MACK-BLACKWELL
Rural Transportation Center

University of Arkansas
4190 Bell Engineering Center
Fayetteville, AR 72701
479.575.6026 – Office
479.575.7168 - Fax

NTSCOE NATIONAL TRANSPORTATION SECURITY CENTER OF EXCELLENCE

MBTC DHS 1105 – Information Enhancement Among Aviation Security Partners

Justin R. Chimka – PI
jchimka@uark.edu

Jing Wu & Ryan Black
University of Arkansas

April 2011

Prepared for
Mack-Blackwell Rural Transportation Center
National Transportation Security Center of Excellence
University of Arkansas

ACKNOWLEDGEMENT
This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2008-ST-061-TS003.

DISCLAIMER
The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

Abstract

Part I: Models previously created by GRA, Inc. for the Federal Aviation Administration to estimate total annual operations by general aviation (GA) airport have been recreated and examined by the authors. Models were originally estimated by GRA to predict the future size of airports, but research described here would go toward detection of unusual GA activity that might be due to a homeland security threat. Toward this end the authors have systematically discovered a statistical model of GA operations that is more efficient than what the literature describes. Part II: Official border-crossing data were collected, assumed to represent usual activity, and summarized with linear regression models to facilitate residual control charts and detect unusual activity that might be attributed to a security threat or some other external variable. Particularly good results are presented for Northern highway, and loaded rail, containers at North American borders with respect to the sigma limits of statistical quality control. Additionally this manuscript describes the estimation of what might be useful model-based control carts for activity across three ports of the State of Michigan: Detroit, Huron and Sault Ste. Marie.

Table of Contents

Part I: Re-estimating and remodeling general aviation operations 4

Part I: Bibliography 18

Part I: Appendices 19

Part II: Regression-based monitors of North American border-crossing activity 22

Part II: Bibliography 30

Re-estimating and Remodeling General Aviation Operations

Ryan Black and Justin R. Chimka

University of Arkansas

Author Note

Ryan Black and Justin R. Chimka, Department of Industrial Engineering, University of Arkansas.

This material is based upon work supported by the U.S. Department of Homeland Security under Award Number: “2008-ST-061-TS0003”. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

Correspondence concerning this article should be addressed to Justin R. Chimka, Department of Industrial Engineering, University of Arkansas, Bell Engineering Center, Room 4027; 800 West Dickson Street, Fayetteville, AR 72701. Telephone: (479) 575-7392. E-mail: jchimka@uark.edu

Abstract

Models previously created by GRA, Inc. for the Federal Aviation Administration to estimate total annual operations by general aviation (GA) airport have been recreated and examined by the authors. Models were originally estimated by GRA to predict the future size of airports, but research described here would go toward detection of unusual GA activity that might be due to a homeland security threat. Toward this end the authors have systematically discovered a statistical model of GA operations that is more efficient than what the literature describes.

Keywords: general aviation, homeland security and linear regression

Re-estimating and Remodeling General Aviation Operations

Since September 11, many steps have been taken to improve security against attacks on commercial aviation, but relatively little has been done to secure general aviation (GA). One reason for the security gap is that GA operates differently than the commercial aviation industry making it difficult to borrow improvements. Another reason for lower GA security standards is that many people did not perceive GA as a serious threat, since planes carry much less fuel and are much smaller in size than their commercial counterparts. However in February 2010 a suicide attacker crashed a single-engine plane onto an Austin IRS building killing one employee and injuring thirteen others. "Thousands of civilian aircraft fly within the general aviation system every day. But there are few regulations, laws, or security procedures that would prevent a pilot with ill intentions from using a plane for evil purposes (Lubold, 2010)."

To accommodate the need for improved GA security, one goal should be to integrate a variety of relevant data formats, "and transform raw data into useful and understandable information that enables productive and efficient analysis (IDS University Affiliate Center for Multimodal Information Access and Synthesis)." Our objective is to understand the variation associated with usual GA activity and operations, so unusual activity can be detected, analyzed and resolved. General techniques include estimation and design of relevant statistical model-based quality control charts. This opportunity to specialize in model-based control for an applied context should eventually result not only in contributions to GA security but also to quality engineering. The research described here improves upon previously existing models of GA operations data and would make possible improved monitoring and detection for GA security.

