Paper 3860-2019

SAS/ETS® Proc Arima used for detecting Anomalies in GPS Time-Series

Richard J Self, University of Derby, UK

ABSTRACT

The GPS system is widely used for identifying the location of people and devices in many circumstances. However, it has been shown both by the author's final year students and others that GPS tends to have an accuracy of about +/- 25 meters 85% of the time, with the outliers ranging up to 1800 km. While the 25-meter accuracy is satisfactory for many situations, it is problematic when GPS is relied on to monitor the position of vehicles, especially near junctions, such is proposed for the Virtual Traffic Light systems, which need accuracies closer to 1 to 3 meters in order to manage the sequencing of movements. The requirement is to be able to use the GPS data, which is recorded in a time series at 1-second intervals, to provide some level of error estimation in order to correct the measured locations relative to the physical location. A group of students undertaking their Final Year Independent Studies projects used a range of approaches to investigate the feasibility of doing this, using both Microsoft Excel and SAS®. It is easy to visually identify the major positional errors when the data is plotted on a map or when a graph of velocity or acceleration is plotted against time. However, the challenge is to use an algorithmic approach. PROC ARIMA was developed for use on financial time series over long timescales but it was shown that it could be applied effectively to intervals as short as 1 second. This presentation demonstrates the power of PROC ARIMA in this unusual field.

INTRODUCTION

IoT sensors are now very significant collectors of information about our lives and our environment in continuous streams of time-series data. This data is part of the many sources of data to which we apply the power of analytics to gain value for our organisations and ourselves.

However, one of the problems is that IoT sensors are not perfectly accurate and, on occasion, supply erroneous data to the analytical systems, which may lead to incorrect analytics and decision making.

One area where it is easy to collect data from such a sensor is in the use of GPS data collected using smart phones, using location tagged photos, or GPS trackers during journeys.

Officially, GPS is generally accurate, in conditions of perfect visibility of the sky and the GP satellites, to within +/- 10 meters **95%** of the time. However, in the conditions that most devices are used to obtain GPS readings, this is not the case. I have had groups of final year BSc Students analyzing the accuracy of GPS data in real-world situations with the result that using over 2500 location tagged photos collected on a range of different smart phones they discovered that **85%** of photos were accurately positioned to within about +/- 25 meters. However, the remaining 15% of tagged photos demonstrated the problem of the range of anomalies from 25 meters up to 4500 km!

Humans can identify these problems by visually inspecting the tagged location on a map and then inspecting the photo to make such a determination as to how big the error actually is.

Between 2014 and 2016, an organization called ThinkNear (www.thinknear.com) carried out analyses of the accuracy of the location targeting of several billions of digital adverts transmitted to smart devices, based on the reported locations of the devices. As can be seen from the following summary of the data, the accuracy is rather low, with up to 14% being mis-targeted by over 100km and less than 40% directed to within 100 meters.

