On Adapting a War-Gaming Discrete Event Simulator with Big Data and Geospatial Modeling Toward a Predictive Model Ecosystem for Interpersonal Violence

Fortune S. Mhlanga
fsmhlanga@lipscomb.edu
School of Computing and Informatics, Lipscomb University
Nashville, Tennessee 37204, U.S.A.

E. L. Perry
eperry@faulkner.edu
Department of Computer Science, Faulkner University
Montgomery, Alabama 36109, U.S.A.

Robert Kirchner
kirchnerr@sbcglobal.net
Cape Girardeau, Missouri, U.S.A

Abstract

The United States leads industrialized countries in rates of interpersonal violence with homicide being the second leading cause of death for people aged 15 to 24 years. In 2010, more than 4,800 youths (ages 10 to 24) received emergency treatment at hospitals due to injuries caused by physical assaults. This problem has taken epidemic proportions with 33% of high school students reporting physical altercations within the last year, 20% reporting being bullied on school grounds, 16% reporting electronic bullying, and 5% declaring that they had taken a weapon to school within the last 30 days prior to completing a survey conducted by the Centers for Disease Control in 2012. This paper presents an approach to adapting a war-gaming discrete event simulator with big data and geo-spatial modeling towards construction of a predictive model ecosystem for interpersonal violence. The ecosystem will be designed and tested using United States data on interpersonal violence collected over the past 20 years. Spatio-temporal data on interpersonal violence will be collected across the entire United States and stored in a Big Data management and analytics facility that will provide the basis for mapping the patterns of historical and current interpersonal violence. The facility will contain both analytical and simulation tools that collectively allow the researcher to input a strategy and observe predicted future states. The adapted discrete event simulation facility is envisioned to use a predictor-corrector method which will make the ecosystem a self-improving model for interpersonal violence prediction.

Keywords: Discrete Event Simulation, Interpersonal Violence, Predictive, Decision Support Systems, Regression.

1. INTRODUCTION

Interpersonal violence (IPV) among youth can result in significant physical, psychological, social, educational and economic consequences. Although rates of violence among youths have been in decline, IPV death remains the number one cause of death among youths aged 10 – 24. Furthermore, according to the Centers of Disease Control (CDC) Fact Sheet of 2012 on understanding youth violence, treatment of nonfatal injuries sustained from assaults caused more than 700,000 emergency room treatment visits in 2011 (CDC Fact Sheet 2012). Although no state is immune to this issue, Tennessee is ranked among the states with highest homicide rates among youth aged 10 - 24.

The U.S. Department of Justice has indicated that the predictors of youth violence can be grouped into five domains: (1) individual, (2) family, (3) school, (4) peer-related, and (5) community and neighborhood factors (Hawkins et. al 2000). Data from the long-term studies that have identified predictors of youth violence can ultimately help determine violence prevention policy and practice.

Despite the fact that Big Data is still a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard statistical software (Snijders 2012), it is a natural choice platform for our envisaged predictive model ecosystem. Although it has been defined in many different ways (Teradata 2013), a generally accepted definition of Big Data comes from Gartner (Sicular 2013). Gartner defines Big Data in terms of high-volume, high-velocity, and high-variety. IPV data meets all of these criteria:

- From a velocity perspective, there was a violent crime roughly every 30 seconds in the United States in 2011, and one third of high school students reported physical altercations in 2011 (CDC Fact Sheet 2012). Based on the 2008/2009 public high school enrollment (Agus 2010) that translates to an altercation roughly every 6 seconds.

- From a volume perspective, crime data is a clear example of Big Data. The FBI Unified Crime Report program dates back to 1930. In 2012 it included law enforcement agencies representing 308 million United States inhabitants (98.1 percent of the total population) (UCR 2013). This represents just one of many large sources of data on crime and aggression.

- From a variety perspective, multiple facets of crime data must be interrelated to understand IPV. Farrington cites multiple causes of IPV including antisocial behavior, aggressiveness, hyperactivity, parental criminality, poor family management, poverty, delinquent peers and more (Farrington 1998). Bringing together data which can represent this wide variety of potential causes of violence is a quintessential big data challenge.

IPV is a by-product of social, economic and political structure (Saenger 2000). Regardless of whether they are, for example, adult or youth, or male or female, people who unfortunately become exposed to this type of violence often find it difficult to put their traumatic experiences behind them. Thus, studies of IPV now go well beyond describing the physical injuries of victims and survivors to include analyses of psychological and emotional impacts (Kaukinen 2004 and Walker 2014). Because violence takes place at particular locations and times, many studies are also looking into the spatio-temporal patterns of this problem (Walker 2014 and Sparks 2011).

In mathematics, particularly numerical methods, a predictor–corrector method is an algorithm that proceeds in two steps. First, the prediction step calculates a rough approximation of the desired quantity. Second, the corrector step refines the initial approximation using another means. It is common to use an explicit method for the prediction and an implicit method for the correction. For example, in the solutions of ordinary differential equations, a simple predictor–corrector method (known as Heun's method) can be constructed from the Euler Method (an explicit method) and the trapezoidal method (an implicit method).

