Data Audit for the Maricopa County Sheriff's Office, Years 2014-2015

Prepared for Maricopa County Sheriff's Office Report

By

Danielle Wallace, PhD
Charles Katz, PhD
Vince Webb, PhD
Courtney Riggs, MS
Richard Moule, Jr., MS

October 2015

About the Center for Violence Prevention & Community Safety

Arizona State University, in order to deepen its commitment to the communities of Arizona and to society as a whole, has set a new standard for research universities, as modeled by the New American University. Accordingly, ASU is measured not by whom we exclude, but by whom we include.

The University is pursuing research that considers the public good and is assuming a greater responsibility to our communities for economic, social, and cultural vitality. Social embeddedness – university-wide, interactive, and mutually-supportive partnerships with Arizona communities – is at the core of our development as a New American University.

Toward the goal of social embeddedness, in response to the growing need of our communities to improve the public's safety and well-being, in July 2005 ASU established the Center for Violence Prevention and Community Safety. The Center's mission is to generate, share, and apply quality research and knowledge to create "best practice" standards.

Specifically, the Center evaluates policies and programs; analyzes and evaluates patterns and causes of violence; develops strategies and programs; develops a clearinghouse of research reports and "best practice" models; educates, trains, and provides technical assistance; and facilitates the development and construction of databases.

For more information about the Center for Violence Prevention and Community Safety, please contact us using the information provided below.

MAILING ADDRESS

Center for Violence Prevention and Community Safety
College of Public Programs
Arizona State University
Mail Code 3120
411 N. Central Ave., Suite 680
Phoenix, Arizona 85004-2115

PHONE

602.496.1470

WEB SITE

http://cvpcs.asu.edu

Contents

1. Data Audit

The purpose of this data audit is to assist the Maricopa County Sheriff's Office in assessing the quality of their TraCs data and assist in developing and maintaining high data quality. Regular examination of data quality enables any future policy and training recommendations to be based on the best quality data that is possible. Without data quality, results from any analyses are seen as questionable.

There are several goals of this data audit. The first is to do a general wellness check of the data, and discern what problems, if any, are contained in the data. Generally, typical problems include missing and invalid data. The second goal is to diagnose these problems and determine where they are coming from; for instance, are the problems originating from a particular district, beat, or deputy or perhaps is it a flaw in the data collection system? The final goal of the data audit is to provide the Maricopa County Sheriff's Office with a clear picture of any problems with their data generally and offer solutions to those problems.

2. Summary of the Data

The data employed in the audit encapsulates one year of deputy initiated traffic stops by Maricopa County Sheriff's Office (MCSO) deputies ranging from July 1, 2014 to June 30, 2015. While MCSO had other calls for service during this period, this data includes only deputy initiated stops, which is the proper unit of analysis for discerning any racial bias or profiling involved in traffic stops. There are two data sources employed in this data audit. The first is CAD data – or data coming from dispatch. The second data source is the TraCs data, which includes the vehicle stop contact form data established as a part of the consent decree. A vehicle stop contact form is used by deputies to collect information about each traffic stop beyond what is collected in each citation, long form, incidental contact report, or warning. In the TraCs data, information is collected about the incident, driver, passenger(s) if there are any, and location of the traffic stop. For ease of reporting, this report will refer to the above datasets collectively as the "TraCs" data for the remainder of the report.

3. General Issues with the Data

The audit revealed several general problems with the TraCs data. Below the details of each problem are presented followed by suggestions and solutions to help remedy the problems.

3.1 Duplicate Event Numbers

The first problem the audit revealed was related to the Event Number variable, which is meant to be an identifying variable for each traffic stop. Typically, identifying variables enable each case, or here each traffic stop, to be uniquely identified. Duplicate event numbers are problematic because without the deletion of duplicate traffic stops, these traffic stops potentially have more

influence on results. Duplicate event numbers typically occurred when more than one vehicle involved in a traffic stop.

While duplicate event numbers are problematic, there is an alternative means of identifying traffic stops as unique. The PrdKey variable, which is a variable created in MCSO's data management system, can be used as an alternative traffic stop identifier. Thus, rather than using Event Number, we suggest using the PrdKey variable to identify traffic stops.

