Paper 3202-2015

Using SAS® Mapping functionality to measure and present the Veracity of Location Data

Richard J Self, Vishal Patel, Daniel Corah, Viktor Horecny, University of Derby

ABSTRACT

Crowd sourcing of data is growing rapidly, based on smart devices equipped with assisted GPS location tagging of photos and mapping other aspects of the users' lives and activities. A fundamental assumption is made when such data is used that the reported locations are accurate within the usual GPS limitations of approximately 10m.

However, as a result of a wide range of technical issues, it turns out that the accuracy of the reported locations is highly variable and cannot be relied on. Some locations are accurate but many are highly inaccurate and this can affect many of the decisions that are being made from the data. As an example of the use of location data, some shops are now tracking their customers in order to offer suitable messages and adverts when they detect that the customer is close to the shop in order to entice them into the shop.

An analysis of a set of data will be presented that demonstrates that this assumption is flawed and will provide examples of the levels of inaccuracy that has significant consequences in a range of contexts.

The paper will demonstrate the quality and veracity of the data and the scale of the errors that can be present using Base SAS®.

This analysis has critical significance in fields such as mobile location based marketing, forensics and law.

INTRODUCTION

Location Based Services (LBS) is a rapidly growing field, with major conferences addressing the topic. However, whilst the proponents claim that GPS is capable of 10m accuracy and is reliable, it turns out that this is not the case. Quarterly reports by ThinkNear (2015) indicate that this is a flawed assumption with only 37% of locations being within 100m and, at the other extreme, 8% being accurate to between 10km and 1900km.

Indeed personal experience with location tagged photos and the Maps+ iPhone app, demonstrate that the errors can be extreme and often suggest spurious activity.

As an illustration, this images demonstrates location errors of the order of 22km.

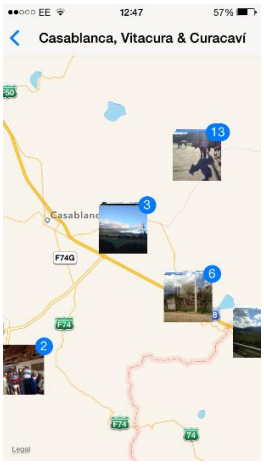

Figure 1 - Extreme location Errors

The set of 2 photos at the bottom left of the Fig. 1 were taken from the identical spot as the set of 13 at the top right.

Fig. 2 demonstrates very significant errors in location tracking in that the dog-leg to the bottom of the image was entirely an artefact of the systems responsible for collecting the location data and then plotting it. The phone was actually stationary at the time inside a large store, whereas the tracking App shows a walk of approximately 300m.

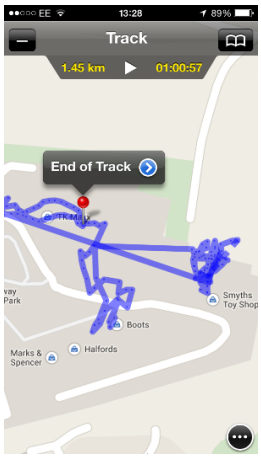

Figure 2 - Spurious Activity

Fig. 3 illustrates the instability of Location Services. All 18 photos were taken from the same place over a period of 2 minutes, the maximum error is of the order of 500m.

Figure 3 - Spurious Movement

As a result of these data, the lead author commissioned a group of 12 final year students to focus their final year dissertation projects on characterizing the degree of inaccuracy of location services. They are each collecting location data under a range of conditions and with a variety of smart devices. Many of the students are using SAS® and JMP® to analyze the accuracy and error levels.

This paper uses data from three of the students, the joint authors, in order to demonstrate a range of analytical and visualization approaches to understanding the accuracy and levels of confidence in location services.

The tools used include both standard SAS® and SAS JMP® statistics procedures, such as Procs TTEST, FREQ and UNIVARIATE, together with a range of the graphics procedures, such as SGPLOT and the lead author is also using GMAP, GREDUCE, GPROJECT and GANNO. Much of the student output was directed to the ODS output streams, where the ability to save the output stream to the .mht file format provides an excellent source of tables and images for copying and pasting into word-processed documents, once the .mht files are opened in Internet Explorer.

EXAMPLE OF ANALYSES AND VISUALIZATIONS

The following analyses have been produced by the three student authors. Additional mapping visualizations and analyses will be presented during the lead author's presentation.

ANALYSES BY V PATEL (SAS)

The analysis is based on some 200 sets of Location tagged photos collected at accurately known locations (based on use of Google Maps), using an iPhone 5S and a Nexus 6. An additional 200 sets of data was collected using a specialist APP which identified the current GPS location, and in the android version, identified the number of GPS satellites visible. The EXIF data from the phones was extracted from the photos.

The GEODIST function was used to calculate the error between the known location and the measured location, based on the

STATISTICAL ANALYSES

Procs Univariate and Ttest were used to produce a range of statistical analyses of the accuracy of the location services

Using all the data, the statistics demonstrate with a high degree of confidence that the aggregated data from both phones demonstrates a significant degree of difference (Pr > |t| <0.0001) between indoors and outdoors means and profiles. The difference between the mean errors is a factor of three.

