The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLI-B2, 2016
XXIII ISPRS Congress, 12–19 July 2016, Prague, Czech Republic

TWEETS AND FACEBOOK POSTS, THE NOVELTY TECHNIQUES IN THE CREATION OF ORIGIN-DESTINATION MODELS

H. K. Malema [a, *], W. Musakwa [b]

[a] University of Johannesburg, Faculty of Engineering and Built Environment, Kingsway & University Road, Johannesburg - hopekoketsomalema@gmail.com
[b] Department of Town and Regional Planning ,University of Johannesburg, Cnr Joe Slovo Drive and Beit Streets Doornfontein - wmusakwa@uj.ac.za

KEY WORDS: Geolocation based services, Big data, Social media, Pattern analysis, Network movements, Origin-Destination models, Kriging, Transportation planning.

ABSTRACT:

Social media and big data have emerged to be a useful source of information that can be used for planning purposes, particularly transportation planning and trip-distribution studies. Cities in developing countries such as South Africa often struggle with out-dated, unreliable and cumbersome techniques such as traffic counts and household surveys to conduct origin and destination studies. The emergence of ubiquitous crowd sourced data, big data, social media and geolocation based services has shown huge potential in providing useful information for origin and destination studies. Perhaps such information can be utilised to determine the origin and destination of commuters using the Gautrain, a high-speed railway in Gauteng province South Africa. To date little is known about the origins and destinations of Gautrain commuters. Accordingly, this study assesses the viability of using geolocation-based services namely Facebook and Twitter in mapping out the network movements of Gautrain commuters. Explorative Spatial Data Analysis (ESDA), Echo-social and ArcGis software were used to extract social media data, i.e. tweets and Facebook posts as well as to visualize the concentration of Gautrain commuters. The results demonstrate that big data and geolocation based services have the significant potential to predict movement network patterns of commuters and this information can thus, be used to inform and improve transportation planning. Nevertheless use of crowd sourced data and big data has privacy concerns that still need to be addressed.

1. INTRODUCTION

Over the past 30 years Origin-Destination models (O-D) have evolved from static to real-time dynamic traffic models (Zhou & Mahmassani, 2007). These models have been crucial in establishing Intelligent Transportation Systems (ITS) for a city, since they provide predictions of traffic flows and network movements of commuters amongst other things (Hu & Liou, 2014). These predictions allows officials to identify road networks which are the most busiest, the times in which traffic is most congested as well as the modes of transport used by commuters. This kind of information is essential in transportation planning since it presents an opportunity for transportation policies to be improved and for ITS to be used in transport management (Hu & Liou, 2014).

An effective transport management strategy is one of the most important elements that contribute towards the sustainability of a city. Not only does it assist with the mobility of goods and people in a city, but an effective transport management system also impacts positively on the economic and environmental aspects of transportation **(**Gao et al., 2012**)**. Most cities around the world, however do not have these effective management strategies in place, hence there is a rising percentage of road accidents, the use of private automobiles and traffic congestions experienced **(**Fernandes et al., 2012**)**. This predicament thus calls for the efficient, reliable and integrated planning of transportation systems, especially in the developing nations.

* Corresponding author

Various scholars have thus, identified Intelligent Communication Technologies (ICT) as efficient tools which have the potential to assist with the effective management of transportation systems **(**European Commission, 2010**)**. This is seen evident in cities such as Barcelona and Dubai, given that both cities are well renowned for being prominent smart cities established to date (Bonnel et al., 2015 & Dassani et al., 2015). These cities have transportation systems that incorporate ICT which is in the form of smartphones, apps, sensors and WiFi in the mobility of their citizens. South Africa has recently caught on to this trend, as it has established its first rapid train system known as the Gautrain, which is a first of its kind to be established in Africa.

1.1 The Gautrain Project

The Gautrain is a mega-engineering project which is the very first rapid-transit train to be launched in the African continent in June of 2010 (GMA, 2010). The train project aims to reduce traffic congestion and encourage the use of public transportation systems in Gauteng, whilst being a means to realise smart mobility through the incorporation of ICT. The Gautrain is built and operated using some of the most advanced technologies in world and this makes for the success of the train project (GMA, 2010). In addition, the Gautrain has a phone application which displays the time schedule and alerts of delayed trains, thus allowing commuters to manage their travelling trips accordingly. Moreover, the app has a feature which permits commuters to calculate the amount of train fares from one station to the other, based on the time of the day. Such an app provides commuters an insight on the trip that is about to be pursued (Liu & Teng, 2015).

Figure 1: The Gautrain and its routes

The railway network of the Gautrain runs through the Gauteng province (Figure 3), linking three metropolitan cities: the City of Johannesburg (COJ), the City of Tshwane (COT) and Ekurhuleni Metropolitan Municipality (EMM). To date there is little known about the travelling behaviours and patterns of the Gautrain commuters, hence the study assesses the feasibility of using modern technologies as tools to create origin-destination models for Gautrain commuters. This approach is being tried and tested as there is evidence of insufficient spatiotemporal data obtained for this project.