Motivation

The Top Ten Challenges Facing The Next Secretary of Homeland Security includes the following: "Continue to improve intelligence and information sharing (Homeland Security Advisory Council, 2008)." However, while the University Affiliate Centers to the Institute for Discrete Sciences (IDS) were established by DHS for advanced methods research in information analysis, IDS activities focus on common author identification, influenza surveillance and text analysis. What is described here is part of ongoing activities that will adopt and/or develop tools to derive knowledge specific to potential attacks against general aviation (GA). Additional activities would extend model-based control of GA to the most appropriate of other contexts chosen among highway, maritime transportation systems, mass transit, pipeline systems, and rail.

Commercial examples relevant to GA include Incident Reports and Surveillance Detection Reports filed by Federal Air Marshals (FAM), and analyzed by law enforcement organizations in a Tactical Information Sharing System (TISS). FAM also place in TISS incident reports by airline employees, and Screening Passengers by Observation Techniques identifies unusual activity by utilizing behavioral analysis.

In the GA domain TSA and the Aircraft Owners and Pilots Association have implemented an Airport Watch Program using pilots for reporting suspicious activity. TSA and the National Response Center (U.S. Coast Guard) have implemented the GA Hotline for airport operators, technicians and pilots to report suspicious activity. However there are not more formal information reporting and sharing systems available to GA. In order to design such effective systems, and make GA a more equal partner in homeland security, the following would seem to be important exploratory activities.

1. Consider what is relevant about commercial examples to GA, and make recommendations for improved intelligence and information sharing which originates at GA landing facilities.
2. Reference the Airport Characteristics Measurement Tool (Transportation Security Administration, 2004) to develop reporting standards, and analyze information that would come from reports.
3. Estimate and/or identify models of usual GA activity that could be used to detect potential attacks.
4. Extend the philosophy that if we can estimate good models of usual activity associated with transportation, then we can effectively monitor operations, and detect unusual activity that may indicate a security threat.
5. Identify the other (in addition to GA) contexts that make the most sense physically for extension of lessons learned from GA. These would seem to be the ones to be most likely affected by unscheduled activity.
6. Explore the concept of a simultaneous, multi-context monitor that would integrate not only information from disparate sources within mode, but also information across modes to enhance transportation security.

The research described here is most relevant to exploratory activity 3. Estimate and/or identify models of usual GA activity that could be used to detect potential attacks.

Literature

Soon after its description of the Top Ten Challenges DHS released an article on strengthening GA security (DHS Press Office, 2008). The article describes an effort to minimize vulnerability of GA flights used to deliver illicit materials, transport dangerous weapons or people, or utilize aircrafts as weapons. DHS is implementing the Electronic

Advance Passenger Information System (eAPIS) which will mandate GA operations to know more detailed information about arriving and departing planes, and the passengers and crew onboard. These data are sent through eAPIS or an approved alternate system one hour prior to departure for flights arriving into or departing from the United States.

NASA has been working on constructing an "Aviation Data Integration System (ADIS)" which provides rapid access to various data sources such as the following (Kulkarni, Wang, Windrem, Patel, & Keller, 2003): weather data, airport operation condition reports, radar data, runway visual data, navigational charts, radar track point records and track deviation, aircraft conditions, and Jeppesen charts. These data are integrated and analyzed along with what is collected by cockpit data recorders (time since flight start, latitude, longitude, altitude) to determine when aircraft are behaving abnormally,

Also taking steps to improve GA security is Transport Canada (2007). Phase II of their Electronic Collection of Air Transportation Statistics (ECATS) allows GA planes to submit air transportation data through web interfaces. This new data integration system should improve the timeliness and availability of air transport data for analysis and interpretation. Transport Canada uses current and secure information technology to collect and distribute data. A collaboration of GA entities and a partnership between the government and industry have allowed this high security information to be shared to improve GA security.

The Federal Aviation Administration releases a terminal area forecast summary each year (FAA Office of Aviation Policy, 2007). This summary predicts the number of enplanements for future years to come for commercial aviation airports, but currently the model is not applied to GA. To approximate this, historical relationships between airport passenger demand and/or activity measures, and local and national factors that influence aviation activity, are examined. The FAA also used regression analysis to reforecast the time series. Regression models including variables that characterize airports and their activities have been used to accurately forecast the number of operations at an airport. These data can aid in building terminal area forecast models for GA airports. It follows that predicting the annual number of operations at GA airports should also aid in identifying unusual activities associated with those airports.