3.2 Issues with GPS in the TraCs system

Of the 28,148 traffic stops in the yearly data, 3,374 or approximately 12% had missing GPS coordinates that originated from the TraCs system. One way of addressing the issue of missing GPS coordinates is to geocode the stop location that is self-reported by the deputy. Using this information, 1,150 additional traffic stops or about 35% of those traffic stops with missing GPS coordinates can now be given GPS coordinates. While this procedure might be helpful where addresses and intersections can easily be identified and recorded, using the deputy reported location of the stop to geocode is less viable when the traffic stop is located in mountainous or more rural areas. For example, when a deputy reports "SR85 AND MILE POST 154" or "USERY PASS MP22" as a stop location, ArcGIS or Google Maps generally does not recognize these as locations.

3.3 Solutions

Next we discuss two potential solutions for these general problems with the yearly data.

3.3.1 Technology Issues in TraCs

An immediate solution for missing GPS data in the TraCs system is to use GPS data coming from the CAD/RMS system or dispatch. This data have a very low missing rate for GPS coordinates, specifically only 3%. MCSO is currently working with the provider of the GPS system to understand why capturing GPS coordinates is so difficult and variable. There are other ways, however, of assuring GPS locations. We discuss this below.

3.3.2 Stress Good Reporting on Traffic Stop Location

When GPS coordinates are not available, the deputy reported stop location becomes critical for determining the geographic location of a stop. Thus, good reporting of the traffic stop location should be stressed to deputies. Here, good reporting would include clear typing, avoiding misspellings, adding directions such as North, West, East and South to the location information, and generally avoiding uncommon abbreviations.

4. Missing Data

Missing data is a common, but significant issue for any law enforcement agency. It is also a complex problem to solve given that missing data can come from many sources. Missing data

may result when reports are not finalized, entries are not finished, technological issues, or from deputy misunderstanding of or unwillingness to compete all field entries. Missing data limits both the types of analyses that can be done as well as the quality of those analyses. Representing what is actually happening in traffic stops requires complete or nearly complete data.

Before detailing the missing data in the TraCs data system, it is important to describe how percentages of missing data were calculated and who it was calculated for. The total percentage of missing data in the traffic stop data by unit was calculated for only those fields in the traffic stop data form that the deputy was responsible for inputting; drag and drop fields and information that could be imported from other forms during the traffic stop were not included in the analysis. Furthermore, this calculation did not include "conditional" fields. These are fields that are dependent on another field; for example, whether or not the driver was Terry searched is dependent on whether or not the driver was searched at all. These conditional fields were excluded from these analysis in order to gain a sense of the primary problem of missing data (versus other missing data issues). Thus, this percentage gives a clear picture of missing information due to the deputy, and not due to the TraCs system or other technologically based problems. Also, this percentage was only calculated for deputies that conducted, on average, ten traffic stops a month. This enables the determination of a pattern of missing data, rather than a fluke occurrence among deputies.

4.1 Missing Data at the Organization level

In brief, missing data is a problem that needs to be addressed. It is generally accepted that for data to be regarded as high quality, only 5% of the data can be missing (Engel et al. 2009; Engel, Cherkauskas and Smith 2008; Engel et al. 2007; Fridell 2004). Table 1 shows that for the first year of data collection, at the organizational level, there were no months where MCSO was beneath a 5% missing data threshold. The range of missing data was as low as 10.6% in June 2014 and as high as 11.54% in July 2015. Figure 1 shows the trend of missing data over time. Here we see that while missing data is a problem, it is a decreasing problem.

As a comparison point, in 2006, the Department of Public Safety (DPS) in Arizona began in intensive investigation into racially biased policing within their organization (Engel et al. 2009; Engel, Cherkauskas and Smith 2008; Engel et al. 2007). AZ DPS uses the same TraCs system as MCSO. During their first year of internal evaluation, AZ DPS reported over a 14% error rate (Engel et al. 2007). Soon, their error rates in stops reduced to 10.4% (Engel, Cherkauskas and Smith 2008) for all data and within the final months of the evaluation, their TraCs data errors were just over 2% (Engel et al. 2009). Put in the above context, MCSO is doing well in regards to overall missing data, though the organization has room for improvement.

