Using Big-Data Analytics for Identifying Hot Spots of Border Security  
Dr. Haibo Wang, Texas A&M International University

This project will design a comprehensive data aggregation and analysis system to provide the decision support for identifying hot spots of border security using a complex network model to present transportation infrastructure in the border region. All these research related data, such as the GIS-based infrastructure data, the economic development and public service level data, the demographic data for the treatment group and control group, will be aggregated on both space and time dimensions and analyzed by using “big data” models and tools developed in this study.

1. Theme Area: Risk Management / Operations Research or Risk and Decision Analysis

2. Principal Investigator: Haibo Wang

3. Institution: Texas A&M International University

4. Co-Investigators: none

5. Research Transition Lead: Dr. Haibo Wang, Texas A&M International University

6. Keywords: border security, big data analytics, complex system analysis, difference-in-difference estimation, optimization, simulation, mathematical programming

7. Brief Description:

This project will design a comprehensive data aggregation and analysis system to provide the decision support for identifying hot spots of border security. PI will use a complex network model to present transportation infrastructure in the border region using the GIS-based infrastructure data from FEMA HAZUS. The economic development and public service level in the border region will be collected from Texas Center for Border Economic and Enterprise Development at Texas A&M International University. The demographic data from US Census Bureau and the immigration database from Mexico, which is compiled and processed by the General Directorate of Protection and Consular Affairs, Directorate General for Communications and Information Technology and the Institute of Mexicans Abroad of the Ministry of Foreign Affairs\(^1\), will be used for identifying the treatment group and control group in the study. All research related data will be aggregated on both space and time dimensions and analyzed by using “big data” models and tools developed in this study. For example, a new difference-in-difference regression\(^2\) and estimation model will be designed and implemented for analyzing the causal relationships of border security and regional economic development. Data will be presented in a more meaningful form using data visualization techniques. PI will present a new complex network model to investigate the influential and disruptive nature of the border security and human trafficking. A set of criteria will be developed to measure the efficiency of border crossing and identify the hot spots of border crossing. PI will develop a predictive model for return of migrants based on the migration data and regional economic development/restructuring from both US and Mexico sides\(^3\). In addition, this project is closely related to PI’s current teaching and research activities on data mining and big data analytics.

8. Research Objectives:

This research will: (a) design a comprehensive data aggregation and analysis system to integrate databases from both US and Mexico and to provide the decision support for identifying hot spots of border security; (b) create data-driven predictive model for voluntary return of migrants due to border enforcement and economic restructuring on both sides of the border; (c) build a complex network model presentation for transportation infrastructure to investigate the influential and disruptive nature the human trafficking pattern\(^4\) and network in the border crossings; (d) design and implement a new difference-in-
difference regression and estimation model to analyze the causal relationships of border security and regional economic development; (e) present data in a more meaningful form using visualization techniques; and (f) develop a model to evaluate a set of criteria to measure the border enforcement methods, which can identify the hot spots in border crossings for the purpose of providing more accurate number of illegal border crossings.

9. Research Transition Objectives:
Currently U.S. Customs and Border Protection (CBP) and other government agencies are using commercial version of a relational database such as ORACLE to store the information on border crossing of human and vehicles and other border security activities. This project will develop novel big data analytics models and tools to be used by DHS personnel in the Capability Development Support (CDS) Group in carrying out their analysis prior to making decisions such as operations and requirements analysis and systems engineering. The predictive models and tools will be developed in close collaboration with CDS and support their current projects on the south Texas border such as the Rio Grande Valley Systems Analysis Project and other on-going projects. In addition, the new big data analytics models and tools will be developed in close collaboration with the Border Situational Awareness (BSA) program and Border Enforcement Analytics Program (BEAP) program within Science and Technology directorate, with the goal of developing comprehensive data aggregation and analysis system to integrate databases from both US and Mexico.

10. Interfaces to Current CREATE Projects:
This work will maintain a close interface with CREATE’s projects such as: 1) Modeling the Dynamics of Risk Perception and Fear: Examining Amplifying Mechanisms and Their Consequences; 2) Modeling the Dynamics of Risk Perception and Fear; and 3) Examining Amplifying Mechanisms and Their Consequences and Working Together for A Safer Tomorrow.

11. Previous or current work relevant to the proposed project, why is DHS interested, identify/who are your expected DHS users.
Currently U.S. Customs and Border Protection (CBP) and other government agencies use a commercial version of a relational database such as ORACLE to store the information, according to the report5 (OIG-13-88) from Office of Inspector General at Department of Homeland Security and a white paper6 from the vendor as in the case study of EXADATA in 2014. For structured and clear data prepared by professional consultants or IT department, ORACLE is the most popular commercial database management software worldwide. Nowadays, most data is collected in real time from in-memory data collecting systems such as scanner (RFID or bar code), surveillance camera, mobile devices, and other electronic devices. The relational database software such as ORACLE is not designed to meet the challenges of data generated from internet and mobile devices in term of scale and agility and application developers have to solve the mismatch issues between the in-memory data structure of big data and the underlying structure of relational database systems by converting the in-memory data to tables and other relational structures.