The FAA administers a GA survey each year to assure safe operation of all aircraft in the National Airspace System. To do this the FAA classifies GA aircraft according to seven different categories that include fixed wing piston, fixed wing turboprop, fixed wing turbojet, rotorcraft, other aircraft, experimental, and light-sport. The survey requests that aircraft owners provide the following information:

- Number of total hours flown in previous year
- Airframe hour reading and the most common place the aircraft was flown in survey year
- Hours flown by flight plan and flight conditions
- Type of landing gear and number of landings
- Fuel type and average fuel consumption
- Percentage of hours flown by person or company other than primary owner
- Avionics equipage

Due to adjustments to the GA survey and the way that it is administered, the response rate has been increasing for the past eight years. The collection of these data would seem vital to understanding baseline GA operations. The information obtained by these surveys should be used to estimate a statistical model of annual number of operations at a GA airport where an operation is defined as a landing or a takeoff.

In 2000, Hoekstra developed a methodology for estimating the annual number of GA operations at an airport, and the annual number of GA operations per based aircraft at an airport (GRA, Inc., 2001). In July 2001, the GRA modified Hoekstra's model to more accurately estimate the number of GA operations for non-towered airports based on data from towered airports. To do this many of the same independent variables were reused, and several were added. The variables used for the regression analysis appear in Table 1.

Table 1. Variables

Variable	Description
OPS	Annual GA Operations at an airport (landings and takeoffs)
BA	Total Based Aircraft at an airport
Pop100	1998 Population within 100 miles
WACAORAK	Categorical variable, 1 if state is CA, OR, WA, or AK, 0 otherwise
BA2	Total Based Aircraft at an airport squared
IN50MI	Percentage of based aircraft among based aircraft at GA airports within
Pop25/100	Ratio of Pop25 to Pop100
IN100MI	Percentage of based aircraft among based aircraft at GA airports within
FAR139	Categorical variable, 1 if airport is certificated for commercial air
POP	County population where airport is located in 1999
Se BA/BA	Single engine based aircraft/All based aircraft
TOWDUM	Categorical variable, 1 if airport is towered airport, 0 otherwise
VITFSNUM	Number of FAR141 certificated pilot schools at an airport
PCI	Per Capita Income in the county in which the airport is located in 1999
EMP	Non-agricultural Employment in the airport's county in 1999
WSTAK	Categorical variable used in place of WACAORAK in Hoekstra's
WST	Categorical variable, 1 if airport is located in FAA Western Region, 0
AAL	Categorical variable, 1 if airport is located in Alaska, 0 otherwise

R12	Categorical variable, 1 if airport is located in FAA New England
VITFS	Categorical variable, 1 if airport has FAR141 certified pilot school, 0
VITFSEMP	Employees of FAR141 certificated pilot schools at an airport
Pop50	1998 Population within 50 miles
Pop25	1998 Population within 25 miles

Methods

Model Recreation

To better understand relationships among airport characteristics and the annual number of airport operations, attempts were made to recreate linear regression models previously constructed by GRA. An equation summary analysis is provided in Table 2. **(**Appendix A contains an equation matrix that describes each equation in terms of the independent variables included.) Each equation is described in Table 2 according to the following.

- Dataset
- Number of airports included
- Whether or not a dummy variable was included to distinguish between towered and non-towered airports
- Number of independent variables
- R-squared value
- R-squared value of the associated GRA model
- Adjusted R-squared value

For models involving non-towered airports R-squared values are slightly yet inexplicably different than the ones estimated by GRA.

Table 2. Model Summary

Eq.	Dataset	Airports	Dummy?	Ind. Vars.	R^2	GRA R^2	Adj R^2
1	Towered	127	No	1	0.556	0.556	0.553
2	Towered	127	No	2	0.640	0.640	0.634
3	Towered	127	No	3	0.666	0.664	0.658
4	Towered	127	No	4	0.703	0.703	0.693
5	Towered	127	No	5	0.723	0.723	0.712
6	Towered	127	No	6	0.735	0.735	0.722
7	Towered	127	No	7	0.744	0.744	0.728
8	Towered	127	No	6	0.742	0.742	0.729
9	Towered	127	No	7	0.748	0.748	0.733
10	All	232	No	8	0.711	0.717	0.700
11	Towered	127	No	8	0.727	0.727	0.709
12	Non-towered	105	No	8	0.645	0.648	0.615
13	All	232	Yes	8	0.739	0.743	0.729
14	Towered	127	No	7	0.748	0.748	0.733
15	Non-towered	105	No	7	0.563	0.569	0.531

New Variable Creation

To further improve the efficiency of our models, we revised those of GRA by creating and including some new variables. Instead of including a ratio of single engine aircraft to total based aircraft (Se BA/BA), a simpler single engine based aircraft (Se BA) variable was created. This variable was created using the data values from the total based aircrafts and the ratio of single engine aircrafts to total based aircraft (a redundant variable in the GRA analysis).