Figure 1. % of Data Points Missing

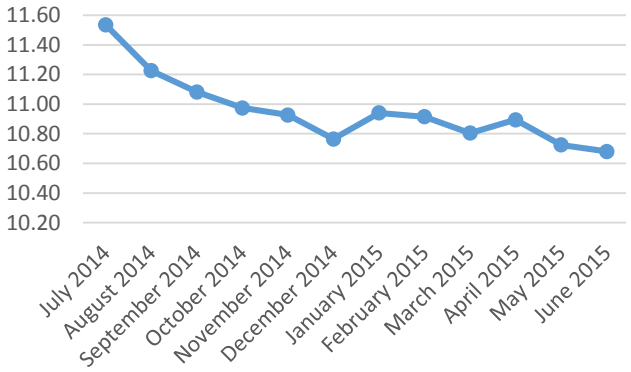

Table 1. % of Missing Data across MCSO by Month

Month	Stops	% of Data Points Missing
July 2014	2544	11.54
August 2014	2538	11.23
September 2014	1904	11.08
October 2014	1835	10.97
November 2014	1917	10.93
December 2014	2733	10.77
January 2015	1964	10.94
February 2015	1903	10.91
March 2015	2188	10.81
April 2015	2496	10.89
May 2015	3234	10.72
June 2015	2892	10.68

4.2 Missing Data among Deputies

Table 2. Top 10 Deputies for the Total % of Missing Data

Deputy Serial Number	Stops	% of Data Points Missing
S0482	310	12.07
S1934	272	11.92
S1681	247	11.81
S1818	128	11.59
S1955	172	11.48
S1905	171	11.45
S1250	245	11.43
S1644	128	11.43
S1294	124	11.42
S1996	130	11.38

The audit showed that some of the missing data problem revolves around specific deputies. It is important to keep in mind that the analysis is focused on those deputies that conduct the most stops (i.e., 10 or more per month). When examining missing data among deputies over the course of the first year of data collection, the average percent missing data per deputy is 10.8%, with the lowest yearly percent of missing data by a deputy being 9.26% and the highest being 12.0%. Put simply, no deputy was within the 5% threshold for data quality. Furthermore, there is variability among deputies with some having higher percentages of missing data than others. Note though that while deputies varied in the amount of missing data in their traffic stops, they did not vary greatly. As an example, Table 2 lists the 10 deputies with the highest percentage of missing data, ranked from highest to lowest; the full table of missing data for all deputies is in Appendix A. Remember that the range of missing data by deputy is between 9.26% and 12.0%; thus, missing data seems to be problem for all deputies not just a select few.

4.3 Missing Data by Beats

Below are the results of the audit with respect to missing data by beat. This analysis was restricted to beats that had at least 10 traffic stops. The analysis of missing data across beats over the course of the first year of data collection shows that the average percent missing data is just over 11%. The lowest percentage of yearly missing data for a beat being 9.64% and the highest is 11.78%. These numbers are very similar to deputy levels of missing data. Table 3 shows the 10 beats with the highest percentage of missing data. The full table of missing data by beat is in Appendix B. Notably, these beats are not those with the highest volume of stops; for example, Lake Partrol (LAK) saw 2059 stops with a missing data percentage of 9.64%, which is lower than any of the ten beats listed in Table 3. Thus, work load and stop volume do not seem to be a component of missing data and data accuracy.

Table 3. Top 10 Beats for Missing Data

Beat	Stops	% of Data Points Missing
344	545	11.78
224	498	11.58
743	180	11.57
371	468	11.52
652	153	11.47
742	216	11.46
342	704	11.41
432	905	11.38
433	1264	11.36
231	405	11.34

4.4 Missing Data by Districts

Table 4. Districts by Missing Data

District	Stops	% of Data Points Missing
1	3892	11.03
2	4767	10.93
3	3709	11.12
4	3963	11.11
6	2694	11.21
7	2140	11.05
Lake Patrol	6519	10.57
Enforcement Support	27	10.49
SWAT or K9	240	11.18
Special Investigations	125	9.77
Major Crimes	61	11.00

Like deputies and beats, missing data is also concentrated in some districts more than others. Table 4 displays the percentage of missing data across stops and the percentage of stops with missing GPS coordinates by district. For all districts there is between 9.8 and 11.2% missing data. Thus, all districts have exceeded the 5% threshold for acceptable levels of missing data.