When a system is driven by the laws of Physics, a predictor-corrector methods that is often used is the Kalman Filter (Kalman 1960), named for Rudolf E. Kalman, one of the primary developers of its theory. Kalman filters are often used in guidance, navigation and control of vehicles, particularly aircraft and spacecraft. The filter forms a prediction of the defining state variables for the system using a time series of noisy input

data from radar, telemetry and on-board sources. The correction is done using a weighted average, with more weight being given to the estimates with higher certainty.

This paper is organized as follows. In Section 2, we present and discuss the high-level components of our Big Data and geospatially-enabled predictive model ecosystem for IPV. Section 3 presents our progress to date. Section 4 presents various ideas that we are currently exploring to support the architecture of our predictive model ecosystem. Finally, Section 5 concludes the paper with an outlook for future work.

2. THE ENVISIONED SYSTEM

Figure 1 presents our first cut at depicting the architectural framework for our predictive model ecosystem for IPV. Our envisaged predictive model ecosystem comprises fundamental components: (i) a generic discrete event simulation facility (DES Facility) which will be adapted to spatio-temporal data on IPV, and (ii) a Big Data management and analytics facility (BDM&A Facility) which will be integrated to the DES Facility. The BDM&A Facility will be designed and built to integrate diverse and aggregated spatio-temporal data on IPV. This input data will represent an aggregation of populations along social, economic or demographic lines. The data will be characterized by a set of attributes that collectively will describe lifestyle, interactions and general quality of life of populations. The BDM&A Facility will be used to facilitate and discover new and unanticipated types of analysis and new information and knowledge pertaining to IPV. The BDM&A Facility is thus an integral component of the knowledge discovery process fostering prediction of future behavior of attributes, identification of the existence of subtle activities or events, and enacting strategies to blunt surprises, which may emerge as unanticipated consequences relative to IPV.

The BDM&A Facility will be used both as input to the initial models of the spatio-temporal data on IPV and as a real world picture of current conditions. When this real world picture is compared with the predicted picture from our model, statistical methods are used to obtain corrections to parameters within the model. This iterative predictor/corrector technique is used to make AIM an ever improving model for the spread of infectious disease.

The DES Facility will be designed and built to have capability of modeling the entire world as one play box, with detailed terrain models of special areas of interest. We will utilize scenario generation tools along with visualization tools to model and visualize the dynamics of IPV in any part of the world. We envisage for the DES Facility to use a predictor-corrector method, which we hope to eventually automate. From our study of spatio-temporal data on IPV, we will be able to identify:

- initial values for a set of parameters used in a discrete event simulation to predict the state of IPV over a given region of study for a future date (perhaps a few weeks or months in the future);
- plausible ranges for each of the parameters; and
- the sensitivity of the predicted state to each parameter. This last factor, the sensitivity of the predicted state to each parameter, is computed using the internal equations of the discrete event simulation.

When the correct amount of time has elapsed, we will do another study to get the actual state of IPV over the region. The differences in the parameters for the actual state and predicted state can be used (along with the sensitivity data) to update the parameters of the discrete event simulation to give a self-improving feature to our predictive model.

3. CURRENT WORK AND INITIAL RESULTS

To avoid "reinventing the wheel," we have chosen to begin with adapting, and running some tests on, an existing generic discrete event simulator (SIMWAR XXI 2004) which will subsequently become the DES Facility for the ecosystem. In the current study, data from The Texas Almanac 2014-2015 (Texas A&M University Press, 2014) on IPV and population statistics of Texas counties has been collected and analyzed using linear regression. We wanted to see if IPV could be predicted from standard population statistics. While the data management, analytical methods and algorithms for the BDM&A Facility have not yet been crystallized and defined, the DES Facility together with historical data from the BDM&A Facility will form an initial model of the IPV.

To date, we have made progress, toward the ability to model and predict levels of IPV, in two areas. First, we have identified an initial set of objects and events for our discrete event simulation. Second, we analyzed some data from a set of Texas counties using linear regression in order to identify a basic predictor equation for IPV within each county.

Initial Set of Objects and Events for the DES

The full set of objects and events in our discrete event simulation is not shown due to page constraints. These objects are designed so that the simulator can be used for (more general) studies related to pandemics (Mhlanga 2013) as well as this initial study on IPV. At present, we have identified more than 35 objects, and their associated attributes, for the discrete event simulation. The overarching (root or superclass) object, called Ecosystem, encapsulates the entire set of objects specific to the study being conducted. It is described by general attributes (such as unique identifier, or ID, for the object together with its long name, or LongName) of all (subclass) objects with a description. Such subclass objects include, for example:

(i) APU (Autonomous Population Unit) – a section of the population that is treated as a single entity. It encapsulates the general information that describes the general attributes of all APUs. (The ID and /or LongName could possibly be formatted such that it maintains a pedigree of its ancestor APUs, e.g., US_TN_NASHVILLE_LIPSCOMB.);

(ii) Demographic – general information that describes the general attributes of all data concerning a specific aspect of the population of an APU, such as gender, marital status, ethnicity, or age range. It can be broken down into subtypes such as male and female for gender;

(iii) Enabler – something that allows a population to affect the ecosystem in the studied way, such as a rifle, knife, personal capability (use arms and legs as weapons), a belief, or a belief system that advocates violence;