In recent years, positioning technologies such as sensors and social media networks have been incorporated into transportation planning and particularly in the creation of O-D models. This novel approach can be credited for its ability to create sustainable and efficient transportation network systems. Detecting the geographical location of consumers from advanced technologies such as the Web 2.0 and big data together with social media network sites, has provided a platform for instant communication on real-time traffic updates to take place (Chen & Liu, 2014). The use of these technologies has made the exchange of data far much easier and as a result, this subject matter has become of great interest in the research domain particularly in transportation planning (Hongyan & Fasheng, 2013) and (Peters et al., 2013).

Likewise, this study aims to assess the feasibility of using geolocation-based services such as Twitter and Facebook as data mining tools to map the movement network patterns of the Gautrain commuters. The remainder of the paper is structured as follows; the next section presents the literature study on Origin-Destination models, which is briefly followed by a summary of the study area. A segment discussing the methods and materials used to conduct the study follows this. Lastly the proceeding section highlights the key results and implications of big data in planning.

1.2 Origin-Destination studies

O-D studies and models are a fundamental element in transportation management and have since been identified as being key tools in informing transit planning (Bohte & Maat, 2009; Kling & Pozdnoukhov, 2012; Gao et al., 2012). Previously, O-D studies used conventional methods such as household surveys and traffic counts as data mining tools (GAO, et al., 2012; Jin et al., 2014). With time, these methods proved to be rather spatially limited, outdated, cumbersome, unreliable, expensive and tedious (Gao et al., 2012; Lu et al, 2013). Hence, this novel approach of using positioning technologies to capture and trace commuter's demands and network movements in the attempt to establish real-time, reliable and spatially expanse transportation datasets used to optimize the use of street networks (Lu et al., 2013).

Positioning technologies such as smartphones, social media networks and Geographic Positioning Systems (GPS) which make up big data and Web 2.0, serve as data collection tools in the creation of O-D studies (Farah, 2014). This application of these technological devices and sensors contributes towards the establishment of smart cities since the technology allows for the easy exchange of traffic information to take place instantly. In addition, these technological devices assist with the monitoring of traffic congestion and road accidents, amongst other things (Caragliu et al., 2009).

Figure 1 depicts the connectivity and transfer of information that is typical in a smart city. Smart cities are described as cities which invest in both human and social capital, traditional as well as modern Information Communication Technologies (ICT) which fuel sustainable economic growth and high quality of life, through having prudent management of natural resources encouraged by participatory governance (Caragliu et al., 2009).

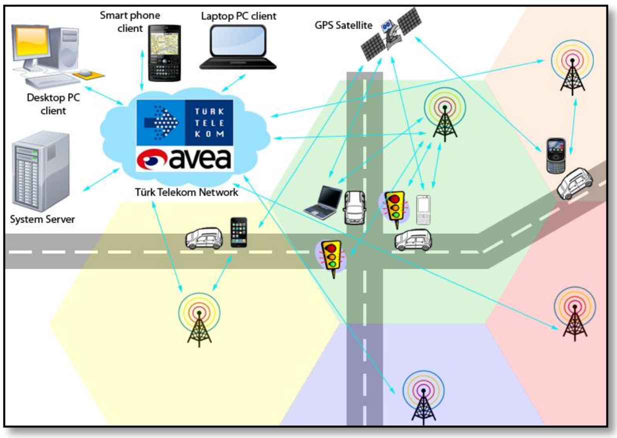

Figure 2: Illustration of smart city. Source: Turktelekomint (2015).

The smart city concept is centred on the idea of mankind incorporating technologies in the planning and operations of cities with the intent to improve the quality of lives of citizens whilst creating sustainable cities. To date, there are numerous smart cities established around the world with the majority of these being in the First World countries. Some of these cities include Barcelona, Amsterdam, London and Dubai to name a few (Bakici et al., 2013).

Dubai is one of the most renowned intelligent cities established in the Third World countries. This city is highly commended for effectively employing ICT in the governance of the city, which subsequently promotes transparency, easy access to information, sustainable transportation systems and so forth (Dassani et al., 2015). Cellphones, have identified as a very important gadget in upholding the standards of Dubai, since this ubiquitous device provides instant access to datasets and information that is crucial for citizens.

1.3 Geolocation-based services

The past 30 years of technological developments has led to an era of social media revolution which has seen almost half of Chinese students using mobile social media sites such as Facebook and Twitter to acquire information on various issues (Xu et al., 2015). The establishments of these communication systems allows for large amounts of data to be captured, stored and exchanged on smartphones, tablets and other mobile technologies, instantly (Peters et al., 2013). This era has brought with it innovative approaches of retrieving data for various uses such as the improvement of service delivery, customer services as well as transportation planning. As a result, social media networks like Twitter, apps such as Waze and sensors like GPS have thus, become popular for their ability to provide real-time traffic updates (Hongyan & Fasheng, 2013).

The acquisition of data through the above mentioned technologies is made effortless and convenient, due to the ubiquitous nature of smartphones. Not only do these technological devices offer access to the Internet, they also provide geographical location of consumers through the use of GPS sensors (Kaya et al., 2014). This feature makes it easy to identify and navigate to the location of a consumer or a loved one. Moreover, information on the activities undertaken by consumers at a specific location is also provided through social media applications (Hasan & Ukkusuri, 2014). These types of mobile applications are commonly referred to as geolocation-based services. The most popular amongst these are Facebook, Twitter and Foursquare as they contain a check-in feature that provides the geographical location of consumers as they navigate in and around cities (de Abreu Freire & Painho, 2014).