4.5 Solutions

Missing data in the TraCs system, for the most part, is a deputy based problem that aggregates to larger units. As such, the solutions discussed below focus primarily deputies.

4.5.1 Mandatory Reporting of all Relevant Fields in TraCs forms

Perhaps the simplest way of alleviating the issue of missing data is to make all the fields in the TraCs form mandatory for deputies to fill out. If this cannot be accomplished, then solution 4.5.2 becomes important. However, at the time of reporting, MCSO is moving towards making the majority of fields in the TraCs system mandatory to report.

4.5.2 Training and Assistance for All Deputies with Missing Data

There needs to be trainings, communication, and assistance provided to deputies with high rates of missing data in their stop reports. Reducing the amount of missing data to under 5% will substantially increase data quality. Furthermore, problems at the deputy level translate up to higher units, like beats and districts. Thus, data quality at the beat and district level cannot be increased until deputies reduce their percentages of missing data.

5. Invalid Data

Data that is not consistent with the operationalization (i.e., how something is intended to be measured) of a variable is problematic. For instance, several important variables have data values that are unable to be coded. An example of this is the Agency variable that documents the district where the traffic stop took place. The audit found nine traffic stops have a value that does not fit a specific district. While this is a trivial amount, reformatting the district to something that is usable takes effort and considerable time among the data analysis teams at MCSO. Invalid data entries also undermines analyses, as invalid values are as equally unusable as missing data.

5.1 Invalid Data among Deputies

The audit showed that only a few deputies have invalid data entry issues. Common variables to have invalid data include the birth date of the driver, the license plate information of the vehicle, and the district the stop occurred in (see Appendix C for a full listing of Deputies by invalid data). For all these variables, bad vehicle license plate information was the most problematic. Table 5 lists the ten deputies with the greatest percentage of stops with an invalid vehicle license plate entry. These ten deputies together account for approximately 39 traffic stops with problematic vehicle license plate entries. While this remains a problem, it is a substantial improvement over missing data.

Table 5. Top 10 Deputies for Invalid Driver Date of Birth Entries

Deputy Serial Number	Stops	% of Stops with an Invalid Vehicle License Plate Entry
S1946	253	1.94
S1818	176	1.85
S1645	184	1.84
S1967	272	1.70
S1936	310	1.69
S1777	233	1.38
S1987	411	1.19
S1768	201	1.09
S1845	489	1.08
S1944	188	1.03

While there were invalid data entries related to the birth date of the driver and the district the stop occurred in, however, they were relatively minimal. The range of invalid data on the birth date of the driver spans from a low of 0% to a high of 0.79%. . The range of invalid data the district the stop occurred in has a low of 0% and a high of 0.53%. Thus, while the entry of data by deputies has its problems, it is unlikely that it impacts data quality to a vast degree.

5.2 Solutions

The problem of invalid data is substantially smaller in magnitude than the problem of missing data. Still, it is a problem that needs to be addressed. The solutions described below might be useful in facilitating better data capture and entry on the part of the deputies.

5.2.1. Increased Auto-Population of Forms

Some variables that have errors in them, such as the district variable, could potentially be auto-populated in the TraCs system or have a pull down menu. This would significantly reduce error in the data. Additionally, this would be helpful in addressing missing data.

5.2.2 Training for Deputies with Chronic Invalid Data

There are a small number of deputies who provide invalid data. Similar to issues of missing data, when a deputy is chronically misreporting information, it compromises data validity. A warning system should be in place to flag deputies who repeatedly provide invalid data. This warning system may enable deputies to correct the issue themselves if they are flagged early, for instance say when they reach 2.5-4.9% invalid data entries. However, deputies who exceed 5% invalid entries should receive training on the importance of quality data as well as on technical issues related to capturing and reporting valid data.