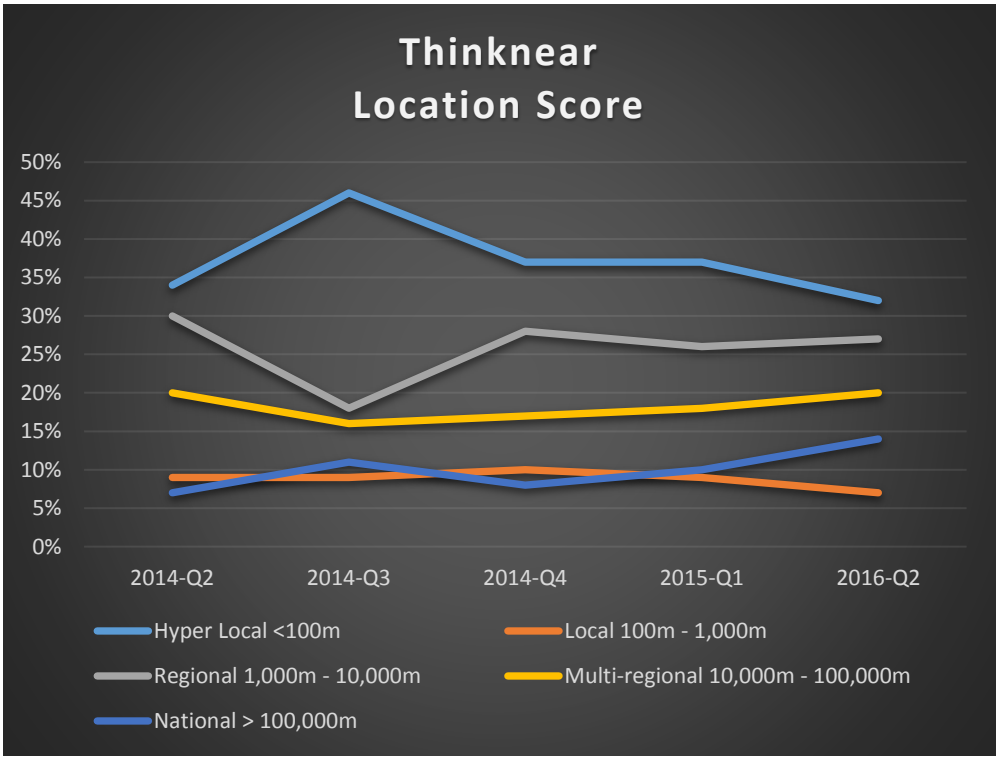

Figure 1, Data Sourced from http://www.thinknear.com/library/research/

What is very clear from these analyses is that attempting to identify measurement errors for single events based just on the location tag is very difficult without additional information, such as accurate knowledge of the location where photos are actually taken. What is needed is a mechanism to determine erroneous measurements in a continuous stream of measurements which will provide some form of context.

The next version of the challenge to final year students was to use time-series data collected by GPS trackers, recording the location every 2 or 5 seconds. This used the data from the MicroSoft Research Asia GeoLife dataset of 17,621 journeys by 182 people in China, see https://www.microsoft.com/en-us/download/details.aspx?id=52367 . Simple visualisations of the data in single journeys shows that whilst most of the points are clearly recorded with some degree of accuracy within the +/- 10m levels of accuracy, there are often single points which are clearly anomalous to the human eye.

As Fig 2 illustrates, when the tracker was stationary at the top left corner for several minutes, the device recorded a considerable journey of erroneous readings stretching some 300m, due to being inside a shop.

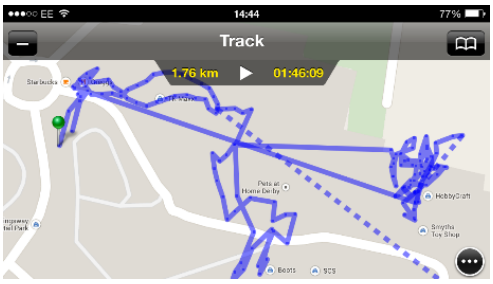

Figure 2,GPS Track Inaccuracies

However, if we are to be able to process significant streams of data, we need to be able to algorithmically detect the anomalous data in the time-series.

ANOMALY DETECTION IN A TIME SERIES USING MS EXCEL

The first group of students to use the Geolife data set analysed individual journeys, treating the data as a series of numbers. They were able to identify some of the most obvious errors in the measured data in the journeys by taking the first and second order differentials of the data to calculate the velocity using MS XL.