We have worked on the critical infrastructure management for water security with optimization and simulation (DOE funded project DE-AC05-00OR22725) and are currently working on the wireless sensor network security and reliability project (Texas A&M International University Research Development Award). We also plan to collaborate with the Texas Center for Border Economic and Enterprise Development and local border-crossing logistic companies on improving the transportation security of cargo in Laredo, which is the second largest port in the North America. We are also working with local public transportation operators to use “big data” analytics for urban sustainable development for economic enhancement zones in Laredo, which is funded by Texas A&M International University.
Research Grant. We plan to analyze the pattern of migration and return rate using cluster analysis and the results can be benefit to both countries. We also plan to collaborate with the researchers of Border Situational Awareness program and Border Enforcement Analytics Program (BEAP) program of the DHS.

12. Major Deliverables, Research Transition Products and Customers:
Project deliverables will consist of a report that will: (a) develop a framework and design of comprehensive data aggregation and analysis system; (b) develop visualization tools using R program to present the data; (c) develop tools using R to present the complex network of transportation infrastructure; (d) develop risk-based forecasting model on the hot spots of border security based on regional economic development and demographic information in the border region cities on both US and Mexico sides; Customers: DHS – IP; DHS – ORD; and State Governments.

Products will include: framework and design of information systems using Apache, Hadoop, MapReduce, XML schema, and MongoDB; R software package for prediction and visualization; and a user manual of the package prepared for DHS customers. Project research deliverables will consist of a report and research publications that will: (a) develop tools using R to present the complex network in a real time manner; (b) develop model to evaluate a set of criteria used to measure the border enforcement methods, which can identify the hot spots in border crossings for the purpose of providing more accurate number of illegal border crossings. Research Transition Products will include: a framework and design of information systems using Hadoop, MapReduce, XML schema and MongoDB; visualization computer programs in R software; tools using R software to present the complex network; Customers: DHS – IP; DHS – ORD; and State Governments.

13. Technical Approach:
The tools and methods the proposed project will utilize for the big data analytics include Hadoop, MapReduce and MongoDB, XML schema, and R software. These open source software provide a number of benefits including free distribution, low cost, worldwide developer community, and are easy to use. Hadoop is an open source software environment for large data storage and distribution that uses low cost computer clusters and MapReduce is an open source functional programming environment for data manipulation and processing. Both Hadoop and MapReduce can be deployed in the cloud or a local datacenter. MongoDB is an open source cross-platform NoSQL (not only SQL) database management software that provides the benefit of improved system productivity through an application oriented data structure and also improves the data utilization efficiency for storing and processing large data volumes, reducing latency and improving input/output performance. XML schema is widely used as an object management standard and is the foundation of semantic systems such as search engines and artificial intelligence. R software is an open source cross-platform programming language with a platform adaptive run-time environment for statistical computing and graphics. There are more than 3,500 application/function based packages developed by worldwide researchers and widely used for a variety of statistical models including statistical tests, linear and non-linear models, time series forecasting, data mining and predictive analytics, etc. There are more functions and algorithms in R software than in the other three most popular commercial statistical software combined. It provides a collaborative environment for new applications or functions to be added on as a cross-platform library.

The PI and team will design an automated system to collect data from different sources including an official database from the Mexican government using the Apache Hadoop server and MongoDB server (Figure 1) and examine the issue of missing data and errors, then apply an imputation technique such as maximum likelihood method or Markov Chain Monte Carlo method on the basis of missing data pattern. This project will develop a tool to present complex network for the transportation system of illegal border
crossings and continuous movement into inland using the traditional pathway and various destinations by different demographic groups. This project will study the overall behavior of the mobility of illegal border crossings in the transportation system to determine the complexity and entropy of the human trafficking network\(^4\) and use excess entropy to measure the complexity. In two dimensions, the entropy density \(h\) can be expressed as:

\[
h = \lim_{n \to \infty} h(n)
\]

Then the excess entropy \(E_c\) is defined as:

\[
E_c = \sum_{n=1}^{\infty} (h(n) - h)
\]

This study will choose \(E_c\) to capture the time and spatial structure of the complex network in transportation system of illegal border crossings.