6. Next Steps

When compared to other agencies like AZ DPS that are using the TraCs software, MCSO is performing well regarding data quality with respect to missing and invalid data in their first year of internal investigation. There is substantial room for improvement, however, particularly among deputies. Below, several solutions and recommendations are detailed to assist MCSO in obtaining lower rates of missing or invalid data and increasing data quality.

- The level of missing data in MCSO is in large part due is to deputies. Many of these issues may be technology based. It is recommended that more intensive feedback, training and assistance be made available to deputies on the use of the TraCs system.
- As next step, a flow chart of how data enters the TraCs system by the deputy and how that information eventually turns into data downstream should be constructed. Such a flow chart would be helpful in trouble shooting any future problems.
- Lastly, MCSO should consider a broad range of strategies and tactics to address problems associated with quality of data. In doing so, thorough records should be retained to detail the methods and frequency of intervention MCSO has proscribed so that it can assess which strategies are most effective in addressing data quality.

7. References

Engel, Robin S., Jennifer Calnon Cherkauskas, Michael R. Smith, Dan Lytle, and Kristan Moore. 2009. "Traffic Stop Data Analysis Study: Year 3 Final Report." Arizona Department of Public Safety.

Engel, Robin S., Jennifer Calnon Cherkauskas, and Michael R. Smith. 2008. "Traffic Stop Data Analysis Study: Year 2 Final Report." Arizona Department of Public Safety.

Engel, Robin S., Rob Tillyer, Andrew J Cherlin, and James Frank. 2007. "Traffic Stop Data Analysis Study: Year 1 Final Report." Arizona Department of Public Saftey.

Fridell, Lorie A. 2004. "By the Numbers: A Guide of Analyzing Race from Vehicle Stops." Washington, DC: Police Executive Research Forum.

Appendix A. Missing Data by Deputy for June 2014 to July 2015

Deputy Serial Number	Stops	% Missing Data Points
S0482	310	12.07
S0793	233	10.89
S0920	411	11.29
S0933	201	10.34
S1005	489	10.02
S1036	188	10.97
S1176	147	10.52
S1210	464	10.49
S1250	245	11.43
S1293	217	10.83
S1294	124	11.42
S1315	130	10.90
S1381	121	10.78
S1428	144	10.85
S1468	162	10.49
S1484	184	9.83
S1521	171	10.92
S1609	513	10.78
S1616	631	10.99
S1642	192	10.72
S1644	128	11.43
S1645	170	10.51
S1681	247	11.81
S1691	131	11.07
S1714	121	9.88
S1727	224	10.99
S1747	300	9.26
S1768	292	10.89
S1770	206	9.55
S1777	262	10.35
S1779	226	10.08
S1782	139	10.04
S1799	334	9.57
S1818	128	11.59
S1834	125	11.30
S1841	147	11.05

Deputy Serial Number	Stops	% Missing Data Points
S1845	228	11.04
S1868	256	10.82
S1872	178	11.19
S1880	160	11.22
S1895	401	10.17
S1905	171	11.45
S1908	124	9.98
S1934	272	11.92
S1935	438	11.36
S1936	147	10.66
S1937	493	11.04
S1938	122	10.76
S1940	221	10.71
S1942	159	10.59
S1944	141	10.82
S1946	251	10.74
S1949	304	11.35
S1951	398	10.72
S1952	166	10.72
S1955	172	11.48
S1967	166	11.14
S1968	141	11.08
S1970	444	10.82
S1978	291	10.77
S1986	216	10.30
S1987	137	10.58
S1993	239	11.18
S1995	176	10.63
S1996	130	11.38
S1999	177	10.40
S2002	253	10.87
S2003	388	10.55
S2004	170	11.13
S2007	134	10.54

