USING CROWDSOURCED DATA (TWITTER & FACEBOOK) TO DELINEATE THE ORIGIN AND DESTINATION OF COMMUTERS OF THE GAUTRAIN PUBLIC TRANSIT SYSTEM IN SOUTH AFRICA

T. Moyo[a*], W. Musakwa[b]

[a] Masters Student, Dept. of Quality and Operations Management, University of Johannesburg. Cnr Siemert & Beit Streets, Doornfontein 0184 Johannesburg, South Africa- thembanijoel@gmail.com

[b] Senior Lecturer, Dept. of Town and Regional Planning, University of Johannesburg, Cnr Siemert & Beit Streets, Doornfontein 0184 Johannesburg, South Africa - wmusakwa@uj.ac.za

KEY WORDS: Origin and destination; geo-location data; commuters; kriging; density

ABSTRACT:

The study of commuters' origins and destinations (O D) promises to assist transportation planners with prediction models to inform decision making. Conventionally O_D surveys are undertaken through travel surveys and traffic counts, however data collection for these surveys has historically proven to be time consuming and having a strain on human resources, thus a need for an alternative data source arises. This study combines the use social media data and geographic information systems in the creation of a model for origin and destination surveys. The model tests the potential of using big data from Echo echo software which contains Twitter and Facebook data obtained from social media users in Gauteng. This data contains geo-location and it is used to determine origin and destination as well as concentration levels of Gautrain commuters. A kriging analysis was performed on the data to determine the O-D and concentration levels of Gautrain users. The results reveal the concentration of Gautrain commuters at various points of interest that is where they work, live or socialise. The results from the study highlight which nodes attract the most commuters and also possible locations for the expansion for Gautrain. Lastly, the study also highlights some weakness of crowdsourced data for informing transportation planning. (208)

1. INTRODUCTION

The study of commuters' origins and destinations (O_D) promises to assist transportation planners with prediction models which inform decision making. Conventionally O_D surveys are undertaken through travel surveys and traffic counts, however the data collection exercise for these surveys has historically proven to be time consuming and causing a strain on human resources, thus a need to aggrandize the data sources (Wolf, et al., 2003). As we are now living in the age of the internet of things (IoT) a need for smart analytic techniques has arisen. Bolstered by the current advancements in web 2.0, humanity has gradually departed from the culture of using the internet to send emails to incorporating it into every aspect of their lives (Gao & Liu, 2013). Chandler (2015, p183) articulated how data is now capable of altering the ways in which knowledge of the world is produced and consequently altering the ways in which it can be governed. This information age has subsequently created new prospects for transportation planning by revolutionising how information is managed, collected and analysed to improve transportation systems.

Accordingly with the dynamics of public transportation being multifaceted, the move towards intelligent transportation systems seems to be the logical solution for addressing situations which may unfold, instead of using the traditional reactive approaches. The general consensus amongst scholars is that it is now possible to model the spatial dependence of commuters using geographical location data to predict areas of clusters and outliners ((Wolf, et al., 2003; Stopher & Greaves, 2009; Gao & Liu, 2013; Hasan & Ukkusuri, 2014).

1.1 The evolution of analytics

The analysis of social media data can best be expressed through an insight of developments in data analysis which has evolved over the years as highlighted in figure 1. Historically data analysis between 1995-2009 scholars used data as a means to an end; and those between the years 2009-2013 analysing data as both a means to an end and also as the end. In the recent years there has been a paradigm shift with scholars from 2013-2016 analysing data as the end and those post 2014 analysing data as a service (Deloitte, 2014). This move from data as an end to empowering cities as a service opens up new possibilities, as data is no-longer only viewed as either a means to an end or just the end, but as an enabler for decision making and collecting feedback for development. Subsequently this offloads the risks and burdens of data management to a third-party cloud-based provider (Deloitte, 2014). The evolution of analytics has greatly changed the manner in which data is managed, as it has led to improvements in data quality, agility and reduction of cost. Furthermore these analytical tools seem to be a viable resource that will improve operational efficiency, while boosting the quality of urban planning.

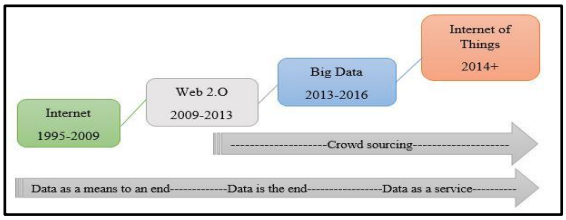

Figure 1: Evolution of analytics

1.2 Big data

In past 5 years there has been a rapid incorporation of social media data in transportation studies. Lorenzi, et al., (2014) have articulate how a middle class individual's life now revolves around the use of smart phones. The continued development of smart phones has led to these devices having in-built mobile location sensors. Furthermore this has given rise to an increase in development of mobile applications which rely on these location sensors (such as Facebook; Instagram; Strava Metro; Twitter and Google Maps). The data generated by these applications has the potential of being used to analyse the day to day movement networks of human beings. However in analysing this data, set backs were identified by Lorenzi, et al., (2014) in that the information measured was subject to noise and uncertainties, hence leading to imprecise results if these were not excluded in the analysis.