These services have become very popular in research because of their ability to provide rich data that has the potential to improve basic service provision such as road infrastructure for a particular area (Hasan & Ukkusuri, 2014; Peters et al., 2013). Despite their use in various fields, the geolocation based services are at the forefront of providing data in the transportation domain. Their ability to provide information on time, date, routes and activities rendered at a specific time, is a distinct feature that makes these services essential for providing information for Origin-Destination models (Filippi et al., 2013). In addition, the information shared on geolocation based services provides commuters the opportunity to plan their trips accordingly should there be an alert made of an accident, road constructions or traffic congestion in a particular road network.

A study revealed that nearly one in five smartphone consumers in the United States of America uses geolocation based services whilst commuting (Comscore, 2011). This statistic demonstrates the potential use of smartphones as data collection tools since they are utilised everywhere and at any given time. The data collected through this device if exchanged and shared through crowdsourcing, can be used for various reasons, including the creation of Origin-Destination models (Filippi et al., 2013). Accordingly, the O-D models will be used to provide an insight as well as predictions on individual travel behaviours (Wechsler, 2014). This predicted travel information is of great significance as it allows transportation planning policies to be drafted based on up-to-date data (Kaya et al., 2014).

1.4 Privacy concerns

The use of crowdsourced data in various fields is becoming increasingly popular. This is due to the ubiquitous traits of cell phones, which allows for the capturing, recording and sharing of spatio-temporal data (Farah, 2014). Crowd sourced data is commended for empowering citizens by allowing them to provide an input on issues and matters concerning development in their communities, amongst other things (Blatt, 2015). Conversely, crowd sourced data has been widely criticized for the reliability or lack of the data obtained. Scholars frequently raise this concern and they argue that data and opinions provided by ordinary citizens with no level of expertise on certain issues cannot possibly provide substantial information that could inform policies (Blatt, 2015; Callister, 2000).
Smartphones and crowdsourced data provide large amounts of significant information particularly relating to traffic (Filippi et al., 2013). However, cell phone use raises privacy concerns when consumers share data on the Internet. This is identified as one of the weaknesses of crowdsourced data since information such as location, identification number, contact details, etc., can be made available for anyone to access especially through social media sites (Blatt, 2014). Once this kind of information is obtained by people with malicious intent, it can consequently be abused.

1.5 The use of Smartphones in South Africa

Smartphone’s are renowned for being efficient data collection tools (Xu et al., 2015), however in South Africa these technologies are yet to be fully exploited for uses other than mere communication devices. It is in the past 10 years, however, that these technological devices, together with crowdsourcing software have emerged and are being explored. Some of these software include echo-social and Mapbox which have the potential to inform transportation planning based on the feeds from commuters. The former has the ability to capture data on the content and location of commuters who use social media sites whilst commuting (Loebal, 2012). The latter, however, is used to capture the location and number of social media users based on the different smartphones used at a specific time period. Both these software have the potential to provide essential information to planning authorities on the movement patterns and behaviours of South African commuters, if used correctly. With South Africa’s poor public transportation management and planning policies, these novelties could be used to improve on the transportation system of the country, whilst providing an attempt at creating smart mobility.

2. STUDY AREA

South Africa is considered to be one of the most developed countries found on the African continent. This country has been ranked amongst the top ten countries to have the highest GDP in the African continent (IMF World Economic Outlook, 2015). With that said, South Africa like many other African countries still lacks essential infrastructure, relevant skills and resources that would aid the development of the country (Douangphachanh & Oneyama, 2014). Consequently, the country experiences a multitude of issues related to transportation planning, urbanisation, poverty and so forth.

Planning in South Africa was previously conducted using the apartheid statutory approaches hence, the settlement planning in the country resembles segregation which current planners are attempting to rectify through modern planning theories (Todes, 2011). Similarly, the transportation planning policies are outdated, resulting in poor transportation management systems that give effect to poor public transportation systems, the preferred use of private automobiles, traffic congestion and heavy emissions of carbon monoxide (RSA, 2008). These policies were often informed by household surveys and traffic counts (RSA, 2008), which were previously considered the main methods to acquiring information on the travelling behaviors of commuters.