Appendix B. Missing Data by Beat for June 2014 to July 2015

Beat	Stops	% Data Points Missing
121	744	11.27
122	809	11.07
123	274	10.86
124	294	11.03
125	787	10.83
126	796	10.68
127	1903	10.90
128	189	11.24
221	776	11.01
222	707	11.14
223	252	11.08
224	498	11.58
225	1242	11.00
231	405	11.34
232	357	10.55
234	359	10.24
235	843	10.17
236	18	10.42
341	443	11.16
342	704	11.41
343	535	11.04
344	545	11.78
345	772	10.99
346	402	10.60
347	314	10.47
351	82	10.06
371	468	11.52
432	905	11.38
433	1264	11.36
434	1279	10.99
435	396	11.16
436	932	11.09
651	1001	11.18
652	153	11.47
653	1121	11.15
654	522	11.33
741	1176	11.00
742	216	11.46
743	180	11.57

Beat	Stops	% Data Points Missing
744	311	11.03
745	41	11.08
LAK	2059	9.64
PNL	86	11.00
PRK	462	10.39

Appendix C. Invalid Data by Deputy for June 2014 to July 2015

Deputy Serial Number	Stops	% Invalid State License Plate	% Invalid Driver's DOB	% Invalid District
S0482	228	0.39	0.00	0.00
S0793	124	0.81	0.00	0.00
S0920	245	0.83	0.00	0.00
S0933	438	0.00	0.00	0.53
S1005	170	0.62	0.00	0.00
S1036	141	0.00	0.00	0.00
S1176	128	0.63	0.00	0.00
S1210	256	0.38	0.00	0.00
S1250	292	0.52	0.00	0.00
S1293	171	0.71	0.00	0.00
S1294	130	0.80	0.00	0.00
S1315	147	0.99	0.00	0.00
S1381	206	0.50	0.00	0.00
S1428	304	0.00	0.00	0.00
S1468	239	0.00	0.00	0.00
S1484	171	0.30	0.54	0.00
S1521	493	0.00	0.00	0.00
S1609	631	0.68	0.00	0.00
S1616	130	0.00	0.00	0.00
S1642	177	0.00	0.00	0.00
S1644	334	0.44	0.00	0.00
S1645	184	1.84	0.00	0.00
S1681	125	0.41	0.00	0.00
S1691	388	0.00	0.00	0.00
S1714	139	0.45	0.00	0.00
S1727	224	0.58	0.00	0.00
S1747	121	0.79	0.00	0.00
S1768	201	1.09	0.00	0.00
S1770	444	0.00	0.00	0.00
S1777	233	1.38	0.00	0.00
S1779	122	0.00	0.00	0.00
S1782	178	0.34	0.00	0.00
S1799	300	0.56	0.00	0.00
S1818	176	1.85	0.00	0.00

Deputy Serial Number	Stops	% Invalid State License Plate	% Invalid Driver's DOB	% Invalid District
S1834	166	0.00	0.00	0.00
S1841	170	0.00	0.00	0.00
S1845	489	1.08	0.00	0.00
S1868	159	0.00	0.00	0.00
S1872	128	0.42	0.00	0.00
S1880	251	0.00	0.00	0.00
S1895	124	0.25	0.00	0.00
S1905	398	0.00	0.00	0.00
S1908	172	0.00	0.00	0.00
S1934	141	0.00	0.37	0.00
S1935	134	0.00	0.00	0.00
S1936	310	1.69	0.00	0.00
S1937	137	0.00	0.00	0.00
S1938	166	0.00	0.00	0.00
S1940	144	0.77	0.00	0.00
S1942	464	0.88	0.00	0.00
S1944	188	1.03	0.00	0.00
S1946	253	1.94	0.00	0.00
S1949	121	0.59	0.00	0.00
S1951	247	0.61	0.00	0.00
S1952	131	0.59	0.00	0.00
S1955	221	0.00	0.00	0.00
S1967	272	1.70	0.00	0.00
S1968	147	0.40	0.00	0.00
S1970	216	0.00	0.00	0.00
S1978	291	0.00	0.00	0.00
S1986	160	0.33	0.00	0.00
S1987	411	1.19	0.00	0.00
S1993	226	0.46	0.00	0.00
S1995	262	0.49	0.57	0.00
S1996	401	0.32	0.00	0.00
S1999	147	0.00	0.00	0.00
S2002	192	0.68	0.79	0.00
S2003	162	0.75	0.00	0.00
S2004	513	0.69	0.00	0.00
S2007	217	0.82	0.00	0.00