The growth of big data analytics has spawned remarkable captivation globally (Riggins & Wamba, 2015), this can be seen with many city officials engaging with the private sector, in a bid to make use of this big data, such as using Echoecho platform to analyse social media data. This has made it possible for city authorities and private companies to understand the multifaceted aspects of social media data. This will led to inside on how people interact with their immediate environment, through insight of social media big data which is amassed by posts made daily by the social media users. Big data has over the years been described as data sets whose size is beyond the ability of commonly used software tools to capture, manage and process the data within a tolerable elapsed time (Riggins & Wamba, 2015).

Consequently social behaviour identification through the quantification of various aspects of human behaviour is now possible through this big data. With regards to urban planning, the unpacking of big data has led to a reduction in time spent to respond to service delivery grievances, as the community can easily inform council of any grievances via mobile applications, thus bridging the gap between the ordinary citizen and local authorities (Hasan & Ukkusuri, 2014). As mobile devices have become more advanced within built sensors, it is now possible to trace and create a digital foot print showing the movement of people, through the collection of big data from mobile network towers, social media platforms and wifi feeds (Lancey, 2001; Yang, et al., 2012; Chatzimilioudis & Zeinalipour-Yazti, 2013). This has the potential to inform planning in that authorities can identify areas which have the potential for investment by analysing the sphere of influence of various land uses. Furthermore big data can be used to improve service delivery. For example Waze through its connected citizens program has assisted authorises such as in Rio de Janeiro to re-direct traffic to other routes to avoid traffic jams during the rush hour as Waze relies mainly on crowdsourced data from users (Waze W10, 2014).

1.3 Crowd sourced data

Crowd sourcing is the activity commonly referred to as a phenomenon in which a large group of people engage in a given task in order to harvest usable information (Estellés-Arolas & González-Ladrón-de-Guevara, 2012). There has been significant growth in utilisation of crowd sourcing as a *modus operandi* to disentangle the multifaceted issues that exist in the real world. The abundance of data on the World Wide Web coupled with the ability to acquire feedback from crowds has the prospective of altering the manner in which data is synthesized and decisions are made. Brambilla, et al., (2013, p1) outlined that "crowd sourcing can be used to answer questions that are inherently hard for machines but can be handled relatively easily with human input". This exploratory technique is generally an information-seeking activity where people gradually acquire knowledge about one or more issues of interest.

Meanwhile Chatzimilioudis & Zeinalipour-Yazti (2013) have tested the prospect of using the user's location as a form of crowd sourcing. Their research creates a trajectory for crowd sourcing activities and data management trends, as the identification of the geographic location of the crowd, will lead to an identification of hot and cold spots in the city. Moreover mobile crowd sourcing platforms such as the Waze connected citizen program have led to an improvement of service delivery and assisted in disaster management (Waze W10, 2014). Crowd sourced data can be analysed from big data collected from social media to identify the various trends circulating on the internet and this data can inform decision making through the storage, processing and analysis of real-time data streams. Map D, Strava Metro, Echo echo and Waze are examples of companies which analyse social media, to analyse people's views, to identify futurist trends and to advise decision makers in planning.

Musakwa (2014) used social media data to determine commuters' perception of the high speed railway network, the Gautrain in South Africa. This was made possible with increased access to the internet by commuters. However, currently most of the research which has been carried out only highlights the numerous ways to collect data through crowd sourcing techniques, and little has been done to incorporate this information with geographic information systems to inform decision making and facilitate sustainable developmental practises.

1.4 Internet of things

The internet allows the contemporary researcher to access information at any time or place, simply because it allows them to literally access various databases. Riggins & Wamba (2015, p1) have outlined how this "emerging IoT allows for the tracking and tracing of any tagged mobile object as it moves through its surrounding environment or a stationary device that monitors its changing surroundings." This opens up new possibilities in O_D analysis as people move around with various mobile devices which are constantly sending information to the internet, such as cell phones, tablets and smart watches. This will allow for more accurate location of trip generation and also tracing of the various movement networks.

With time and place having been the two major constraints in origin and destination surveys, the advancements of in crowd sourcing, big data and internet of things present an untapped gold mine of geo-location data. Consequently social behaviour identification through the quantification of various aspects of human behaviour is now a possibility in real time. With

technological advancements machines are now able to handle big data, plus through their ability to process algorithms at real time, they reveal insights about the social media users. These algorithms generate classes which can be used for sieving data according to predefined orders hence leading to a means to analysis existing patterns in the dataset (Big data privacy report, 2014). An example is twitter which uses learning algorithms for analysing big data to inform various twitter users based on their interests the latest trends and news stories.

1.5 Data interpretation technique

Kriging models have over the years been used in fields of mining, remote sensing and environmental disciplines to predicate spatial patterns (Cressie, 1991; Auston, 2002; Chahouki, et al., 2010). Kriging can be defined as a geo-statistical local interpolation procedure that utilizes the known locations of data points and distance between them to predict density patterns (Bonaventura & Castruccio, 2005).

Scholars such as Mohammad & Adnan (2011) have utilized kriging to predict bird species occurrences using observed records. Using density maps in GIS which are model-based estimations of data distributions, they used kriging to create ideal and impartial approximations models to predict the location of hot and cold spots for bird species. Their work, consequently forms the basis of this study, as a pre-analysis of the data needs to be done before selecting the appropriate parameters for kriging as a means of ensuring optimal estimates and minimum error.