(iv) Contributor – Ecosystem specific thing that causes an individual to be more likely to resort to a studied activity. If someone were abused as a child, they are unemployed, they are impulsive, their father uses drugs, etc., that person may be more likely to engage in IPV;

(v) Influence – what affects one object, A, has on another object, B, from the point of view of object A. For example, a long hot spell (A) could cause an increase in the number of physical assaults (B);

(vi) IncidentType - a possible result including incidents such as murder, rape, assault, death, etc., of the use or activation of an Enabler in the study. This will most likely be a class hierarchy. This is because we need to be able to model deaths, because they change the demographics of the APUs. Also, other incidents may affect the data of other objects – so, modeling this as a class hierarchy will allow those types of incidents to be processed differently;

(vii) Zone – a defined geographic area. It can be used to model area-wide things such as weather, economic conditions, political conditions, etc., that would not necessarily be attributed to a single APU;

(viii) Condition – a physical, environmental, social, etc., set of circumstances in effect in a Zone or for a specific APU at a specific time. This could include weather conditions, social conditions, political conditions, etc.;

(ix) Impact – the effect (i.e., impact) an incident has on other objects. For example, a death incident should at least decrease the population of an APU but it may also have an effect on certain demographics within the APU.

Possible events at this time include the following:

(i) Interact – one APU or Actor has some kind of interaction with another APU or Actor;

(ii) EnablerEvaluationEvent – an event to evaluate a specific Enabler of a specific APU to determine if it is to be used or activated;

(iii) IncidentEvent – a result of an enabler being used or activated. For example, a murder, rape, assault, etc.;

(iv) ImpactEvent – makes an impact effective;

(v) Move – when one APU moves from one place to another;

(vi) Spawn – when part of an APU breaks off into a separate, independent APU;

(vii) Merge – when an APU joins another APU to become a single APU;

(viii) Condition Change – one or more attributes of a Condition changes.

Initial Results

For this initial study, our index of IPV in a county is the sum of the number of murders, the number of assaults and the number of rapes during a given time period. Table 1 shows a portion of the initial data set collected from the Texas Almanac (Texas Almanac 2014).

The full set contains data from a randomly selected set of 36 out the 254 counties in the state of Texas. The last column, labeled Tot IPV, is the sum of the number of murders, rapes and assaults in each of the 36 counties during 2012. This column represents the dependent variable for our study. We want to predict it from the independent (or explanatory) variables shown in columns 2 – 7. These variables represent the population of the county, the percentage of the population that is Anglo, the percentage that is Black, the percentage that is Hispanic, the per-capita income of the county, and the percentage of the population that is unemployed. Linear regression runs using these initial data did not produce acceptable results. The page limitations on this paper do not allow enough space to describe the general process of Stepwise Multiple Regression in detail. However, it is described in most of the textbooks (see (Garson 2013), for example) on the subject. (In general, one begins with the independent variable best correlated with the dependent variable. In Stage 2, the remaining independent variable with the highest partial correlation to the dependent is entered and a new regression is completed. This process continues until either (1) the addition of the new variable does not significantly increase r-squared; or (2) all variables are used. If the process terminates and the value of r-squared is not sufficiently high then the researcher looks for new independent variables. The ultimate goal is to get a set of independent variables that are not highly correlated among themselves but are highly correlated to the dependent variable and have the R-squared value above 0.95.)

While we hoped to find a predictive equation with r-squared above 0.95, we found that it could not be done using combinations of the variables from Table 1. This led to a series of experiments in which we added new independent variables and dropped old ones in our regression runs. During these experiments, we used data from the 2014 edition of the Texas Almanac (Texas Almanac 2014), the Texas Department of Public Safety Databases, and the Texas Education Agency Public Records.

A portion of the most successful of these experiments is shown in Table 2 and Figure 2. We used county population (x_1), per-capita income (x_2), public school drop-out rate per 100 students (x_3), the number of incarcerated persons in the county for 2012 (x_4), and the number of concealed carry weapon permits issued in 2012 (x_5) in the county to predict our index for IPV for 2012 (y).

Figure 2 shows a portion of the summary of this regression. Note that the r-squared value is 0.97, indicating that 97% of the variation in the index of IPV is explained by Equation 1.

$$y = (0.001352*x_1) + (0.000599*x_2) + (22.83063*x_3) + (1.396537*x_4) - (0.38155*x_5) + error$$

Equation 1. Predictor equation

Further research is needed to see if this is unique to Texas. (We suspect that it is but much more work is to be done.) In Equation 1, *error* is a random variable which is normally distributed about 0.

We refer to Equation 1 as the *predictor equation*. It leads to the following observations:

(i) IPV can be expected to increase by 1 for each 1000 person increase in population.

(ii) Each $10000 increase in per-capita income will lead to an average of about 6 new cases of IPV.

(iii) An increase (or decrease) of 0.1 in the dropout rate in the public schools will lead to

a corresponding increase (or decrease) of two cases of IPV per year.

(iv) An increase (decrease) of 1 in the county prison population will lead to and increase (decrease) of 1 in the county IPV cases.

(v) The negative sign in the coefficient for x_5 indicates that for each increase of 10 in the number of concealed weapon permits in a given year, one can expect a decrease of about 4 in the cases of IPV in the county.