The TTEST Procedure

Variable: error (error)

Indoor_Outdoor	N	Mean	Std Dev	Std Err	Minimum	Maximum
Indoor	116	64.6165	84.8614	7.8792	2.6500	552.5
Outdoor	284	21.9058	21.6400	1.2841	0.3000	220.9
Diff (1-2)		42.7108	49.1304	5.4137		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	398	7.89	<.0001
Satterthwaite	Unequal	121.16	5.35	<.0001

Table 1 - Proc Univariate Statistics – Indoor vv Outdoor Mean Error

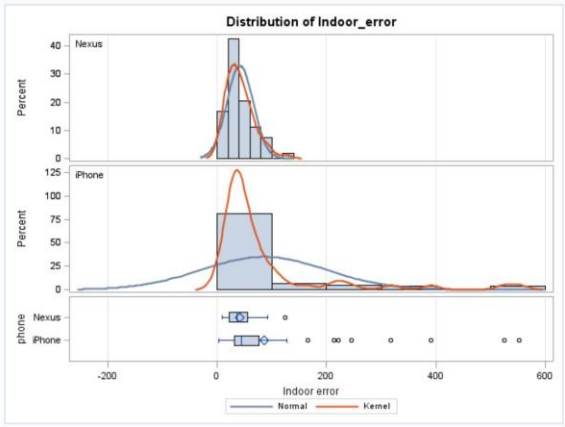

Figure 4 - Indoor - Outdoor Profiles - TTEST

Fig 4. Illustrates a significant frustration with the default settings in using very different “bin” sizes for the two profile plots which makes it very difficult to determine the real profile and also the fact that the Normal and Kernel curves extend into negative territory.

Comparing the accuracy of the two phones in the indoor condition demonstrates that there is a significant difference in accuracy between the two phones, as seen in the following output.

The TTEST Procedure

Variable: Indoor_error

phone	N	Mean	Std Dev	Std Err	Minimum	Maximum
Nexus	54	41.5629	24.1146	3.2816	9.5050	124.7
iPhone	58	85.5101	113.8	14.9403	2.6500	552.5
Diff (1-2)		-43.9472	83.5987	15.8088		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	110	-2.78	0.0064
Satterthwaite	Unequal	62.476	-2.87	0.0055

Table 2 - Indoor Comparison - Nexus vv iPhone

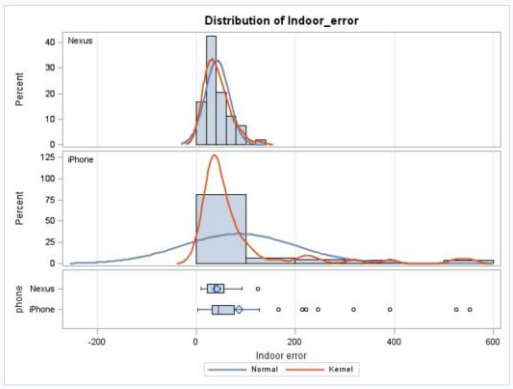

Figure 5 - Indoor Error - Nexus vv iPhone

Using an Android API to identify the number of GPS satellites that are visible, this set of data disproves the general assumption that accuracy improves until 8 are visible, after which the accuracy stabilizes at the most accurate level. This set of data suggests the contrary.

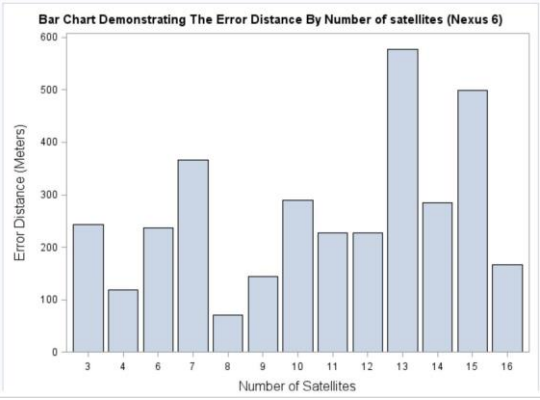

Figure 6 - Mean error with Satellite Numbers

ANALYSES BY D CORAH (SAS)

This student collected 200 sets of photos at precisely known locations using an iPhone 5S with the wi-fi connection switched off and using only 3G cell connections. This is a recommended option for smartphone users for security purposes. This experiment is, therefore, representative of users with a high degree of security awareness and practice. The location data was extracted from the EXIF data and processed in SGPLOT.

The following Proc SGPLOT vertical bar chart plots the mean error in meters, of the measured location, in a variety of locations (both indoors and outdoors) in 11 different villages and towns in Derbyshire in England. It can be seen that the locations in Duffield are comparatively accurate, whilst Ambergate provides a mean error of more than 1km, this is an isolated village with stone built houses.