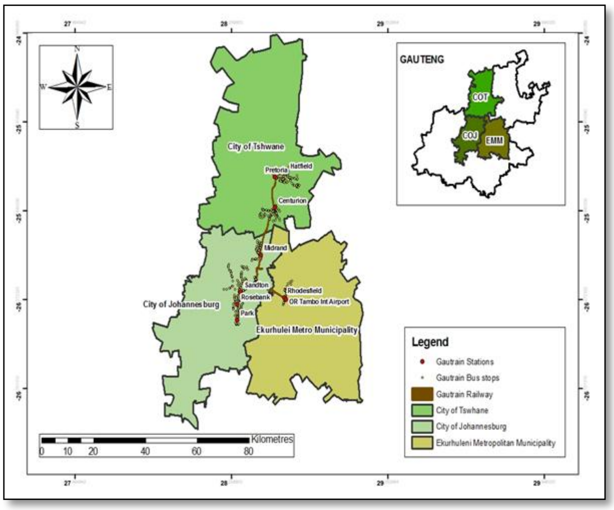

Figure 3: Study area

The Gauteng province is the smallest province in South Africa and has a land cover area of 18 178 m^2. The province is reported to ironically have the highest population size of 12 272 263 million residents which is comprised of 77.4% of the African population, 2.9% Asian, 3.5% Colored and 18.3% White (South African Census, 2011). The large populace in this province puts pressure on the available resources and movement systems of the province, thus creating traffic congestion, unemployment, water scarcity and housing challenges (City Of Johannesburg, Integrated Development Plan (COJ, IDP), 2012; Republic of South Africa (RSA), 2014). This population is concentrated in areas such as the Johannesburg CBD and surrounding areas, and likewise the province experiences urban sprawl, high urbanisation rates and spatial polarization. This predicament has thus created a need for the deployment of smart city concepts and different planning mechanisms, which could potentially assist with managing the challenges, encountered in the province and in the process promote sustainability.

The Gauteng province is the ‘Economic hub’ of South Africa and contributes 3.3% of GDP towards the country’s economy (RSA, 2012). This province further boasts with diversity, which ranges from a variety of ethnic groups from across Africa, to the various economic activities that take place in the major cities (GMA, 2010). The Gauteng province is reported to have 58% of its populace being economically viable. It is comprised of three metropolitan cities: the City of Johannesburg (COJ), City of Tshwane (COT) and Ekurhuleni Metropolitan Municipality (EMM) as illustrated in Figure 2 above. These metropolitan cities combined create an economic region, which is mostly comprised of the Finance, Real Estate and Business sector as well as the manufacturing sector. These metropolitan cities, in addition also create a corridor, which is clustered with numerous industrial developments along the N1 highway. Furthermore, the cities share the Gautrain railway, which is Africa’s, first high-speed railway.

3. METHODS AND MATERIALS

3.1 Methodology

Spatial and quantitative techniques were used to determine the trip distribution of Gautrain commuters. Tweets and Facebook posts containing information on the location of Gautrain commuters were used to derive the concentration levels of the commuters in various neighbourhoods of Johannesburg, Ekurhuleni and Tshwane. The Exploratory Spatial Data Analysis (ESDA) was employed as the research design since it assists in revealing patterns, trends and distributions spatially.

3.1.1 Data collection and preparation: Cadastral data of the municipal boundaries for the three metropolitan cities and social media feeds (tweets and Facebook posts) were collected from municipalities and the Echo-social software, respectively. This software allows for the active monitoring and engagement between customers and company owners to take place in the social media space. The data obtained from Echo-social contains information on the thoughts and opinions of Gautrain commuters.

The data used for this study was captured over a 6 month period starting from 1st January 2015 to 1st June 2015. The Echo-social software contained relevant data which was only captured for the above mentioned dates; hence the study used information recorded within the 6 month period. This data was captured in an excel spreadsheet, which contained crucial information on the coordinates and location of consumers who posted and tweeted about the Gautrain as seen in Figure 3 below.

1	Message	User	Lat	Long	City	Source	Date	Time
2	@Sadie_theeRebel Hahaha I sa\	Nkulu_S	-30.5595	22.93751	Somewhe	Twitter	01-01-15	42:07.0
3	RT @Nkulu_S: @Sadie_theeReb	Sadie_the	-25.7313	28.21837	Pretoria,S	Twitter	01-01-15	51:43.0
4	@Kaizer_Chiefs All I can say is r	KwinikaZa	-26.2041	28.04731	Johannesl	Twitter	01-01-15	58:06.0
5	@TheGautrain is the train opera	ick_Sesho	-26.2041	28.04731	johannesl	Twitter	01-01-15	35:43.0
6	Homeward bound... #knysna (@	PieterClo	-34.0351	23.04647	Knysna	Twitter	01-01-15	16:41.0
7	Me and Your Mom re ka stop'a (	BuddyTeb	-25.7234	28.42215	Mamelodi	Twitter	01-01-15	54:10.0
8	I'm at Gautrain Marlboro Statior	PieterClo	-34.0351	23.04647	Knysna	Twitter	01-01-15	59:26.0

Table 1: Enumerator values.

The coordinates contained in Table 1, were used to visualise the geographical locations of the tweets and Facebook posts using the ArcGIS software. Identifying the location of the tweets and Facebook posts highlights the areas which have the most social media activities.

The excel spreadsheet data (Table 1) with the enumerator values initially contained certain columns and rows with invalid information and duplicates of entries needed to be deleted, prior to running the GIS analysis. There were 18 634 entries left from the original 64 043 entries after deleting the invalid columns and rows. Subsequently, the spreadsheet was loaded onto the ArcGIS software and this allowed for the location where the tweets and posts were made to be visualised and analysed spatially. One of the first analysis to be completed on the data using the ArcGIS software, was the kriging analysis, which requires one to create Z-values using point data and digital elevation model (DEM) data.