1.6 Public Transport in Gauteng, South Africa

The history of public transport provision (PTP) in Gauteng is driven by various forms of social, economic and political forces that have moulded and shaped it to form its current nature. From the horse drawn cart of the early colonial era; to motorized systems; to the present multi-faced modes encompassing motorized and rail transportation. Khan (2014, p 173-174) have articulated how "the transport landscape in South Africa was largely shaped by colonial and apartheid social and spatial engineering to serve primarily the economic wants and social well-being of the minority white ruling class". Hence over the years since gaining independence most transportation policies (such as the National Land Transport Transition Act) have tried to advocate for more sustainable means of ensuring the provision of public transportation.

One of the solutions used to regulate transportation is travel demand management (TDM). This with regards to PTP seeks to reduce the amount of motorised travel (Del Mistro & Behrens, 2008), and this has been done in Gauteng through the implementation of the Rea vaya; Metro rail and bus; Ari yang; Putco; Gautrain and Gaubus. However TDM has not been fully implemented as people still prefer to use the mini-bus taxis, as they argue they carter more to their needs as they have more flexible operating hours and that they have successfully penetrated into various their POI. Thus there is still a need to make the formal forms of public transportation more attractive to the commuter.

To address such issues, the National Government identified the use of Intelligent Transportation Systems (I.T.S). ITS refer to "application of data processing, data communications, and systems engineering methodologies with the purpose of improved management, safety and efficiency of the surface transportation network." (Gauteng 25-year Integrated Transport Master Plan, 2013, p. 6). However ITS are heavy reliant on the collection and analysis of data, to make improvements in travel demand prediction, traffic modelling and O_D surveys. Hence to meet the goals of I.T.S, the following objectives were identified namely being safety; mobility; efficiency; productivity; energy and environment; customer satisfaction (Gauteng 25-year Integrated Transport Master Plan, 2013). Consequently the incorporation of I.T.S in PTP presents a new untapped source of data, and also introducing new aspects to origin and destination surveys such as big data, crowd sourcing, and internet of things. Accordingly there seems to be large market of social media users in South Africa, with Twitter and Facebook having 6.6 million and 11.8 million users respectively (World Wide Worx & Fuseware, 2015). Hence, the aim of the paper was to explore how geographic information systems and geo-location based analytic techniques can be used to define trip generation for the Gautrain nodes.

2 STUDY AREA

The paper mainly was focused on Gautrain Rapid Railway Link (GRRL) and their commuters (figure 1.1). As the Gautrain has been identified as the backbone for public transit provision in the province (Du Plessis, 2010; Gautrain, 2009), and the current gap in knowledge systems exists in how this can become a reality, as the Gauteng province has many inherent problems with regards to public transportation provision. Also given how the Gauteng City Region (GCR) is a cohesive cluster of cities, towns and urban nodes that collectively make up the economic hub of South Africa, generating more than 36% of the country's Gross Domestic Product (GDP), whilst covering less than 2% of the country's total surface area (Gautrain, 2009), an improvement public tansporation system planning becomes necessity to ensure continued sustainable development.

This economical hive is constantly drawing as an influx of commuters traverse through the province on a daily basis, leading to congestion becoming a norm on highways during the peak hours of the day. Bohlweki Environmental (2002) has outlined how the Gauteng Provincial Government identified the Pretoria CBD, Johannesburg CBD and the airport in East Rand as the most important nodes to be linked by the Gautrain this over time has led to the growth of activities on the other nodes such as Rosebank, Sandton and Hatfield as evident in figure 2 (Ruwanpathirana & Perera, 2015). In addition the Gautrain project is still at its inception as the project was only implemented in 2010, and how less than 6 years later, the railway line is still not near completion, with only 10 fully functional train stations, a need for its expansion and integration to other parts of the province still exists.

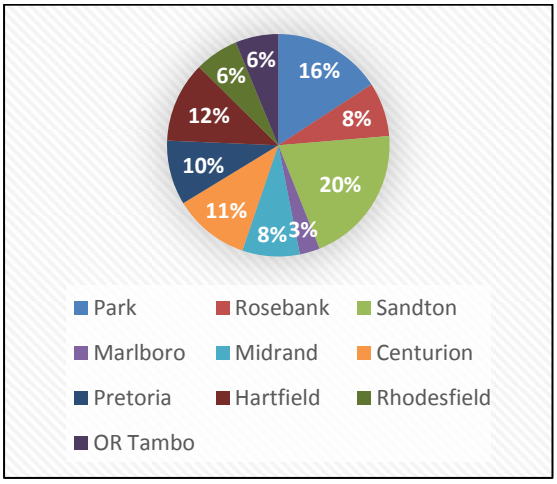

Figure 2: Number of passengers entering the station per month for January to June 2015