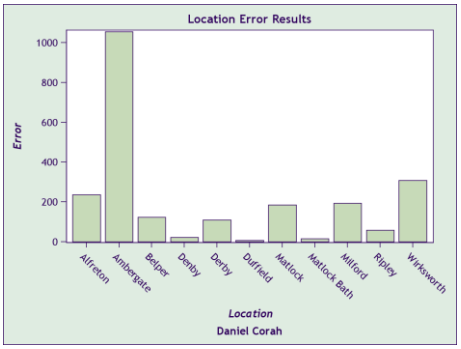

Figure 7 - Mean Error (m) by Location

The following Proc SGPLOT characterises the difference in accuracy for indoors and outdoors. It shows that outdoor location accuracy is of the order of 30m, compared to the claimed 10m accuracy capability, and that indoor accuracy for an iPhone 5S is of the order of 200m.

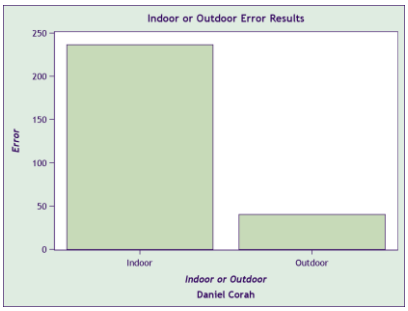

Figure 8 - Comparison of Locations (m)

ANALYSES BY V HORECNY (JMP)

Data was collected using two different HTC smartphones, the Desire S and the M8. The location errors were calculated using MS Excel and the data visualized using both Excel and SAS JMP. For the purposes of the analysis two outliers, where the error was of the order of 500m to 1000m, were excluded from the following charts.

The most interesting visualization was generated in SAS JMP comparing the accuracy in three different locations for the two phones. Whilst the profile of accuracy was similar for two types of location (rural and urban), the profiles for the two phones was dramatically different indoors.

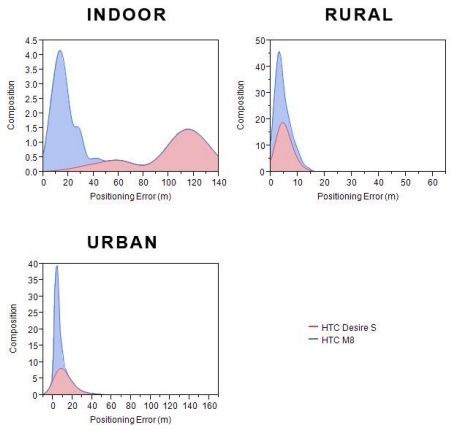

Figure 9 - Error profiles for M8 and Desire 8 (SAS JMP)

The overall profile for the indoors for the HTC Desire S is shown below using Excel graphics.

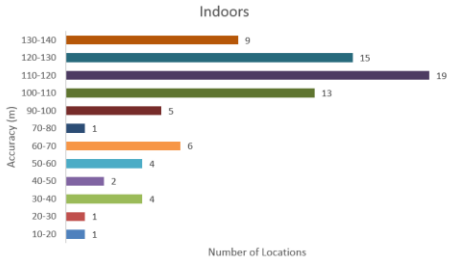

Figure 10 - HTC Desire S Indoors Profile

Comparison with the Urban and Rural accuracy profiles provides an interesting perspective of the impact of factors such as multi-path reflection in the urban setting, compared to the rural setting.

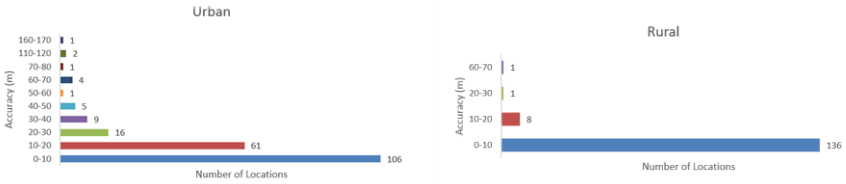

Figure 11 - HTC Desire - Urban and Rural Error Profiles

CONCLUSION

The work by the student authors demonstrates the value of using SAS based analytical techniques to evaluate the extent of the accuracy of location services on a range of current smartphones. The analyses demonstrate, with high degrees of statistical certainty that the actual accuracy is considerably less in many circumstances that that claimed by the proponents of LBS.

There are significant consequences for businesses and crime detection for this impaired accuracy. The questions posed by the 14 Vs of Big Data can provide a clear governance framework for the analysis that is required in order to ensure that value is gained from LBS.

The analyses carried out so far, in these projects, strongly suggests that there are additional factors which have not been captured and could form the basis of further research.

REFERENCES

ThinkNear (2015) **Location Score Index**, http://info.thinknear.com/mobile-advertising-location-data-accuracy (accessed 26 March 2015)

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Richard J Self
University of Derby
r.j.self@derby.ac.uk
http://computing.derby.ac.uk/wordpress/people-2/richard-j-self/
http://uk.linkedin.com/pub/richard-self/a/93a/829