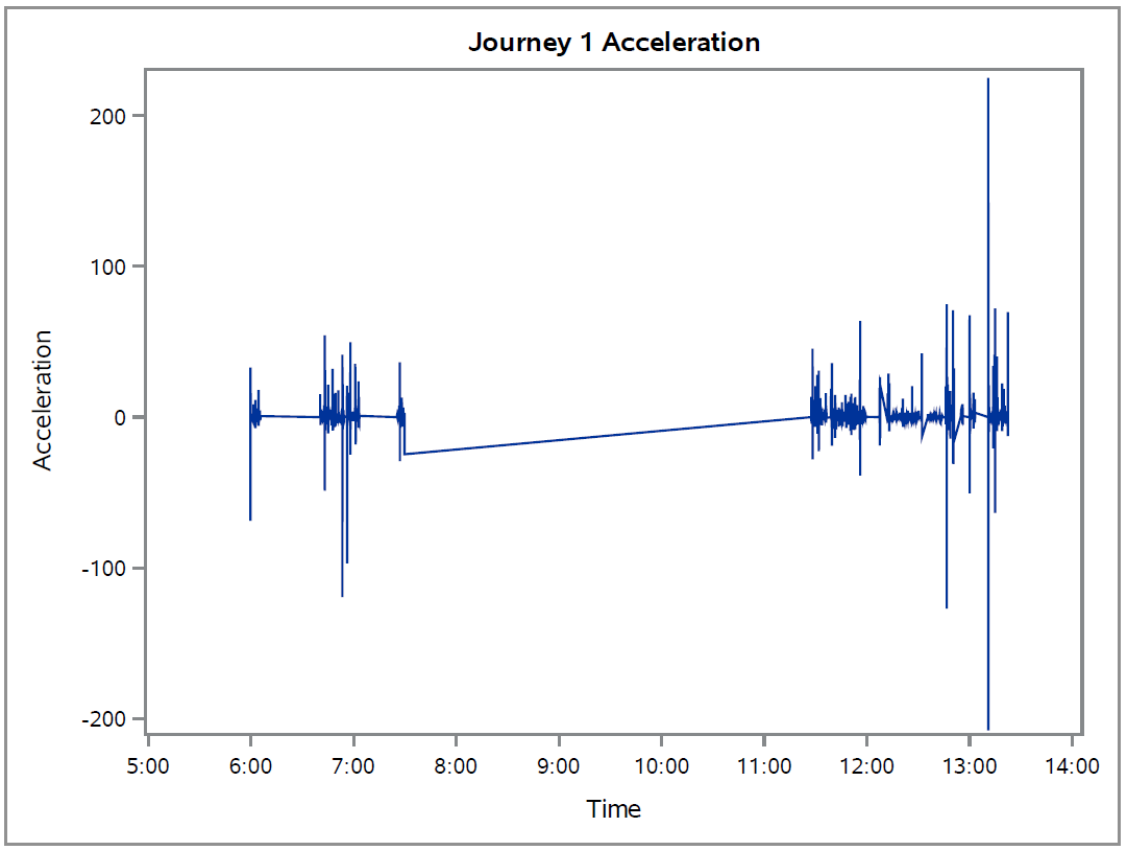

Figure 3. Acceleration (George, M. 2017)

It is clear that we can see some potentially interesting spikes in velocity and acceleration in the graphs. The question is whether there are algorithmic solutions to this. The fundamental question raised is the decision about the threshold for identifying the absolute value of the acceleration of velocity that indicates an anomalous reading. Visually, it is also possible that

there are signatures for these errors, such as a strong acceleration immediately followed by a strong deceleration. The use of MS XL for this is somewhat complex and was not attempted by that group of students.

STAGE TWO APPROACH TO ANOMALY DETECTION USING PHYSICS AND SAS®

The following year, 2017 -2018, the students were guided to investigate the procedures available in SAS®. In addition, they were sensitized to the fact that the time series of data points were surrounded by a physical context that allowed for the development of filters for the search for anomalous readings in the data stream.

Too often, data scientists treat their data as data without a context and try to extract meaning by applying their various tools. However, without domain understanding, this is a challenge. In general, the greatest problem for educators in the field of data science is not teaching the techniques, but the domain understanding that is required in order to become a useful data scientist and this tends to take years of working in the field.