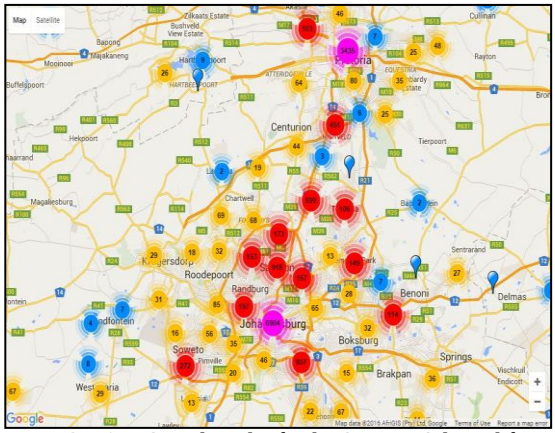

Figure 3: 2014 Social media feeds; Source: Adapted from Echo social (2015)

In the previous year in 2014 the terms Gautrain and station were mentioned 83195 and 19561 times respectively (figure 3), on social media platforms namely Twitter and Facebook. With such a high interaction of users, interfacing on social networks, it can be assumed that these are either existing or potential commuters of the Gautrain. Drawing from the background of the statics of how South Africans and the global world have embraced the use of Web 2.0, in their daily lives with Twitter and Facebook having 5 500 00 and 9 600 000 users respectively (Meier, 2013) and the a lack of integration within the public transport provision in Gauteng, which namely encompasses the rail, bus and mini-bus taxis services. The results of the paper shall be used as a means of identifying the extent of trip generation of the various nodes, whilst highlighting areas of crowd clusters, which through collaborative planning could be used as a basis of integrating the existing public transit systems.

Accordingly as this study is premised on the utilisation of social media big data to monitor the points of interest of Gautrain users, that is the demarcation of the sphere of influence of the Gautrain, it becomes evident that privacy concerns arise. As the data under analysis carries with it sensitive personal data of the users, that is the user's name and unfiltered tweet or facebook post, the research had to ensure that the data was only used for academic and planning purposes. Also another ethical issue becomes evident that is confidentiality, although information shared on social media platforms is public knowledge, the researcher still could be held liable for any misuse of the data, especially the geographic locations of the posts. Hence the researcher utilised the university's ethics and code of conduct policy to guide and inform how the data would be safe guarded to protect the interest of the Gautrain and also the social media users.

3 METHODOLOGY

To achieve the goal of the study, the research design adopted an experimental approach which used spatial and quantitative data, in a bid to explore Geographic Information Systems (GIS) techniques which can be used to define trip generation for the Gautrain nodes. This research design hence formed the blueprint of the study from inception to its epilogue. As spatial phenomenon in the real-world is made up three spatial dimensions namely the 'x'; 'y' and 'z' with x; y representing geographical co-ordinates and 'z' representing elevation, a means to incorporate these in the research was indispensable. Consequently the proposed criteria used for delineating trip generation was established using the model (figure 4) which was developed through strenuous trials of various analytic and visualisation *modus operandi*, until one which showed a real life interpretation of the data was found.

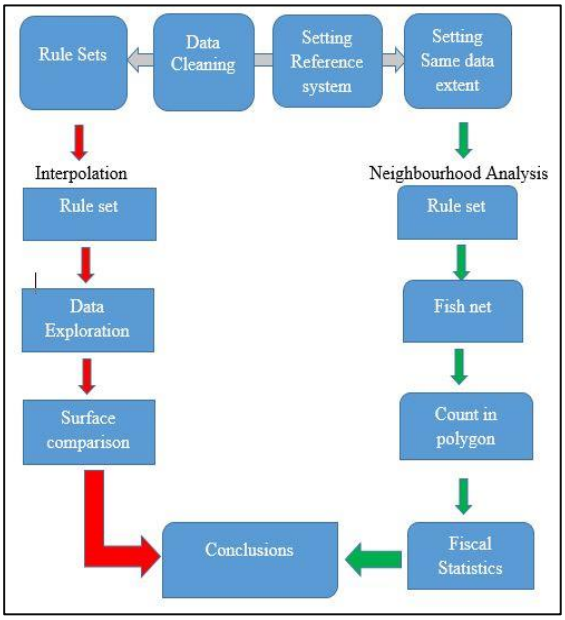

Figure 4: Model

Moreover, the model was used as the germane for distinguishing patterns of spatial association (clusters) and atypical spatial locations (outliers) for the various nodes. The execution of the various analyses carried out relied largely on the reliability of the information recorded that is all potential errors had to be minimised despite quality assurance being embedded in all the analytic processes, such as data collection and editing, errors may exist. Hence to reduce errors accumulating

the editing process was repeated until the researcher was satisfied that all the records used in the analysis reflected a true representation of the real world feeds. The raw data was edited as follows, with the following raising red flags, namely being fields missing content and these were removed from the data to be analysed:-

- No X co-ordinates
- No Y co-ordinates
- Missing X and Y co-ordinates
- Missing twitter comment
- Missing facebook comment

3.1 Data preparation

Figure 5 shows a summary of the data which was used in the analysis. After data cleaning 18633 geo-location social media big data records for the period January to June were uploaded onto Arc Map 10.3, and converted to vector points. The data was then projected to a common projection system (TM 29 Hartebeesthock 1994). After projecting the data was clipped to the extent of the various sites that is Gauteng; Johannesburg; Pretoria; East Rand. The 'z' value for all the sites was determined, by extracting the 'z' using the DEM raster layer for the various site extents. Consequently the model was developed using the rule set:-