A new regional variable was also created to more efficiently describe the location of GA airports. In the GRA models, four dummy variables are used to describe location. It was found, controlling for relevant independent variables, that significant differences between locations existed only where Alaska is involved. In other words dummy variables that describe location in detail greater than Alaska versus not Alaska would not contribute to efficient statistical models of GA operations. Therefore categories other than Alaska were collapsed. Next, a new regression model was created that included new variables AAL and SEBA. Also, the demographic variables PCI and EMP were added back to the model in order to determine if they contributed significantly to the model. These two variables were not included in any of the GRA models. (Equation 1 of Appendix B shows more details of this new regression model.)

Introducing Second Order Terms

Many of the *p*-values for this model were above 0.10; however disregarding their interaction with other variables would be unwise. On the other hand a full second order model is not practical, because it would leave the observation to variable ratio at less than two. In order to consider interaction, we estimated that between just original continuous independent variables with a *p*-value greater than 0.10. The variables that satisfied this rule were VITFSNUM, VITFSEMP, IN50MI, IN100MI, Pop50, and Pop25. (Rather than include the newly introduced demographic variables PCI and EMP in the interaction terms we instead continued to arbitrarily control for them simply as main effects throughout the rest of the study.) Remember an explanation of these variables can be found in Table 1. We created fifteen new variables by taking the products between each of those named above. The variable FAR139 was also removed because it has a great *p*-value in the previous model and was not continuous. Next a regression model was estimated which included the fifteen additional variables that were created in order to assess interaction. The adjusted R^2 value improved from 0.7220 to 0.7753. (This equation is recognized as equation 2 in Appendix B.)

Results

The next step in our analysis was to determine what variables contributed appreciably to the model and what variables might still be contributing to relatively inefficiency. Examination of *p*-value for each independent variable in the regression model revealed that VITFSEMP was the only variable remaining that was not statistically significant as a main effect, nor were any of the interaction terms including it. Therefore, the variable VITFSEMP and the second order variables that included VITFSEMP were removed from the model. This regression model is displayed as equation 3 in Appendix B. When the regression model was re-estimated without these variables, the adjusted R^2 value surprisingly decreased from 0.7753 to 0.7734. The dropped variables apparently contributed to the efficiency of the model in less than obvious ways, and they were retained to be included in the finally recommended regression model. A summary of the results from the final regression model as compared to that of GRA is presented in Table 3. The coefficient estimates and *p*-values of the variables used in our final model are displayed in Table 4.

Table 3. Final Model Comparison with GRA

	# Of Airports	# Of Independent Var.	R^2	R^2_{adj}
GRA's Best Model (eq. 13)	232	8	.7386	0.7292
Black-Chimka Best Model	232	29	.8036	0.7753

Table 4. Final Regression Variable's Coefficients and *P*-Values

Variable	Coefficient	*P*-value
TOWDUM	13901.43	0.000
BA	162.48	0.033
POP	-17.58	0.032
PCI	0.26	0.137
EMP	43.00	0.018
AAL	-17229.20	0.038
VITFSNUM	774.60	0.438
VITFSEMP	285.41	0.618
IN100MI	4083.09	0.914
IN50MI	33887.68	0.001
Pop100	0.002	0.000
Pop50	-0.003	0.162
Pop25	0.008	0.055
VITFSNUM * VITFSEMP	1.16	0.995
VITFSNUM * IN50MI	31407.45	0.025
VITFSNUM * IN100MI	-61379.23	0.020
VITFSNUM * POP50	-0.0005	0.038
VITFSNUM * POP25	0.0005	0.038
VITFSEMP * IN50MI	-540.39	0.753
VITFSEMP * IN100MI	-899.46	0.640
VITFSEMP * POP50	-0.00002	0.950