Difference-in-Difference (DiD) estimation has become an important tool to study causal relationships\(^3\). Researchers can compare the difference in outcomes from different time periods on the group affected by the intervention known as “treatment group” to the unaffected group known as “control group”. In general, the outcomes of intervention are observed for two groups for two time periods with the help of a big data tool on data aggregation. The treatment group A can find jobs in the border regions due to the regional economic development/restructuring in the second time period but not in the first period, the control group B cannot find jobs in the border regions during either period and have to return or move out of the traditional destinations. The enforcement of border security can also be a factor of intervention and used to study the causal relationship between two groups.
The outcomes such as employment opportunity, access to public services and other benefits for group B at time period $t$ can be written as:

$$y_{i,t} = \delta(t) + \alpha \cdot D(i, t) + \theta(i) + v(i, t)$$

where $\delta(t)$ is a time-specific component and is the coefficient in the model, $\alpha$ represents the effect of the treatment, $\theta(t)$ is a space (group)-specific component and indicates possible differences between the treatment and control groups prior to the economic development/restructuring, and $v(i, t)$ measures the transitory shocks for the group and presents aggregate factors that would cause changes in outcome $y$ over time, even in the absence of regional economic development/restructuring.

Then the estimation model for difference-in-difference is:

$$\delta_1 = (\bar{y}_{B,2} - \bar{y}_{B,1}) - (\bar{y}_{A,2} - \bar{y}_{A,1})$$

In the project, the DiD model is programmed in software R. PI and team will also develop a new procedure to control for the covariates to measure the resource allocation on a set of outcomes. In this project, a new DiD estimation and regression model will be designed and implemented in software R for analyzing the aggregated data to examine the relationship of economic development/restructuring and voluntary return of voluntary return of migrants or moving out of traditional destinations. The same procedure can be applied to analyze the enforcement of border security.

PI and team will also develop a new procedure to control for the covariates to measure the resource allocation on a set of outcomes. The results will be validated at least 10 folds using other advanced regression models. PI will develop complex network model for transportation infrastructure based on the
results of DiD estimation and identify the key nodes in the network as the hot spots of border security. We will also perform cost-benefit analysis on the current enforcement methods.

14. Major Milestones and Dates:
1. PI and team will design an automated system to collect data from different sources using Apache Hadoop server, MapReduce program, XML schema, and MongoDB server which can aggregate data on both space and time dimensions – July 1 to August 31, 2015.
2. PI and team will prepare R program on basic data analysis to understand nature of “big data”--September 1 to October 31, 2015.
3. PI and team will design “big data” analytics models based on the nature of big data and implement the models with open source software R, and prepare the summarized data for simulation and optimization --November 1 to December 31, 2015
4. PI and team will develop visualization package in R to present the data in the meaningful form and present complex network model in R to understand the transportation methods in the illegal border-crossings – January 1 – April 30, 2016
5. Summarize and evaluate the prediction models in R package with a user manual, disseminate results for grant report and continuous study on related issues– May 1 – June 30, 2016.

15. References:
1. Institute of Mexican Abroad (2013), http://www.ime.gob.mx

16. CV

BIOGRAPHICAL SKETCH
NAME: Haibo Wang
POSITION TITLE Killam Distinguished Associate Professor
EDUCATION/TRAINING

<table>
<thead>
<tr>
<th>INSTITUTION AND LOCATION</th>
<th>DEGREE</th>
<th>MM/YY</th>
<th>FIELD OF STUDY</th>
</tr>
</thead>
<tbody>
<tr>
<td>South China University of Technology, Guangzhou, China</td>
<td>B.S.</td>
<td>1991</td>
<td>Biochemical engineering</td>
</tr>
<tr>
<td>University of Mississippi, University, MS</td>
<td>M.S.</td>
<td>1996</td>
<td>Chemistry</td>
</tr>
<tr>
<td>University of Mississippi, University, MS</td>
<td>M.S</td>
<td>1997</td>
<td>Computer &amp; Information Sciences</td>
</tr>
<tr>
<td>University of Mississippi, University, MS</td>
<td>Ph.D.</td>
<td>2004</td>
<td>Business Administration</td>
</tr>
</tbody>
</table>

A. Professional Positions
Killam Distinguished Associate Professor (2014–present), TAMIU
Wang, Using Big-Data Analytics for Identifying Hot Spots of Border Security

**B. Selected Peer-reviewed Publications**

1. “An Ejection Chain Approach for the Quadratic Multiple Knapsack Problem” accepted by European Journal of Operational Research, 2014 (with Bo Peng, Zhipeng Lu, Gary Kochenberger)
4. “Single Machine Scheduling to Minimize a Modified Total Late Work Function with Multiple Due Dates” *Production and Manufacturing Research*, 2(1), 624-639, 2014 (with Bryan Kethley and Bahram Alidaee)