunviolencearchive.org/>