3.1.1.1 Kriging and fishnet analysis: Kriging is an interpolation analysis which uses point data to determine the spatial correlation of points in relation to each other (ESRI, 2015). This particular analysis is significant for this study since the study uses social media feeds (in the form of point data) to determine the concentration levels of tweets and Facebook posts made by Gautrain commuters. These concentration levels were used to determine the areas of origin and destination for the Gautrain commuters. Below (Figure 3) is the Kriging mathematical formula which is used to calculate the point data prior to executing the results from the analysis.

$$\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i)$$

where:

$Z(s_i)$ = the measured value at the ith location

λ_i = an unknown weight for the measured value at the ith location

s_0 = the prediction location

N = the number of measured values

Source: ESRI 2015

The Kriging analysis was executed on both the Gauteng boundary and the three municipal boundaries, using a maximum distance of 1000m/1km. This radius distance is large enough to cover the social media feeds from commuters coming in and going out of the Gautrain stations, hence it was used.

There were more analysis' conducted on the social media point data after the Kriging analysis were completed on both the Gauteng province and the three municipal boundaries. These new analysis included the fishnet and counts in polygon which uses the Geospatial Modelling Environment (GME) software. The former analysis was used to create grid cells of a $5000m^2$ area for both the Gauteng and municipal boundaries. The rationale behind creating these grid cells is to ensure that the concentration levels of the point data are calculated on equal grids so as to enhance the accuracy of the results. The latter analysis however, was used to create the cold and hotspots of social media activities through converting point data to concentration levels within the grid cells created using the fishnet analysis. The results obtained from both the fishnet and counts in polygon analysis are useful in identifying the neighbourhoods which have more social media activities as indicated in Figure 6.

4. RESULTS AND DISCUSSION

This section of the paper presents the results obtained from the Kriging, Fishnet and Count in polygon analysis. The results have been presented in the form of maps. Following this is a section that provides an interpretation of the results obtained. The results obtained have been used to infer a rationale and to provide logic for the analysis obtained pertaining to the concentration levels of tweets and Facebook posts.

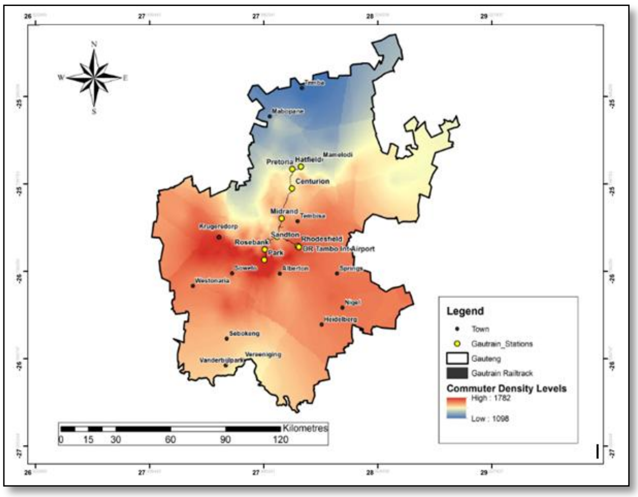

Figure 4: Maps showing the concentration levels of the tweets made in the Gauteng province.

The output from the Kriging analysis is maps indicating commuter density levels. These density levels are measured using high and low concentration values, which represent hot and coldspots, respectively. The hotspots are represented by the light and dark orange colour and the coldspots by the colour yellow and blue. The former are symbolic of the locations with high commuter density levels, whereas the latter are typical of locations with low commuter density levels. All the locations with high commuter density levels have been used to refer to both the origin and destination since the data used for this study did not stipulate the areas of origin and destination of commuters. Hence, it was difficult to identify which areas were the origin and destination of the commuters using social media networks.

Figure 4 demonstrates commuter density levels of social media activities within the Gauteng province. The concentration levels of commuters using social media in the province ranges from the value 1097 to 1782 with the former representing coldspots and the latter hotspots, respectively. The areas shaded in all shades of orange are highly concentrated and have commuter density levels close to the maximum concentration value of 1782. Likewise, the locations shaded in yellow and blue represents commuter density levels which have a minimum concentration value that is close to the commuter density level of 1097.

Running the Kriging analysis on such a large scale as the province produced useful results since the outcome indicated the locations within the province which had high and low commuter density levels. Nonetheless, it was also necessary to run the analysis on a smaller scale so as to identify the neighbourhoods with the most social media activities. Thus, there is Figure 5 & 6, which is the outcome of both the Kriging and Counts in polygon analysis run on the three metropolitan municipalities.

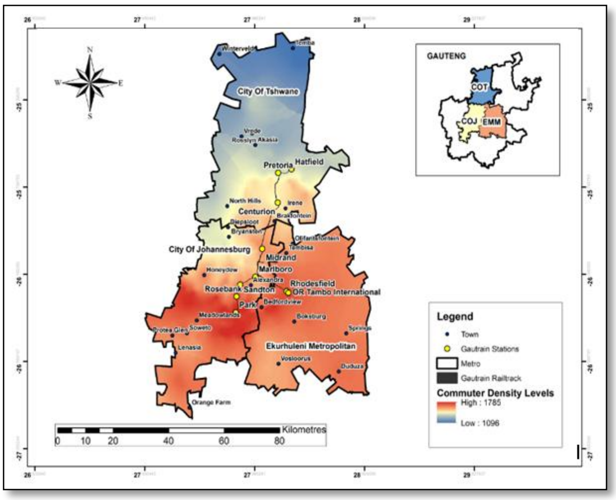

Figure 5: Maps showing the concentration levels of the tweets made in the Gauteng province.