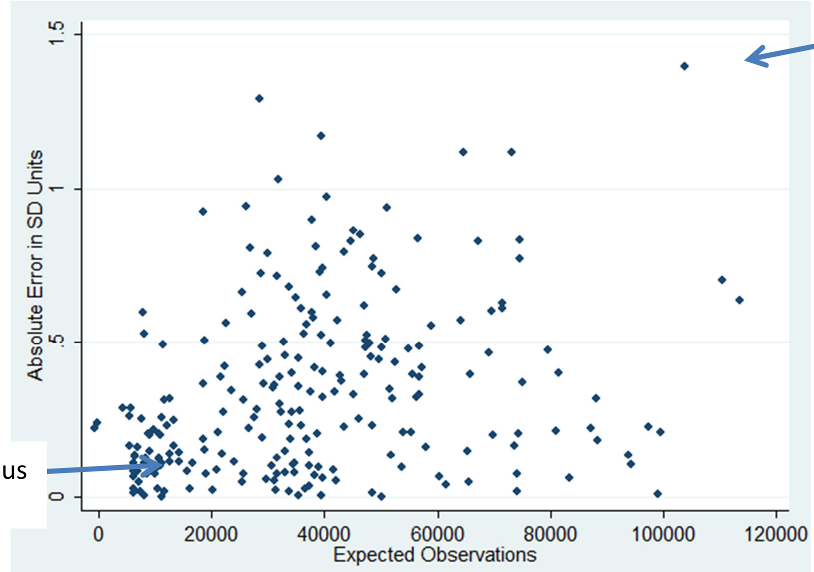
VITFSEMP * Pop25	0.000547	0.375
IN50MI * IN100MI	-64341.07	0.165
IN50MI * Pop50	-0.07	0.045
IN50MI * POP25	0.06	0.473
IN100MI * Pop50	0.344	0.000
IN100MI * Pop25	-0.290	0.109
Pop50 * Pop25	-7.73E-10	0.004
[Constant]	-6641.06	0.123

Next, face validity of the coefficients was considered in hopes to make some logical and physical sense of the model, and to confirm there is no evidence of interpretation problems related to inter-dependence among independent variables. The regional variable AAL, which is a categorical variable that represents whether an airport is located in Alaska, has a large negative coefficient. This means that if an airport is located in Alaska, then it will most likely have a very small number of total annual operations. Perhaps this makes sense because Alaska is sparsely populated. Another variable named IN50MI, which represents the percentage of based aircraft among based aircraft at GA airports within 50 miles, has a large positive coefficient. This seems consistent if we should expect more prominent airports (airports with high percentage of based aircraft at GA airports within 50 miles) to also have a large number of total annual operations. For another example the coefficient of the variable BA, for total based aircraft at an airport, has a fairly large coefficient, which seems valid since one might expect an airport with a large fleet of total based aircraft to have a large number of annual operations.

Conclusions and Future Considerations

The research conducted in this report has produced a more accurate and efficient model for estimating the annual number of operations at a GA airport. This information can of course be used to create better terminal area forecast summaries for GA airports. But more importantly it could possibly be used to detect unusual behavior based on the annual number of operations at an airport.

One future objective of this project was to create quality control charts that could be used to monitor general aviation activity. We do this by analyzing the residuals from the regression model. Figure 1 demonstrates how quality control charts could be used to monitor GA behavior. The x-axis of this figure represents the estimated total annual number of operations for a given airport according to our model. The y-axis displays how far, in standard deviations, the actual values are in comparison to the predicted values estimated by the model. Airports with a large x-value and y-value might be of greatest concern because they represent highly unusual behavior at supposedly large airports. Airports located in the lower left hand corner might be of least concern because they represent usual behavior at supposedly small airports.



Most

Dangerous

Least Dangerous

Figure 1. Model Errors versus Expectations

Another future objective of this project was to provide recommendations for multiple data stream integration applied to transportation security. Methods must be developed to improve monitoring across collaborative data sources and modes. Further improved information technology for GA could lead to even better recommendations for early detection decision aids for GA security. All of these activities would exist under with a common philosophy that if good models of usual activity fail to predict, then unusual activity may indicate a security threat. The model-based control of GA security described in this article may also be extended to other contexts such as highway, maritime transportation systems, mass transit, pipeline systems, and rail.