With reference to the predictor equation observations (i) – (v) above, it is important to note the big difference between "prediction" and "causation" and that we have simply observed that, in the Texas data, there is, from (ii) for example, a positive relationship between per-capita income and IPV. This does not mean that increasing income causes violence. We suspect that this is unique to Texas. In the past few years, there has been a large increase in per-capita income in the oil regions of West and South Texas. Crime has also increased dramatically as oil field workers from around the world have scrambled for jobs in these areas. The same comments are appropriate for observation (iv). Crime is on the rise in many areas of Texas due to the big money brought into the state by the oil companies. We are not suggesting that putting people in jail causes increased violence. However, there is a positive relationship that could be used for predictive purposes. Our overall goal in the study was to produce a predictive equation for the state of Texas. It was not to determine the causes of IPV.

4. IMPLEMENTATION IDEAS

At this early stage of project conception, we are still exploring appropriate storage and management, implementation, modeling and simulation approaches befitting to support our architectural framework. We are currently exploring a Hadoop cluster for the BDM&A Facility along with other statistical tools for the analytics. We are also studying available tools that normalize data, exclude outliers and determine correlations, especially for the so-called "*data munging*" and for extracting the behavior model for the environment.

One approach that has come to mind is to treat IPV as a dynamic system. In such a model, the reference would be the current socio-economic environment and bullying (or being bullied) comprising the output. Measurement of the inputs and output of the previous state would lead to a model able to predict future state. The model could be trained on individual measurements from students in temporal order. After training, an input function could be established able to predict methods which would end the literal cycle of violence. Implementation could be fairly simple as, once data was organized by subject first, then year, the data could be sequentially parsed by the model with very little data existing in memory at any one time.

We are also considering a graph-based human behavior prediction model in which a Bayesian behavior graph could be created based on observed action paths taken by subjects attempting to obtain goals. These paths would be combined into a Bayesian network or even a partially observable Markov decision process (POMDP) if probability becomes a big part of the calculations. The values of probabilities in the conditional probability tables associated with each node could be learned through analysis of the paths taken by the subjects and their associated attributes. As the graph would be relatively small, but would need traversing repeatedly, it's storage as a simple program object would facilitate calculation. Each subject's temporal activities could be projected onto the graph. This approach would require knowledge of subjects' sequential actions.

Another means of looking at bullying is to look at it as an economics transaction. In this case, the bully purchases power from the bullied. The 'cost' of bullying is effectively an externality on the bullied. The unit of the purchase is 'power'. Based on the survey data, bullying or being bullied could be modeled as a transaction and a wealth of 'power' possessed by each subject could be recorded. In such a model, indicators could be considered as factors effecting the volume of 'power' transferred during the bullying transaction or as source of outside 'power' affecting the wealth of an individual. As each subject would need to have their personal wealth of 'power' tracked, persistent storage will be required in the BDM&A Facility. An initial model could be postulated with testing of the sensitivities to specific indicators used to refine the model.

Bullying could also be modeled as a disease which spreads from subject to subject. A dynamic network representing the subjects would be created and updated for each time slice of the data. The probability of spread of the bullying 'disease' would then be calculated based on the indicators also resident on the subject. To accomplish this model, inter-personal relationships would have to be somehow established. This could be gleaned from social network data, or geographic location data. The storage of a large dynamic network is problematic as most graph tools are not well suited for large dynamic networks. Instead, the graph and node attributes may best be represented in a relational database. If the model were to simply promulgate the disease over the graph without changing the underlying architecture, the graph could be stored in a graph database with node attributes identified both by their attribute and time slice. Either way, the relational database or graph database will be a sub-component of the BDM&A Facility.

Although geographic information systems (GIS) are de facto tools for analyzing, interpreting and presenting spatio-temporal information (Longley et. al., 2011), few studies have exploited the capabilities of GIS to help understand, address problems, predict the distribution of and make decisions concerning IPV. One reason is that there are still some unknowns regarding the application of GIS concepts and methods in studies of the spatio-temporal patterns and causes of violence (Pridemore 2010). GIS have, however, been employed in many studies involving crime in general (Wang 2005). Pain et al. (Pain 2006), for example, used GIS to address simple but important "[w]hen, where, if, and but" questions about the effects of street lighting on crime and the fear of crime occurrence. Through GIS, Walker et al. (2014) conducted an exploratory spatio-temporal analysis of the distribution of violent trauma hotspots many of which were correlated with night club areas and Saturday night times.

The beginning point for the discrete event simulator in our study of IPV is an existing simulation which was previously used for combat simulation and war games within the United States Air Force. This simulation is completely data driven, which makes it extensible to other domains. Although we will discard much of the combat portion of this model, we plan to retain the ground and terrain model which can be used to model the entire world (or any portion of it) as a tiled region of variable-sized hexagons and pentagons. The new objects and events will be general enough to facilitate the use and extension of this tool for other studies including world pandemics (Avian-flu, AIDS, etc.), political issues (greenhouse gases, fresh water, etc.), drugs and drug trafficking, and others. The existing simulation also includes two expert systems that may be useful for defining rules for population interaction in all of the studies and as a basis for the corrector method. We will add code to the discrete event simulator to allow a feedback or predictor / corrector loop as shown in Figure 1. A finite set of parameters $\{x_1, x_2, \ldots, x_n\}$ will determine the "state" of the system. For example, we might choose xi as the percent of IPV relative to population Pi. This state can then be easily compared to the actual state at a given time. The differences between the actual and predicted values will determine corrections which can be applied to model parameters and processes to get better predictive capability in the next time cycle (Gershenfeld 1999).