Figure 5 illustrates that the stations with most commuter density level and social media activities are Park, Rosebank and Sandton. The commuter density levels in these stations are high because of the locations which they have been established in. For instance, the Gautrain Park station is located near the largest train station in Africa. This station is commonly known as Park station and it is renowned for offering a variety of modes of transport to commuters from all over the county and Africa.

These modes of transport range from buses to taxis and trains. It is for this reason that Park station has been recognised as a good exemplary of Integrated Transportation System (ITS) as there is an integration of the BRT station, Gautrain and Gaubuses, Taxi's and the Metrorail trains all located within the same proximity.

Both the Rosebank and Sandton stations are located in close proximity to the Rosebank mall and Mandela Square mall, respectively. These are the two most popular nodes in the Gauteng province since they are known to provide a variety of entertainment. Apart from the entertainment aspect, most people travel to and from Rosebank and Sandton on a daily basis because of work. As a result of their popularity, these nodes attract the greater population and since the Gautrain stations are located in close proximity to them, most people choose to travel by the Gautrain to get to their desired destinations.

The City of Johannesburg according to Figure 5 has the most concentrated levels of social media activities as a result of having 5 of the 10 Gautrain stations located within its jurisdiction. The Johannesburg CBD together with neighbourhoods such as Cresta, Parktown, Braamfontein, Ormonde and Randburg to name just a few are some of the commuter density hotspot areas in the municipality. The high concentrated levels of social media activities in these neighbourhoods can be attributed to the fact that most of the neighbourhoods are located close to the Gautrain stations mentioned above. For example, Braamfontein is in the same vicinity as the Gautrain Park station, thus there is such a high concentration level of social media activities.

There are numerous neighbourhoods south of the Johannesburg city which have high commuter density levels. These neighbourhoods include Soweto and Meadowlands which do not have Gautrain services in operation. Therefore, one can assume that the high concentration levels of social media activities about the Gautrain are a result of having the residents of these areas using the Gautrain services to travel up north to places such Centurion, Midrand and the likes. For such people, the Park Gautrain station is the first point of departure and this station can thus be assumed to be their place of origin once they reach the Johannesburg CBD.

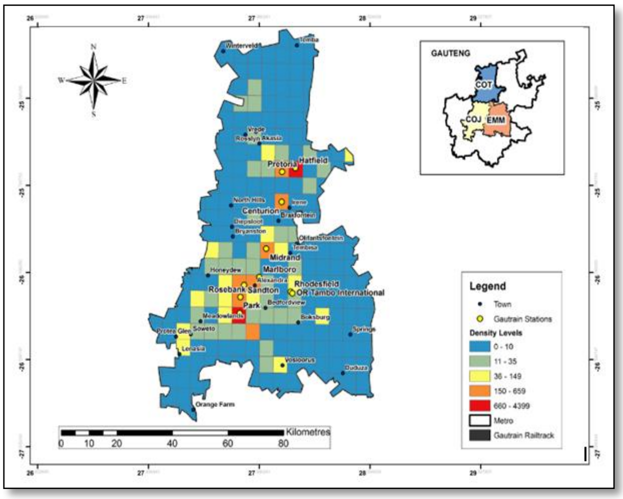

Figure 6: Concentration levels per 5000m-block area.

Figure 6 is the results obtained from running the fishnet and counts in polygon analysis. As mentioned above, both the fishnet and counts in polygon allow for an analysis to be made on a smaller scale by creating the grid cells which enhance the level of accuracy of the results, as shown in Figure 6. The counts in polygon analysis basically calculate the number of points found within a polygon. In terms of this paper, the point data refers to the twitter and Facebook posts which have been calculated within every 5000m^2 grid cells. The results obtained thereof, range from 0-4399 and these represents the commuter density levels.

The results obtained in Figure 6 demonstrates that there are more social media activities taking place in neighbourhoods located within the vicinity of the Gautrain stations. With that said, Braamfontein and Brooklyn are some of the neighbourhoods which are illustrated to have the most highest commuter density levels. These neighbourhoods are located around the Park and Hatfield station, respectively. The number of tweets and posts made in and around these stations range between 660 and 4399. The location of the Hatfield station encourages the use of the Gautrain for people travelling down south to places like Centurion, Sandton and even to the OR Tambo international airport. The Hatfield station is located north of the province and close to the University of Pretoria as well as affluent neighbourhoods such as Hillcrest and Brooklyn and this makes travelling to the south using the Gautrain more time effective as one is able to avoid the traffic congestion found along the N1, especially during peak hours.

The Rosebank, Sandton, Midrand and Centurion stations have commuter density levels ranging from 150 – 659 according to Figure 6. These moderate hotspots indicate that these stations are either the origin or destination stations for the Gautrain commuters. Contrary to these results, are the commuter density levels found in Marlboro, Rhodesfield and OR Tambo International stations. These stations have low concentration levels of social media activities due to the following possible reasons: The Marlboro station is located on the outskirts of a township called Alexander, which is notorious for having high crime rates. The crime aspect could be the determining factor for whether a commuter chooses to use the Marlboro station or not, hence it is a coldspot. Both the Rhodesfield and OR Tambo stations are situated in and around the OR Tambo International Airport and as a result they attract social media activity from tourists coming in and out of the country. These stations serve as areas of origin and destination for those tourists and citizens using the OR Tambo Gautrain station. With that said, it can be assumed that there is a small percentage of population which use the Gautrain in these stations to access places in the east of the province.