1. Identify the spatial analytic technique
2. Running the spatial analytic technique
3. Rate the results from the analyses.
4. Comparison of Spatial patterns
5. Identification of clusters and outliners

Dataset	Type	Spatial Reference
Social data	Excel	Undefined
ZA Dem 90m	Raster	WGS1984 UTM zone 35S
Wards	Vector	GCS Hartebeesthoek 1994
Gauteng	Vector	GCS Hartebeesthoek 1994
Johannesburg	Vector	GCS Hartebeesthoek 1994
Pretoria	Vector	GCS Hartebeesthoek 1994
East Rand	Vector	GCS Hartebeesthoek 1994
Income	Vector	WGS 1984 UTM zone 35S
Gautrain Stations	Vector	HH94 Lo29
Gautrain rail tracks	Vector	HH94 Lo29

Figure 5: Dataset

Using the geostatitsic wizard, a histogram with 10 bars was created using the z-axis values of the social media big data. An expeditious analysis of the data set was then undertaken this which would show whether the data represented either a normal or abnormal distibution. The histogram shown in figure 6 consequently shows a very distinct unimodal (one hump) and skewed right, this hence relating to the existance both clusters and outliers existed in the data set. In geo-statistical analysis, outliers are considered to be sampling errors however in this analysis these outliers will be analyzed using the criteria weighting.

The ideology of the variogram is premised on the hypothesis that the spatial relation of two sample points does not only depend on their absolute geographical location, but rather on their relative location (Wackernagel 2003). Also Webster & Oliver (2007, p 65) have outlined how "the variogram as a geo-statistical method is a convenient tool for the analysis of spatial data and builds the basis for kriging". The cloud produced in the variogram in figure 7 represents the lag distances in the data set (Wackernagel, 2003). Subsequently if the data set had produced a discontinuity at the origin, then the height of the discontinuity being the nugget effect would be included in the krig, however there seems to be no discontinuity in the data set; and for the sill the value for the data set was $\gamma.10^{-5}$; whilst the range was 0.80.

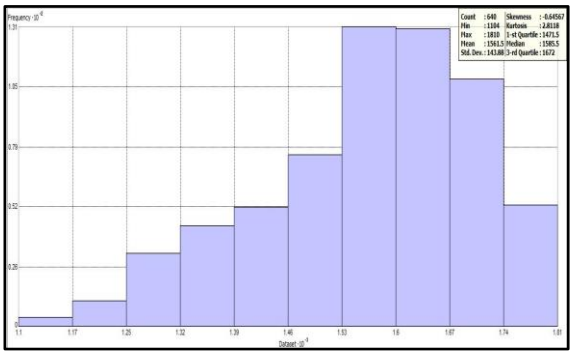

Figure 6: Histogram

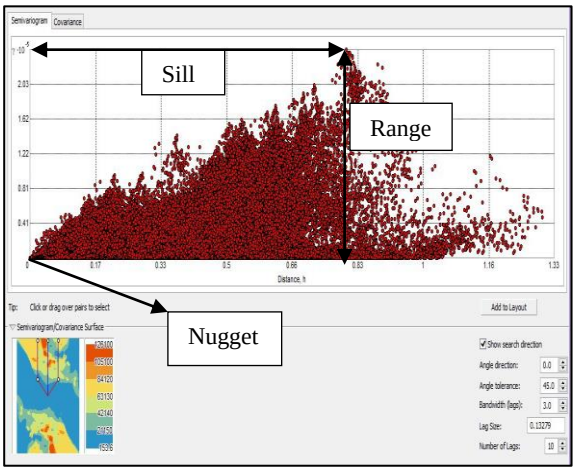

Figure 7: Semi-variogram in North-South direction

Kriging was then carried out to visualise these trends identified in the histogram and semi-variogram. The results of kriging from the simulated social media data show that for the various spatial resolution distance bands namely 1km; 1,5km and 2km, there is little variation in the density sizes of the various Gautrain nodes, hence only the 1km spatial resolution distance band was used in the analysis. To conduct the neighbourhood analysis, two methods were done, one using a focal statistics for a fishnet and the other using a count in polygon for the wards data. Initially a fishnet of 5kmx5km was initially created for the Gauteng region

and through the use of Geospatial Modelling Environment (Beyer, H.L, 2015), a count in polygon of the social media data was run for the fishnet and the wards data respectively. This created a new field called 'count in poly' in the fishnet and ward data which was then converted to raster format for the focal statistics to be run for the two layers. The analysis on the fishnet was carried out with the neighbourhood being rectangular with a height and width of 5x5km; and secondly with a circular neighbourhood with a radius of 2850 metres for the fishnet. Lastly for the wards using the count in polygon for the social media data. The data was then re-classified using an interval of 5 classes to visualise the results.

In view of the results an evaluation criterion for the model was then developed through brain storming and a review of literature. Tools such as list reduction and multi-voting were used to assess which criterion could be used to establish the extent of the clusters and outliers. A respective ranking was consequently produced based on the criterion was then developed with 5 representing areas of high cluster with potential for development and 1 representing areas with little to no potential for development.