If the existing discrete event simulator proves difficult to adapt, we can consider implementing a neural network or classification system such as the support vector machine which could be used to perform the predictive aspect of the ecosystem. Both of these systems support the predictor/corrector technique or method.
The predictive system is also envisioned to follow a process that employs reinforcement or machine learning techniques. Such systems manipulate the data to produce a model and predict the behavior of the environment and then display the results in a simulation. When real data comes in and comparisons are made, the systems receive feedback that may require them to re-analyze the data and predict a more accurate model.

5. CONCLUSION

We have presented ideas towards development of a Big Data and geospatially-enabled predictive model ecosystem for IPV. This approach in which we combine a robust big data management and analytics capability with predictor / corrector methods to forecast IPV is unique and novel. It extends the use of the Kalman Filter predictor / corrector methods to new domains and the use of data mining and knowledge discovery technologies (commonly used in areas of retail and marketing, banking

and finance, manufacturing, and healthcare).

The capabilities of GIS to handle large volumes of data can be augmented through geovisualization tools and techniques. Unlike GIS which are powerful largely in the area of geocomputational analysis, geovisualization places the user squarely at the center of geospatial data analysis, interpretation and sense-making (Hodza 2009). The goal is to exploit the over 50% of our brain neurons that are primarily for supporting our visual sense. The goal is also to augment human cognition through the use of highly interactive, dynamic and multidimensional visual displays like maps, charts, tables and graphs. This in itself is important because there are many cases where the human eye-mind combination is more effective and efficient at uncovering spatio-temporal patterns, relationships and trends embedded in large and complex data (Byrne 1999). Heer (2013) cites John Tukey, the famous mathematician as having said "Nothing – not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers – nothing can substitute here for the flexibility of the human mind."

Our goal is to facilitate geospatial and geovisual thinking by exploiting the combination of the geocomputational capabilities of GIS and the geovisual analytics power of geovisualization. Both GIS and geovisualization are useful tools and techniques in geospatial data mining (Valencio 2013) which is also of primary importance in this study.

Although we plan to initially study the problem of IPV within some region, our methods and tools are general enough to apply to other medical-socio problems involving the spread of disease (Mhlanga 2013) and (possibly) other domains.

While we have some interesting results from our initial study, we need to extend our data set to include randomly selected counties for randomly selected other states and see if we get similar results. We also realize that the predictor equation in Section 4 does not allow for population dynamics. This will come from the use of the discrete event simulation. Population movement, weather, other dynamic local situations can increase or decrease IPV in the given locality. We view the discrete event simulation as the ideal tool for dynamic analysis. While our initial study used the county as the basic population unit (APU), this may not be the best one for our simulation.

Our collection of objects and events will continue to evolve as we collect more and more data on IPV and begin to build the BDM&A Facility to test these objects and events. We are also garnering a better understanding of the objects and events themselves. For instance, we are looking at ways to specify what a change to an attribute of an object does to other objects when there is an Influence relationship between the objects, or what affect an Incident occurring would have on other objects. For example, a death Incident should affect the population count of one or more APUs, and may also have an effect on one or more Demographics of the population of the APU, affect some aspect of a Contributor, Influence, Condition, Enabler, etc.

Once determined, the BDM&A Facility will also define the schema to accommodate the real-world and intermittent results of the simulation. We will also gradually get a better grasp at defining how the simulation would actually work. We are currently entertaining ideas to determine how close the result of a simulation are to real-world conditions, and what data changes would need to be made to get the results of a simulation run closer to the real-world results. Our goal is get a simulation where we can test strategies for reduction in IPV.

As we continue this work, we also plan to add more geo-temporal analysis techniques to better understand IPV.

6. REFERENCES

Agus, Jessica (2010). "National High School Center at AIR", High Schools in the United States, Quick Stats Fact Sheet, December 2010, (http://www.betterhighschools.org/pubs/documents/HSInTheUS_1210.pdf).

Byrne, Christina A. and Heidi S. Resnick and Dean G. Kilpatrick and Connie L. Best and Benjamin E. Saunders (1999). "The Socioeconomic Impact of Interpersonal Violence on Women", Journal of Consulting and Clinical Psychology, Vol. 67 (3), pp. 362-366, 1999. doi: 10.1037/0022-006X.67.3.362.

CDC Fact Sheet (2012). "Understanding Youth Violence", National Center for Injury Prevention and Control, Division of Youth Violence, Centers for Disease Control, (http://www.cdc.gov/violenceprevention/pdf/yv-factsheet-a.pdf), 2012.

Farrington, David P (1998). "Predictors, Causes, and Correlates of Male Youth Violence", Journal of Crime and Justice, Vol. 24, pp. 421-475, The University of Chicago Press (http://www.jstor.org/stable/1147589), 1998.