The city of Tshwane has low concentration levels of social media activities, despite having one of the largest Gautrain station in the city, which is the Centurion station. This could be a result of having most of the commuters in this city having limited contact with the Gautrain services, hence so many coldspots. Areas such as Bronkhorspruit, Ga-rankuwa and Temba are marginally concentrated thus, representing the coldspots. Such areas are furthest away from the Gautrain services; hence they are less users of the Gautrain located there. It can therefore, be assumed that people residing in those areas use other modes of public transport and do not commute with the Gautrain on a regular basis.

The northern parts of the Gauteng province have low commuter density levels. This may be alluded to the fact that people residing in the northern parts of the COT work in the Tshwane CBD, hence they do not utilise the Gautrain services as much as the people travelling from the south of the Gauteng province. This observation implies that areas such as Ga-rankuwa and Bronkhorspruit are the places of origin for the commuters and the CBD which is located north of the Pretoria station is their final destination. Hence, most of the residents do not have much contact with the Gautrain.

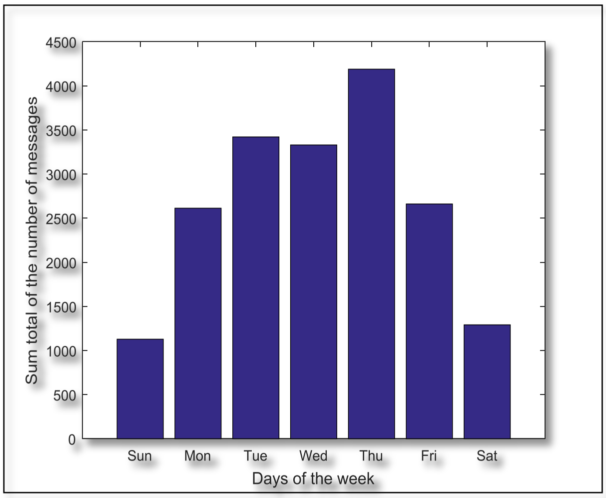

Figure 7: Sum total of the number of messages for each day of the week.

Figure 7 indicates the number of tweets and Facebook posts made by commuters during the six-month study period. It is evident from the graph that most messages about the Gautrain were posted during the weekdays than they were during the weekends. This is most likely because most people use the Gautrain during the week for business purposes such as travelling between home and work as compared to using it for leisure purposes during the weekend. Such information is important as it has the potential to give an indication to company owners of the busiest day of the week. From this information, the operators will then know whether to increase staff or more trains on that particular day or not.

The contents of the messages posted by commuters or consumers are crucial for any business, irrespective of whether it is in the private or public sector. Therefore, their ability to provide real-time data makes social media sites the relevant tools to assist with improving transportation planning by analysing the crowdsourced data. Based on the analysis of the contents of messages, company owners and planning authorities can provide the necessary transportation services and infrastructure in areas where most complaints are coming from. The contents of the tweets and facebook posts could also be used to inform transportation planning authorities in the public sectors of the need for providing services such as the Gautrain app and timetable for other modes of transport used in the Gauteng province. Such features are necessary as they synchronise the times of the public transportation systems thus allowing consumers to plan their travelling trips accordingly.

5. LIMITATIONS OF THE STUDY

The Echo-social software used to provide data for this paper has highlighted some weaknesses, which may be assumed to be generic with other big data software. The data obtained from the Echo-social software contained various rows and columns of data which were invalid. Consequently, the data had to be cleaned to enhance accuracy in the analysis. Similarly, there where cases where the location (coordinates) could not be verified, because users deactivated the location features on their cell phone. Deactivating the location feature on ones smartphone could pose a challenge for companies or authorities who would like to obtain data on the locations of their commuters or citizens.

Another weakness identified was that the location points representing the tweets and Facebook posts did not specify which areas are the origin and which are the destinations. Therefore, it became difficult to separate the two, once the Kriging analysis was conducted. This Kriging analysis highlighted the commuter density levels of social media activities for the Gautrain commuters. And as a result, the study assigned the concentrated areas to represent both locations of origin and destination.

Lastly, ward data for Gauteng was used to identify which wards have more concentrated levels of social media activities. Wards are not similar in size hence the results may be distorted. Consequently, a fishnet analysis was therefore conducted in order to assign equal grid cells to the municipality boundary so as to obtain more accurate results.

6. IMPLICATION FOR PLANNING AND CONCLUSION

The content of the social media data obtained from Echo-social proved to be proficient tools to provide an insight on the travelling behaviours of commuters. This data was used in a density analysis called Kriging, which enabled one to identify areas which areas have a concentrated level of social media activities and commuter density. This information could thus, be used in transportation planning to alert and inform planning authorities about roads which are commonly used and congested. Furthermore, the times of day for which the congestion takes place can also be brought to attention. In light of this information, authorities may choose to expand lanes, establish new road network links or work on improving the already existing public transportation systems as an attempt to reduce traffic congestion.