4 RESULTS AND DISCUSSIONS

4.1 Neighbourhood Analysis

As C. Fiorina has articulated concerning big data, "the goal is to turn data into information, and information into insight." Hence from the analysis of the focal statistics and comparisons of the results, the wards data shows a clear distribution of hot and cold spots within the study area, with areas near or around the stations having hot spots as shown in figure 8. However this representation does not visually represent the real world, as these hot spots take the form or shape of the ward, thus making it difficult to compare density per km^2. The focal statistics for the fishnet however enables for an unbiased analysis over the surface area per 25 km^2 as shown in figure 9. However using the circular neighbourhood analysis produced a perspicuous hot spots. These hot spot sites can be used as means to justify developing more in these sites. More research is however indispensable to identify whether these as cluster sites of commuters, have any economic; social or historical influence which is attracting them there. Once it has been determined the frequency of the influx of commuters to these sites, the Gautrain railway line or Gaubus may be extent to these sites.

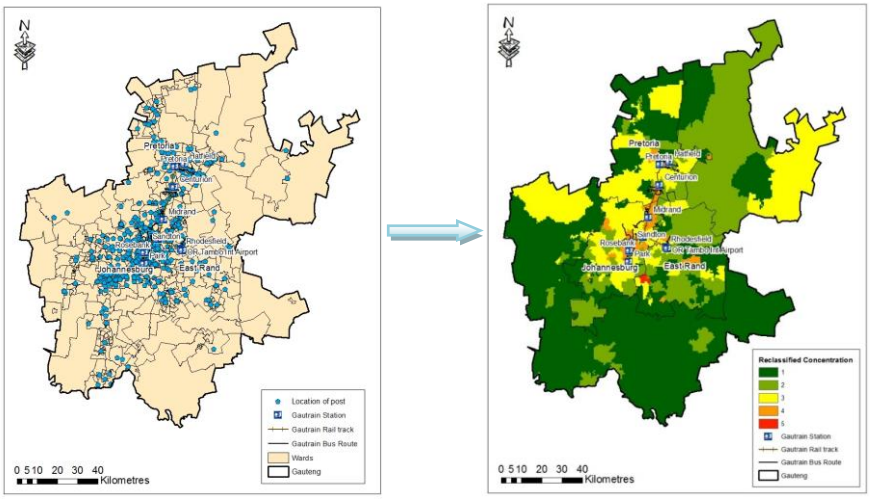

Figure 8: Gauteng wards and Focal statistics for wards

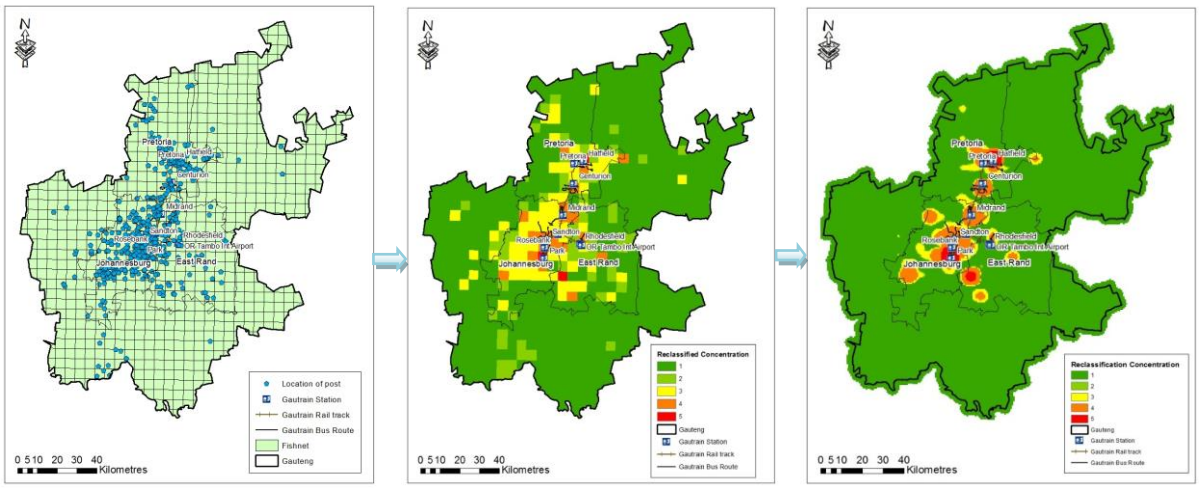

Figure 9: Gauteng fishnet and Focal statistics for fishnet

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-III-2-143-2016

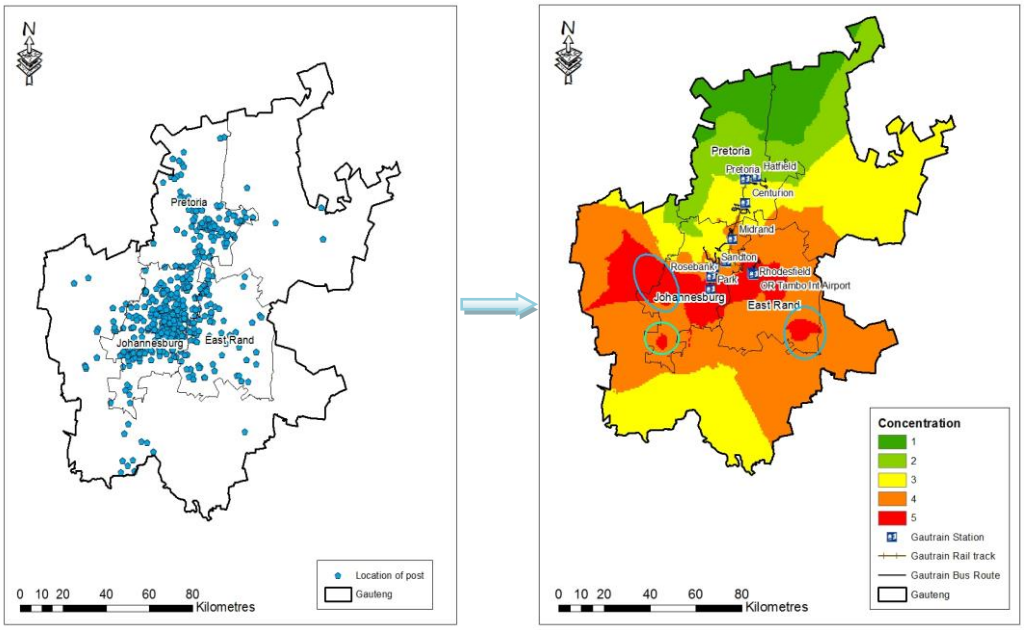

Figure 10: Gauteng Krig

4.2 Interpolation

Interpolation was simulated over 100 times using the empirical bayesian kriging technique to reduce the occurrence of errors. By means of the 'z' values extracted from the DEM data for the Gauteng province, prediction maps were consequently produced to visualise the trends identified in the preliminary analysis of the social media big data set as shown in figure 10. Also it seems that kriging seems quite sensitive to the presence of outliers or misfit values, as some clear indications of cold spots can be seen in northern parts of Pretoria, with the lowest values being recorded falling in the rank 1 in the criteria weighting. Also a hot spot belt seems to emerge in Johannesburg moving towards the East as highlighted in figure. Accordingly from the kriging results the majority of the users seem to be located near the train station locations as shown in figures, this could be due to that the current Gautrain stations are located in melting points of commuters. Examples include the Park node, which is located in the centre of the CBD and acts as an entry point for most regional and local commuters, also given the close proximity to Bree taxi rank and Mtn taxi rank being located only 10minutes away, this node has a high connectivity level.