Garson, G. David (2013). "Multiple Regression". Statistical Associates Publishing. Blues Book Series, 2013, (http://www.statisticalassociates.com)

Gershenfeld, Neil (1999). "The Nature of Mathematical Modeling", Cambridge University Press, 1999.

Hawkins, J. David and Todd Herrenkohl and David Farrington and David Brewer and Richard Catalano and Tracy Harachi and Lynn Cothern (2000). "Predictors of Youth Violence", Juvenile Justice Bulletin, Office of Juvenile Justice and Deliquency Prevention, U.S. Department of Justice, April 2000.

Heer, Jeffrey (2013). "Interactive Visualization of Big Data", O'Reilly Strata, (http://strata.oreilly.com/2013/12/interactive-visualization-of-big-data.html), December 20, 2013.

Hodza, Paddington (2009). "Evaluating user experience of Experiential GIS", Transaction in GIS, Vol. 13 (5-6), pp. 503-525, 2009.

Kalman, Rudolph E. (1960). "A New Approach to Linear Filtering Prediction Problems", Transactions of the ASME - Journal of Basic Engineering, Vol. 82 (Series D), pp. 35-45, 1960.

Kaukinen, Catherine (2004). "Status Compatibility, Physical Violence, and Emotional Abuse in Intimate Relationships", Journal of Marriage and Family, Vol. 66 (2), pp. 452-471, 2004.

Longley, Paul A. and Michael F. Goodchild and David J. Maguire and David W. Rhind (2011). "Geographic Information Systems and Science", 3rd edition, Wiley, London, 2011.

Mhlanga, Fortune S. and E.L. Perry and C-S Wei, and Peter A. Ng (2013). "Towards a Predictive Model Architecture for Current or Emergent Pandemic Situations", 2013 Summer Simulation Multi-Conference (Summer Sim'13), Society for Modeling & Simulation International, Toronto, Canada, ACM DL 2013 ISBN 978-1-62748-276-9 (57), July, 2013.

Pain, Rachel and Robert MacFarlane and Keith Turner and Sally Gill (2006). "'When, Where, If, and But': Qualifying GIS and the Effect of Street Lighting on Crime and Fear", Environment and Planning, 38, 2055-2074, 2006.

Pridemore, William A. (2010). "Using GIS and Spatial Analysis to Better Understand Patterns and Causes of Violence", 2010 Annual Meeting of the American Association for the Advancement of Science (AAAS), 18-22 February, 2010, San Diego, USA.

Saenger, Sieglinde A (2000). "Family Violence : A Review of the Dysfunctional Behavior Patterns", Minnesota Center Against Violence and Abuse, MINCAVA electronic clearinghouse, 2000, (http://www.mincava.umn.edu/documents/familyviolence/familyviolence.html#idp30185040).

Schiller, Daniel and Ingo Liefner (2007) "Higher Education Funding Reform and University-Industry Links in Developing Countries: The Case of Thailand." International Journal of Higher Education and Educational Planning, vol. 54, no. 4, pp. 543-556, October 2007.

Sicular, Svetlana (2013). "Gartner's Big Data Definition Consists of Three Parts, Not to be Confused with Three "V"s", Gartner, Inc., (http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/), March 27, 2013.

SIMWAR XXI (2004). "Simulation Engine Analyst Manual", Air Force Wargaming Institute, Maxwell AFB, AL. April 2004.

Snijders, Chris and Uwe Matzat and Ulf-Dietrich Reips (2012). "'Big Data': Big Gaps of

Knowledge in the Field of Internet Science'", International Journal of Internet Science, Vol. 7 (1), pp. 1-5, 2012.

Sparks, Corey (2011). "Violent Crime in San Antonio, Texas: An Application of Spatial Epidemiological Methods", Spatial and Spatio-Temporal Epidemiology, 2 (4), pp. 301-309, DOI:10.1016/j.sste.2011.10.001, 2011.

Teradata (2013). The Big Data Conundrum: How to Define It? (http://www.technologyreview.com/view/519851/the-big-data-conundrum-how-to-define-it/), October 3, 2013.

Texas Almanac (2014). Published by Texas Historical Association, 1155 Union Circle, #311580, Dention, Texas, 76203.

UCR (2013). "Law Enforcement Officers Killed and Assaulted, 2012", Summary of the Uniform Crime Reporting (UCR) Program, U.S. Department of Justice – Federal Bureau of Investigation, (http://www.fbi.gov/about-us/cjis/ucr/leoka/2012/standard-ucr-info/about_ucr_2012.pdf), Released Fall 2013.

Valencio, Carlos R. and Thatiane Kawabata and Camila A. de Medeiros and Rogeria C. G. de Souza and José M. Machado (2013). "3D Geovisualisation Techniques Applied in Spatial Data", MLDM'13 Proceedings of the 9th international conference on Machine Learning and Data Mining in Pattern Recognition, 57-68, Springer-Verlag Berlin, Heidelberg, 2013.

Walker, Blake B. and Nadine Schuurman and S. Morad Hameed (2014). "A GIS-based Spatiotemporal Analysis of Violent Trauma Hotspots in Vancouver, Canada: Identification, Contextualisation and Intervention", BMJ Open, 4 (2), DOI: 10.1136/bmjopen-2013-003642, 2014.