References

Department of Homeland Security Press Office (2008), Fact Sheet: General Aviation, *U.S. Department of Homeland Security*

FAA Office of Aviation Policy (2007), Terminal Area Forecast Summary: Fiscal Years 2007-2025, *Federal Aviation Administration*

GRA, Inc. (2001), Model For Estimating General Aviation Operations at Non-Towered Airports Using Towered and Non-Towered Airport Data, *FAA Office of Aviation Policy*

Hoekstra, M. (2000), Model for Estimating General Aviation Operations at Non-Towered Airports, Federal Aviation Administration

Homeland Security Advisory Council (2008), Top Ten Challenges Facing The Next Secretary of Homeland Security, *U.S. Department of Homeland Security*

Kulkarni, D., Wang, Y., Windrem, M., Patel, H., & Keller, R. (2003), Aviation Data Integration System, *NASA Ames Research Center*

Lubold, G. (2010), Plane crash in Austin points to vulnerabilities from small planes, *Christian Science Monitor* (February 18)

Transport Canada (2007), *General Aviation Economic Footprint – Measurement.* Roundtable discussions of the Air Transport Association of Canada, Ottawa

Transportation Security Administration (2004), Information Publication A-001: Security Guidelines for General Aviation Airports, *U.S. Department of Homeland Security*

Appendix A. Matrix of Equations Recreated from GRA, Inc.

Eq.	BA	POP100	WACAORAK	BA2	%in50mi	POP25/100	%in100mi	FAR139	Pop	SeBA/BA	TOWDUM	VITFSnum
1	1	0	0	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	0	0
3	1	1	1	0	0	0	0	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0	0	0	0
5	1	1	1	1	1	0	0	0	0	0	0	0
6	1	1	1	1	1	1	0	0	0	0	0	0
7	1	1	1	1	1	1	1	0	0	0	0	0
8	1	1	1	1	0	1	1	0	0	0	0	0
9	1	1	1	1	0	1	1	1	0	0	0	0
10	1	1	1	1	0	0	0	1	1	1	0	1
11	1	1	1	1	0	0	0	1	1	1	0	1
12	1	1	1	1	0	0	0	1	1	1	0	1
13	1	1	1	1	0	1	1	0	0	0	1	1
14	1	1	1	1	0	1	1	0	0	0	1	0
15	1	1	1	1	0	1	1	0	0	0	1	0

Appendix B. Black-Chimka Matrix of Equations

Variables	Equation 1	Equation 2	Equation 3
	$R^2_{adj} = 0.7220$	$R^2_{adj.} = .7753$	$R^2_{adj.} = .7734$
towdum	1	1	1
ba	1	1	1
pop	1	1	1
pci	1	1	1
emp	1	1	1
far139	1	0	0
aal	1	1	1
vitfsnum	1	1	1
vitfsemp	1	1	0
in50mi	1	1	1
in100mi	1	1	1
pop100	1	1	1
pop50	1	1	1
pop25	1	1	1
seba	1	1	1
vitfsnum~vitfsemp	0	1	0
vitfsnum~in50mi	0	1	1
vitfsnum~in100mi	0	1	1
vitfsnum~pop50	0	1	1
vitfsnum~pop25	0	1	1
vitfsemp~in50mi	0	1	0
vitfsemp~in100mi	0	1	0

vitfsemp~pop50	0	1	0
vitfsemp~pop25	0	1	0
in50mi~in100mi	0	1	1
in50mi~pop50	0	1	1
in50mi~pop25	0	1	1
in100mi~pop50	0	1	1
in100mi~pop25	0	1	1
pop50~pop25	0	1	1

Regression-based monitors of North American border-crossing activity

Justin R Chimka[1]

Department of Industrial Engineering, University of Arkansas, Fayetteville, AR, USA

Official border-crossing data were collected, assumed to represent usual activity, and summarized with linear regression models to facilitate residual control charts and detect unusual activity that might be attributed to a security threat or some other external variable. Particularly good results are presented for Northern highway, and loaded rail, containers at North American borders with respect to the sigma limits of statistical quality control. Additionally this manuscript describes the estimation of what might be useful model-based control carts for activity across three ports of the State of Michigan: Detroit, Huron and Sault Ste. Marie.

Key words: *Quality control, residual control charts, transportation security, border crossing, regression*

INTRODUCTION

North American border crossing and entry data were collected from the Bureau of Transportation Statistics web site, Tran Stats (www.transtats.bts.gov). They include monthly counts by mode and subject, from 1995 through 2009. For

[1] jchimka@uark.edu

example modes are aviation, maritime, highway, transit, rail, pipeline, bike / pedestrian and other. The work described here is guided by the following proposition: If good statistical models of usual activity can be found, by monitoring their errors in time we can detect unusual activity which might be attributed to a security threat or some other external variable.