As a result the high levels of social media posts around these locations do not necessary mean that the users reside around the train station, but that the train station is located in one of the commuters' major points of interest. Furthermore the kriging surface and neighbourhood analysis presents hot spots of areas which are currently not being directly serviced by the Gautrain or Gaubus. These are easily identified in areas such as the western and southern parts of Johannesburg; central and eastern parts of East Rand. Accordingly these locations could be areas worth investing into by either expanding the railway lines or bus routes to these as there is clearly a ready market these locations represent points of interest of the potential commuters. The combined maps from the neighbourhood analysis and krig support the study's hypothesis, and should be used as they visualise the points of interest of the Gautrain commuters using a prediction model with respect to a three dimensional analysis.

5 CONCLUSION

Using the model the study compared and contrast their merits and demerits, depending on the input datasets (Anselin, 1996). The GIS techniques offered the researcher various control elements to assist in determining spatial relationships which exist in the datasets. Accordingly the study revealed that the focal statistics presented the most visually accurate means of identifying clusters in the geo-location social media data per square metre. Hot spots were identified in areas near some stations such as Park Station and Sandton, this could mean these have the highest concentration of commuters. Also new hot spots were identified that is areas which are currently not serviced by the Gautrain and these are Soweto and Randburg in Johannesburg; Germiston and Alberton in East Rand; Montana Park in Pretoria. Subsequently these could be possible locations the Gautrain could further investigate as viable locations to expand the railway tracks to. Also through the results from kriging, hot and cold spots are easily identifiable, hence locations with hot spots should be further invested in, and as these are clearly points of interests of the commuters. However further research is still needed, such as running the model whilst incorporating other control factors to determine variations using a time-series analysis, to identify any variations in hot and cold spots over time, thus areas which would present a constant hot spot would clearly be worth investing into.

Also as the purpose of study practices vary, also the applicability of the model should differ. Different scholars conduct research for numerous objectives. Accordingly the current study, was based on demarcating the locations were Gautrain commuters, were coming from, and not what affects their mode of transportation. In this cause the model efficiency is being judged on

whether it was able to visualise the sphere of influence of the Gautrain. Using the assumption that people who post about the Gautrain are either existing or potential commuters, then yes the model does show the sphere of influence of the Gautrain. Nevertheless in the real world situation, this may not always be the case. Hence the shortcoming of the study appears, is there is currently no clear manner to ensure that everyone who posts about the Gautrain is a current or potential commuter.

REFERENCES

Anselin, L., 1996. The Moran scatterplot as an ESDA tool to assess local instability in special associations. pp. 1-118.

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling, 157, pp.101–118.

Beyer. H.L, 2015. Geospatial Modelling Environment (Version 0.7.4.0). (software). URL http://www.spatialecology.com/gme.