Wang, Fahui (2005). "Geographic Information Systems and Crime Analysis", Hershey, P.A.: Idea Group Publishing, 2005.

ISSN: 2167-1508
v7 n3305

APPENDICES

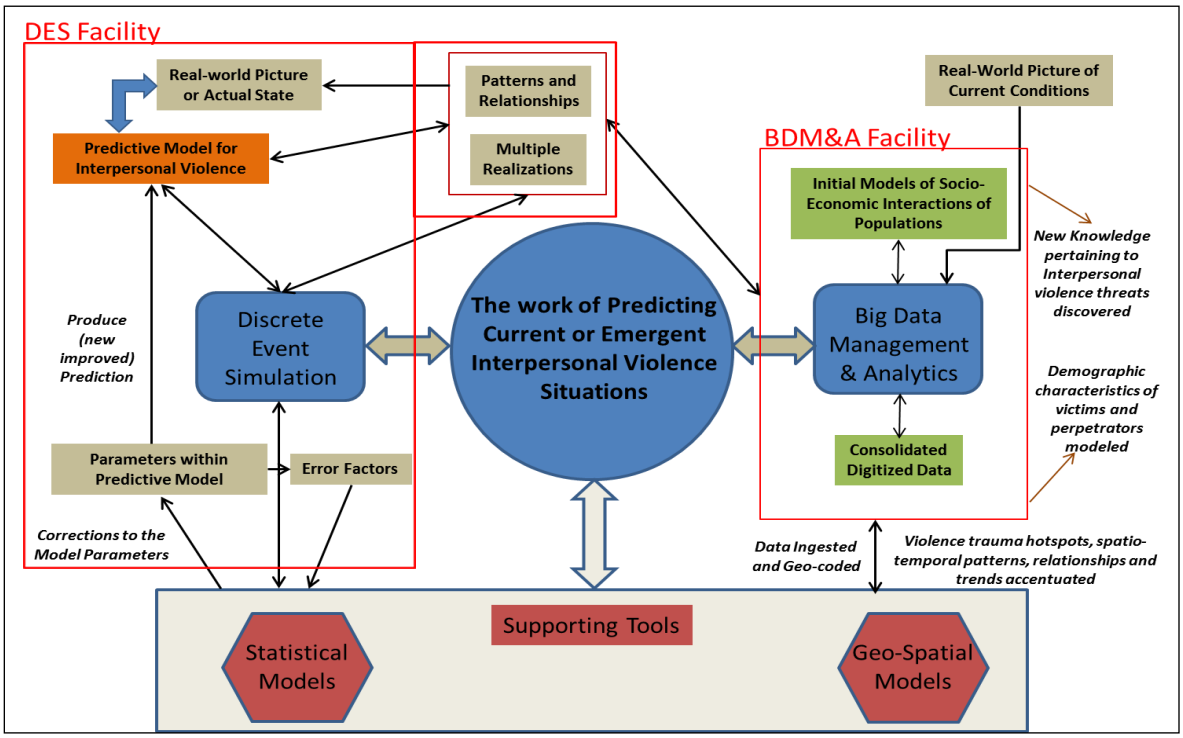

Figure 1. Predictive Model Ecosystem for IPV

Factors							Interpersonal Violence Study			
County	Pop	%Anglo	%Black	%Hisp	PC Income	%Unemp	Murder	Rape	Assault	Tot IPV
Bowie	93148	65.72	24.03	7.05	35360	6.7	7	13	454	474
Brazos	200665	58.75	10.54	23.86	29045	5.7	5	52	670	727
Briscoe	1561	69.11	2.48	26.35	27769	8.1	1	0	0	1
Castro	8164	36.11	2	60.79	48285	5.4	0	0	7	7
Colorado	20696	59.05	12.57	26.97	39030	5.8	3	12	39	54
Crane	4562	39.43	3.31	55.51	36362	5.1	9	1	8	18
Deaf Smith	19360	29.59	0.96	68.36	35880	5.2	0	1	26	27
Denton	707304	63.71	8.46	18.73	42371	5.9	5	133	484	622
Falls	17610	52.21	24.71	21.47	28073	8.9	1	4	23	28
Freestone	19515	67.98	15.88	14.49	31573	6	1	1	39	41
Grimes	26783	59.87	16.08	22.2	31418	6.6	1	7	85	93
Hall	3293	57.86	6.84	34.08	23662	8.4	0	0	5	5
Hamilton	8307	86.69	0.86	10.93	20238	5.8	0	2	11	13
Hill	35115	72.55	6.57	18.93	32266	7.1	1	10	71	82
Leon	16803	76.6	7.54	13.9	35114	7.3	1	4	9	14
Matagorda	36547	46.72	10.95	39.22	33287	10.1	1	11	90	102
Maverick	55365	3.2	0.22	95.24	22188	14.8	3	4	156	163
Menard	2240	62.94	0.75	35.42	30157	7.2	0	0	4	4
Montague	19565	86.92	0.57	10.3	40161	5.1	0	4	20	24
Montgomery	485047	70.26	4.36	21.43	48508	5.8	12	48	547	607
Moore	22313	37.32	1.76	53.02	34060	4	1	8	43	52
Morris	12787	66.24	22.44	8.65	34904	9.5	0	1	53	54
Orange	82977	82.22	8.68	6.36	38163	11	1	16	231	248
Panola	24020	72.94	16.26	8.82	39654	6	0	7	57	64
Potter	122335	48.36	9.82	35.99	33714	5.7	10	111	897	1018
Rains	10943	86.3	2.62	8.36	30131	7.4	0	7	14	21
Randall	125082	76.49	2.63	17.58	40001	4.3	1	2	53	56
Real	3369	70.34	0.82	26.5	30296	7.7	4	6	1	11
Reeves	13798	19.71	4.99	74	23505	9.9	2	0	14	16
San Saba	6002	66.66	3.45	28.36	31384	8.4	1	0	12	13
Scurry	17126	56.91	4.64	39.96	37970	4.3	1	9	49	59
Taylor	133473	50.42	0.73	46.77	37132	5.3	3	44	335	382
Titus	32663	48.16	9.29	40.58	28542	7.3	0	0	62	62
Travis	1095584	50.29	8.09	33.87	43198	5.7	33	246	2703	2982
Upton	3283	46.2	1.64	50.36	45030	3.7	0	0	1	1
Williamson	456556	63.09	6.13	23.61	40067	5.9	2	89	349	440
Wilson	44370	58.2	1.72	38.52	34810	6.2	3	6	34	43
Yoakum	8075	38.08	0.94	59.39	41060	3.5	1	5	0	6
Zapata	14290	6.4	0.36	92.83	25162	6.9	0	1	28	29