The general concept of residual control charts encompasses regression adjustment developed by Hawkins (1991), and residual control charts to detect security threats have been developed for the general aviation context (Black and Chimka). In other related literature Espenshade (1995) examined how the continuous stream of undocumented migrants crossing the southern United States border is related to data on Immigration and Naturalization Service (INS) border apprehensions, while Espenshade and Acevedo (1995) showed how apprehension probabilities are determined by INS effort and number of migrants attempting to cross illegally. Weeks, et al. (2011) contributed research on the geographical origins of Mexican immigrants by creating a migration propensity index. Regression analysis by the authors finds that a Mexican state's index is predicted by the death rate from violence and accidents among men aged 20-34.

The following section describes analysis of containers by border assuming usual activity. We search for a useful regression model, such that its errors or residuals conform well to expectations associated with sigma limits of individuals control charts. Follow-up analysis of containers by port is described to show the reader that a finer, more realistic level of detail could also be monitored with residual control charts. In practice what is presented here should provide a methodological and validation framework for potential real-time monitors of border-crossing activity.

METHODS

First we chose to search for a good time series model of containers that controls for border (Northern versus Southern), mode (rail versus highway), and whether or not the container is empty. The lagged variables chosen to consider were containers last month to account for a trend, and containers this month last year to account for seasonality. Following are some details about the original main effects model assuming containers have the normal distribution and constant variance. North is 1 if the border is Northern, 0 otherwise. Rail is 1 if the mode is rail, 0 if the mode is highway. Empty is 1 if the container is empty, 0 otherwise. Last month is the number of containers crossing last month, and Last year is the number of containers crossing this month last year.

$$\begin{aligned} E\text{ (containers)} \approx\ & 4268 + 1738\text{ (north)} - 3284\text{ (rail)} - 2409\text{ (empty)} \qquad (1) \\ & + 0.80377\text{ (Last month)} + 0.014512\text{ (Last year)} \end{aligned}$$

This multiple linear regression model is statistically significant with R-squared approximately equal to 97.2% ($p \approx 0.000$), but only the lagged variables are significant ($p < 0.050$). Therefore we stratified according to the external variables north, rail and empty, and fit eight different models of the same container data controlling for lagged variables:

1. Southern containers as a function of the variables rail and empty
2. Northern containers as a function of the variables rail and empty
3. Highway containers as a function of the variables north and empty

4. Rail containers as a function of the variables north and empty
5. Loaded containers as a function of the variables north and rail
6. Empty containers as a function of the variables north and rail

Models 1, 3 and 6 immediately above are significant, but they have no significant main effects other than lagged variables, so the models are disqualified from participating in regression-based monitors.

In Model 2 the only significant external variable is empty ($p \approx 0.026$), so we decide to stratify again according to the insignificant rail variable, and fit two more models of the same Northern container data controlling for lagged variables: 1) Northern highway containers as a function of empty, and 2) Northern rail containers as a function of empty. All of the main effects in the model of Northern highway containers are significant, so we declare it a good candidate to monitor border-crossing activity. In the model of Northern rail containers only the lagged variables are significant, so it is disqualified from participating in regression-based monitors.

In Model 4 the only significant external variable is north ($p \approx 0.006$), so we decide to stratify according to the insignificant empty variable, and fit two models of the same rail container data controlling for lagged variables: 1) loaded rail containers as a function of north, and 2) empty rail containers as a function of north. All of the main effects in the model of loaded rail containers are significant, so we declare it a good candidate to monitor border-crossings. In the model of empty rail containers only the lagged variables are significant, so it is disqualified from regression-based monitors. All of the main effects in Model 5 are significant, and we declare it a good candidate to monitor activity.

RESULTS

In summary we have found three relatively good models of containers:

1. Loaded containers in general (R-squared ≈ 96.5%)
2. Northern highway containers (R-squared ≈ 95.8%)
3. Loaded rail containers (R-squared ≈ 93.7%)

Following are the expected values assuming normal distribution and constant variance.