Each of the Geolife journeys were recorded by the volunteers on foot, on bicycles, in busses and taxis or on the light rail transport in Beijing. Each of these modes of transport have physical and mechanical capabilities that govern both velocity and acceleration. It turns out that invoking these physical models that describe the maximum capabilities for acceleration and deceleration. For cars the limits are approximately 0.5g (5m per sec^2), for trains it is about 1 m per sec^2 etc.

Using this approach we can begin to see how to address the problem of algorithmic identification of the most important anomalies, by setting the limit at +/- 5m per sec^2 for all modes of transport, although we could do better, if we know the form of transport, which is the case for a few of the Geolife journeys. The work by Sultani and Taylor was based on the use of the more generic limits.

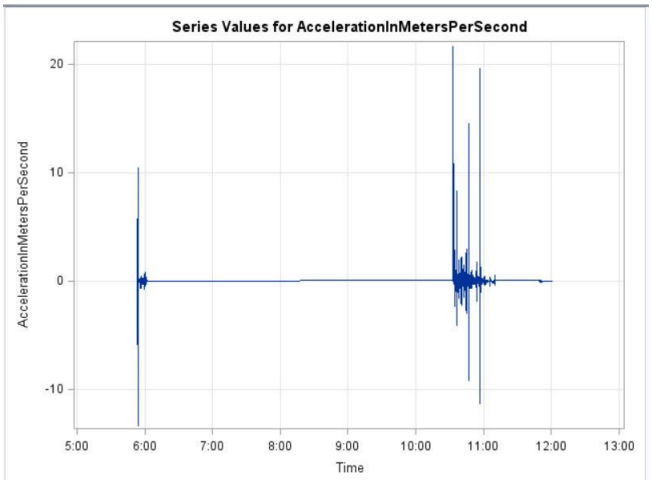

Figure 4, Acceleration (Taylor, A., 2018)

Here, we can see several excursions above the +/- 5m per sec^2 per limits.

PROC ARIMA AND ANOMALY DETECTION

Sultani and Taylor chose to investigate the use Proc ARIMA for this task. It turns out, that although PROC ARIMA is designed for Econometric situations with time intervals of days and months, it is also possible to use it with time intervals of seconds, which is what is required in this case.

They were successful in detecting a good range of the anomalies, as can be seen in Fig 5

Outlier Details

Obs	Type	Estimate	Chi-Square	Approx Prob>ChiSq
152	Additive	21.63231	7434.40	<.0001
564	Additive	19.56616	6473.82	<.0001
379	Additive	14.48824	4097.55	<.0001
18	Additive	-13.36983	3557.89	<.0001
563	Additive	-11.34498	2461.86	<.0001
19	Additive	11.38742	2545.00	<.0001
378	Additive	-9.08644	2073.25	<.0001
201	Additive	8.28001	1892.76	<.0001
562	Additive	-6.73110	1298.42	<.0001
7	Additive	-5.67657	910.30	<.0001
8	Additive	5.56920	719.97	<.0001
200	Additive	-4.25663	606.60	<.0001
561	Additive	-3.41097	403.22	<.0001
559	Additive	2.91095	301.32	<.0001
350	Additive	2.85471	294.14	<.0001
565	Additive	2.82286	289.85	<.0001
167	Additive	2.80759	290.68	<.0001
348	Additive	-2.65504	262.25	<.0001
4	Additive	2.71696	250.76	<.0001

Figure 5, Proc ARIMA Outliers (Taylor, 2018)

Sultani then used the Proc Arima functionality to both list the anomalies, as did Taylor and also use the forecast functionality (fig 6) and to plot the anomalies on the graph (fig 7).