Table 1. Initial data from Texas counties (Texas Almanac 2014)

County	Pop	PC Income	DORate/100	Incar2012	CCPermits2012	Tot IPV
Bowie	93148	35360	0.7	338	687	474
Brazos	200665	29045	2.4	585	1177	727
Briscoe	1561	27769	0	1	7	1
Castro	8164	48285	1.1	16	29	7
Colorado	20696	39030	1	52	178	54
Crane	4562	36362	1.2	11	40	18
Deaf Smith	19360	35880	0.8	72	101	27
Denton	707304	42371	0.7	1093	4716	622
Falls	17610	28073	2.7	34	105	28
Freestone	19515	31573	0.8	53	113	41
Grimes	26783	31418	1.7	73	186	93
Hall	3293	23662	0	9	14	5
Hamilton	8307	20238	0.7	15	75	13
Hill	35115	32266	0.4	145	214	82
Leon	16803	35114	0.4	22	190	14
Matagorda	36547	33287	0.6	125	234	102
Maverick	55365	22188	1.3	70	46	163
Menard	2240	30157	0.6	6	9	4
Montague	19565	40161	0.3	63	157	24
Montgomery	485047	48508	0.1	1145	4223	607
Moore	22313	34060	1	42	114	52
Morris	12787	34904	0.2	31	79	54
Orange	82977	38163	1.5	181	767	248
Panola	24020	39654	1.2	52	174	64
Potter	122335	33714	2.3	491	551	1018
Rains	10943	30131	0.4	26	80	21
Randall	125082	40001	0.5	277	1191	56
Real	3369	30296	4.1	6	43	11
Reeves	13798	23505	0.9	33	10	16
San Saba	6002	31384	0.2	10	52	13
Scurry	17126	37970	0.7	45	98	59
Taylor	133473	37132	2	523	944	382
Titus	32663	28542	0.2	105	150	62
Travis	1095584	43198	2	2314	4546	2982
Upton	3283	45030	0	9	7	1
Williamson	456556	40067	0.6	574	3022	440
Wilson	44370	34810	0.6	66	361	43
Yoakum	8075	41060	0.2	13	44	6
Zapata	14290	25162	2.2	36	48	29

Table 2. Final data from Texas counties

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.98633
R Square	0.97284 7
Adjusted R Square	0.96873 2
Standard Error	90.7009 7
Observations	39

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	*Upper 95.0%*	*Upper 95.0%*
Intercept	-31.276	89.61255992	-0.34901320 2	0.729296596	-213.59409 05	151.04 22	-213.594	-151.042 2	207.278 3	312.980 4
X Variable 1	0.00135 2	0.000382242	3.53822744 5	0.001220996	0.0005747 82	0.0021 3	0.00 0575	0.00213 1	0.00524	0.00544 8
X Variable 2	0.00059 9	0.002562329	0.23396538	0.816457299	-0.0046136 01	0.0058 13	0.00 461	-0.00581 3	0.01396 1	0.01256 1
X Variable 3	22.8306 3	18.07397649	1.26317684 8	0.2153778	-13.941153	59.602 41	-13.9 412	-59.6024 1	133.862	-0.31876
X Variable 4	1.39653 7	0.144064212	9.69384971 7	3.50639E-11	1.1034359 74	1.6896 38	1.10 3436	1.68963 8	-	0.29273
X Variable 5	-0.38155	0.04080004	-9.35158773 2	8.44079E-11	-0.4645534 59	-0.2985 4	-0.46 455	-0.29854		

Figure 2. Regression summary output