E (Loaded containers) ≈ 14,301 + 9698 (north) – 15,865 (rail)
+ 0.79663 (Last month) (2)
+ 0.011544 (Last year)

E (Northern highway containers) ≈ 31,051 – 26,247 (empty)
+ 0.80736 (Last month) (3)
+ 0.00955 (Last year)

E (Loaded rail containers) ≈ 1740 + 9686 (north)
+ 0.79324 (Last month) (4)
+ 0.007279 (Last year)

Next we highlight two important assumptions about our multiple linear regression models: 1) They are based on data that do not indicate unusual border-crossing activity, and 2) their errors or residuals should have the standard normal distribution.

For each of the candidate linear regression models of containers we can find the absolute observed and expected errors greater than one, two and three.

These values are associated with one, two and three sigma limits of the quality control chart assuming standard deviation is equal to one.

For example among the 624 observations used to fit the model of loaded containers, where Z is the standard normal random variable, the number of associated residuals we expect to be outside of one sigma limits is approximately equal to 2 (624) P (Z > 1) or 198. However as it turns out only 115 of model errors are beyond the same limits. See Table 1 for loaded container results that indicate this regression-based monitor of loaded container data in is not very useful; the normal distribution assumption about errors for some reason does not seem good.

Table 1. Observed and expected errors beyond sigma limits associated with the loaded containers model

Sigma limits	1	2	3
Loaded containers	115	33	19
Expected errors	198	28	2

In Table 2 we have present similar results for the more specific models of Northern highway containers and loaded rail containers, and hope to find models of more constrained data will also be more useful to monitor border-crossing activity. The same number of observations (312) was used to fit each these models.

Table 2. Observed and expected errors beyond sigma limits associated with the Northern highway and loaded rail containers models

Sigma limits	1	2	3

Northern highway containers	97	10	2
Loaded rail containers	91	14	2
Expected errors	99	14	1

Results seem to indicate that regression analysis of the more constrained data provide a relatively good fit between them and monitors of border-crossing activity. In other words an unusual number model errors compared to associated expectations or predictions about Northern highway and loaded rail containers would seem to indicate unusual border-crossing activity which might be attributed to a security threat.

DISCUSSION

We followed up analysis of containers by border with similar analysis of specific ports Detroit (Detroit = 1, Huron = 0), Huron (Detroit = 0, Huron = 1), and Sault Ste. Marie (Detroit = 0, Huron = 0), Michigan, chosen for their relation to the US Customs and Border Protection SBI Northern Boarder Project (dummy variable values to describe the port of interest). We again have empty = 0 to described loaded containers, rail = 0 to describe highway traffic, and lagged variables to describe relevant border-crossing activity last month, and this month last year. In the original main effects model only the lagged variables were significant, so this time we fit a full second order model to investigate interaction. Unfortunately inconsistencies between the main effects and full second order models suggested problematic inter-dependence, so we estimated the variance inflation factors associated with each of the main effects. Since the variable to describe relevant border-crossing activity last year was the most inter-dependent it was dropped from the analysis, and the main effects model was refit. This time without variance inflation each of the main effects was significant: Detroit, Huron, empty, rail, and last month. Future research would include diagnostics related to monitoring and detection like those presented in Tables 1 and 2 in the analysis of containers by border.

In summary we have presented a system for estimating models of crossings used to identify unusual activity by border that might be attributed to a security threat or another external variable. A finer application to crossings by port showed how the regression-based monitor concept might be applied in a more realistic setting. Even more practical would be comprehensive port-level analysis of data collected daily or even weekly as opposed to monthly, but unfortunately they are not freely available.

ACKNOWLEDGMENT

This material is based upon work supported by the United States Department of Homeland Security (DHS) under award number 2008-ST-061-TS0003. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies either expressed or implied of the US DHS. Thanks to Morgan Ulesich and Ryan Black for assistance with the regression analysis.

REFERENCES

Espenshade, TJ (1995). “Using INS border apprehension data to measure the flow of undocumented migrants crossing the U.S.-Mexico frontier,” International Migration Review 29(2): 545-565

Espenshade, TJ and D Acevedo (1995), “Migrant cohort size, enforcement effort, and the apprehension of undocumented aliens,” Population Research and Policy Review 14(2): 145-172

Hawkins, DM (1991), “Multivariate quality control based on regression adjusted variables,” Technometrics 33(1): 61-75

Weeks, JR, J Stoler and P Jankowski (2011), “Crossing the border: New data on undocumented immigrants to the Unites States,” Population, Space and Place 17(1): 1-26