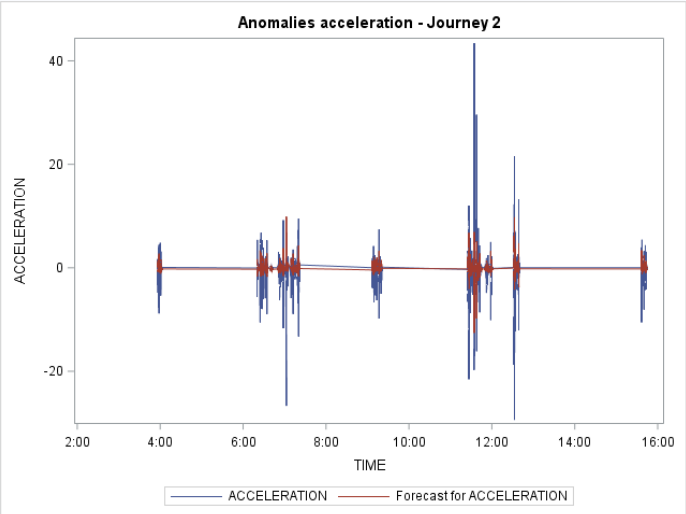

Figure 6, ARIMA Forecast of Acceleration

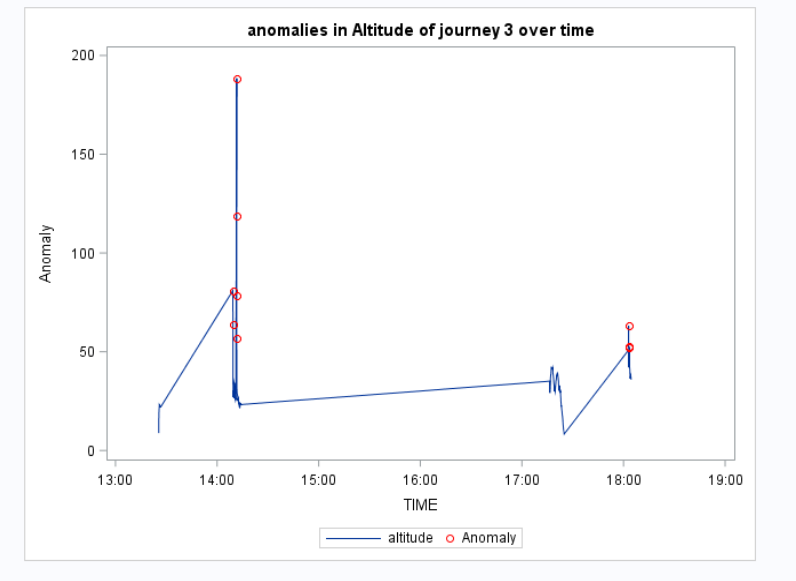

Figure 7: ARIMA Anomalies for Altitude in meters

It should be noted that GPS measurements of altitude can be very inaccurate. Physics models of travel can be invoked to identify the physical limits in rate of change of altitude and Proc ARIMA can, without prompting identify many of these anomalies

CONCLUSION

Identifying Anomalies in streams of IoT data is becoming a critical requirement due to the lack of accuracy of the sensors. This paper demonstrates that the SAS® ETS procedure ARIMA can be used with time steps as small as 1 second to process location data streams with a considerable degree of success.

REFERENCES

George, M, 2017, Identifying Anomalies in Location Data and Movement Pattern Analysis, University of Derby

Sultani, T, 2018, How can users of Location Data identify points which are erroneous and how can these data streams and the errors be visualised?, University of Derby

Taylor, A, 2018, Using Econometric Time Series Tools For Anomaly Detection Regarding Location Data, University of Derby

ACKNOWLEDGMENTS

This paper is based on the work of the following students of the author as they completed their final year BSc Information Technology Independent Studies Dissertation projects.

- Mark George
- Alex Taylor
- Tahera Sultani

CONTACT INFORMATION <HEADING 1>

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

Richard J Self
University of Derby, UK
r.j.self@derby.ac.uk
Web https://computing.derby.ac.uk/c/people-2/richard-j-self/ or
https://www.linkedin.com/in/richardselfllm/