Big data privacy report for the Executive Office of the President, 2014. Big Data: Seizing opportunities, preserving values. pp. 1-69.

Bonaventura. L & Castruccio. S, 2005. Random notes on kriging: an introduction to geostatistical interpolation for environmental applications p 1-28

Bohlweki Environmental, 2002. Environmental Impact Assessment for the proposed Gautrain Rapid Link Project. Issues Report, Pp. 1-151.

Chahouki, M.A.Z. Azarnivand, H. Jafari, M. and Tavili.A, 2010.Multivariate statistical methods as a tool for model-based prediction of vegetation types. Russian Journal of Ecology, 41, pp.84–94.

Chatzimilioudis, G. & Zeinalipour-Yazti, D., 2013. "Crowdsourcing for Mobile Data Management". Proceedings of the IEEE 14th International Conference on Mobile Data Management (MDM), Volume 2, pp. 3-4.

Cressie, N.A.C., 1993. Statistics for spatial data (1st ed.). New York: John Wiley and Sons.

Del Mistro & Behrens, R., 2008. How variable is the variability in traffic? How can TDM succeed? In Annual Southern African transport conference 7-11 July.

Deloitte, 2014. Operationalizing the Analytics. s.l., Kelley School of Business, pp. 1-22.

Du Plessis, J., 2010. Injecting a rapid rail link into a metropolis. pp. 1-10.

Estellés-Arolas, E. & González-Ladrón-de-Guevara, F., 2012. Towards an integrated crowdsourcing definition. Journal of Information science, 38(2), pp. 189-200.

Gao, H. & Liu, H., 2013. Data analysis on location-based social networks.. In A. Chin, & D. Zhang (Eds.), Mobile Social Networking: An Innovative Approach, pp. 164-194.

Gauteng 25-year Integrated Transport Master Plan, 2013. 25-YEAR INTEGRATED TRANSPORT MASTER PLAN. Annexure K: Intelligent Transport Systems, pp. 1-56.

Gautrain, 2009. Gautrain to be the backbone of integrated public transport;Gautrain News. [Online]
Available at:
http://www.gautrain.co.za/newsroom/2009/02/gautrainto bethebackboneofintegratedpublictransport/
[Accessed 20 April 2015].

Hasan, S. & Ukkusuri, S., 2014. Urban activity pattern classification using topic models from online geo-location data. pp. 1-19.

Khan, S., 2014. Historical evolution of Durban's public transport system and challenges for the post-apartheid metropolitan government. New Contree, Volume 70, pp. 173-194.

Krige, D.G., 1951. A statistical approach to some mine valuations problems at the Witwatersrand. Journal of the Chemical, Metallurgical and Mining Society of South Africa, 52, pp.119-139.

Lancey, D., 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety.

Matheron G, 1963. Principles of geostatistics, Economic Geology, 58, p. 1246-1266

Meier, G. (2013, September 29). Retrieved October 12, 2015, from Blue magnet:
http://www.bluemagnet.co.za/blog/the-state-of-social-media-in-south-africa-2013

Mohammad. S & Adnan. R, 2011, Geostatistical techniques for predicting bird species occurrences: Master's of Science Thesis in Geoinformatics

No. 5 of 2009: National Land Transport Act, 2009.

Riggins, F. & Wamba, F., 2015. Research directions on the adoption, usage and impact of the Internet of Things through the use of Big Data Analytics. The 48 Hawaii International Conferences on System Sciences (HICSS), 5-8 January.pp. 1-10.

Ruwanpathirana, S. & Perera, I., 2015. CDME – Crowd-Sourced Data Mapping Engine. System that Analyzes, Maps & Publishes Crowd-Sourced Data on Environment Facts, pp. 1-6.

Stopher, P. & Greaves, S., 2009. "Missing and inaccurate information from travel surveys – pilot results". 32nd Australasian Transport Research Forum.

Wolf, J., Oliveira, M. & Thompson, M., 2003. Journal of the Transportation Research Board. "Impact of underreporting on mileage and travel time estimates: Results from global positioning system-enhanced household travel survey transportation research record", Volume 1854, pp. 189-198.

Musakwa. W, 2014. The use of social media in the Gautrain in Gauteng Province, South Africa : analysis and lessons learnt. REAL CORP 2014 Proceedings.

Wackernagel, H, 2003. Multivariate Geo-statistics: An Introduction with Applications (3rd ed.). Berlin, Heidelberg: Springer.

Waze W10, 2014. Launch Event Panel: Connected Citizens Program, hosted by Baratunde Thurston. An Evening of Discussion and Celebration in New York. https://www.youtube.com/watch?v=AMqbh3rqZRs

Webster, R & Oliver M A, 2007. Geo-statistics for Environmental Scientists (2nd ed.). Statistics in Practice. Chichester: John Wiley & Sons, Ltd. pp. 60-70

World Wide Worx and Fuseware: SA Social Media Landscape, 2015
http://www.worldwideworx.com/wp-content/uploads/2014/11/Exec-Summary-Social-Media-2015.pdf

Yang, D., Xue, G., Fang, X. & Tang, J., 2012. Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing. In Proceedings of the 18th annual international conference on Mobile computing and networking, Volume ACM, pp. 173-184.