The areas which were found to be highly or moderately concentrated with social media activities need to be assessed for possible Gautrain extension sites. The concentration levels could be an indication that there is a threshold of commuters who could benefit from having the Gaubus services, extended to those areas. According to the results obtained, these potential areas include Soweto, Randburg and Springs, to name a few.

Social media could also be used as a platform for civil society to voice out their level of satisfaction (or lack of) with regards to the services provided by the government and business entities. This is a cheap and convenient way of collecting data from commuters without having to leave the comfort of their homes. Such mediums could also be used to provide recommendations in terms of how to improve transportation planning in South Africa based on the views and opinions of commuters from the Gautrain, as it is renowned for being the first high speed train in the country. This kind of data may probe authorities to improve on transportation facilities and infrastructures for other modes of transportation. From a business aspect, big data could be used to identify customer profile and their respective places of origin.

In conclusion, social media data which contains information on the location of origin and destination of commuters, accurate times and all the other information included in Figure 3, has the ability to be analysed and thus be used to draw travelling pattern and behaviours of commuters.

REFERENCES

Blatt, A., J. (2015). Data Privacy and ethical uses of volunteered Geographic Information. *Health, Science and Place, Geotechnologies and the Environment,* 12: 49-59.

Bohte, W. & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Policy*, 17, pp. 285-297.

Callister, T., A. (2000). Media Literacy: On-ramp to the literacy of the 21st century or cul-de-sac on the information superhighway. *Advances in Reading/Language Research*, 7: 403-420.

Caragliu, A., Del Bo, C. & Nijkamp, P. (2009). Smart Cities in Europe. University of Amsterdam: Amsterdam.

Chen, M., Mao, S. & Liu, Y. (2014). Big Data: A Survey. *Mobile Network Application*, 19, pp.171–209.

City of Johannesburg. (2012). Integrated Development Programme. Johannesburg: Gauteng.

Comscore. (2011). Comscore press release: Owners access check-in services via their mobile devices. Available from: http://www.comscore.com/Press_Events/Press_Release/2011/5/Nearly_1_in_5_Smartphone_Owners_Access_Check-in_Services _Via_ Their_ Mobile (12th May 2015).

Dassani, N., Nirwan, D. & Hariharan, G. (2015). Dubai- a new paradigm for smart cities. KPMG.

de Abreu Freire, C., E. & Painho, M. (2014). Development of a mobile mapping solution for spatial data collection using open-source technologies. *Procedia Technology*, 16, pp. 481-490.

Douangphachanh, V. & Oneyama, H. (2014). A study on the use of smartphones under realistic settings to estimate road roughness condition. *Networking*, 1(114), pp. 1-11.

ESRI. (2015). Point density (spatial analyst). Available from: http://resources.arcgis.com/en/help/main/10.1/index.html#//009z0000000v000000 (25 July 2015).

European Commission. (2010). Intelligent Transportation Systems: EU-funded research for efficient, clean and safe road transport. European Union, Belgium.

Farah, J. (2014). Crowdsourced monitoring citizens empowerment and data credibility.

Fernandes, J., Oliveira, P., Silva, C. & Marcelino, L. (2012). Route Social Network. *Procedia Technology*, 5,pp.547-555.

Filippi, F., Fusco, G. & Nanni, U. (2013). User empowerment and advanced public transportation solutions. *Procedia-Social and Behavioural Sciences*, 87, pp.3-17.

Gao, S., Yang, J., Yan, B., Hu, Y., Janowicz, K & McKenzie, G. (2012). Detecting origin-destination mobility flows from geotagged tweets in greater Los Angeles area.

Gautrain Management Agency. (2010). Gautrain Annual Report. GMA: Gauteng.

Hasan, S. & Ukkusuri, S., V. (2014). Urban activity pattern classification using topic models from online geolocation data. *Transportation Research* Part C, 44. pp.363-381.

Hongyan, G. & Fasheng, L. (2013). Estimating freeway traffic measures from mobile phone location data. *European Journal of Operational Research*, 229, pp.252-260.

Hu, S. & Liou, H. (2014). A generalized sensor location model for the estimation of network origin-destination matrices. *Transportation Research*, 40, pp.93-110.

Jin, P., J., Chebalek, M., Yang, F., Ran, B. & Walton, M. (2014). Location based social networking data: an exploration into the use of a doubly constrained gravity model for origin and destination estimation.

Kaya, S., Kilic, N., Kocak, T. & Gungor, C. (2014). From Asia to Europe: Short-term traffic flow prediction between continents. International Conference on Telecommunications (ICT).

Kling, F. & Pozdnoukhov, A. (2012). When a city tells a story: Urban Topic Analysis.

Lu, C., Zhou, X. & Zhang, K. (2013). Dynamic origin-destination demand flow estimation under congestion traffic condition. *Transportation Research*, 34, pp.16-37.

Loebal, M. (2012). Is Privacy Dead ? – An inquiry into GPS Based Geolocation and Facial Recognition Systems. *International Federation for Information Processing*, 386: 338-348.

Peters, K., Chen, Y., Kaplan, A., M, Ognibeni, B. & Pauwels, K. (2013). Social Media Matrics-A framework and guidelines for managing social media. *Journal of Interactive Marketing*, 27, pp. 281-298.

Republic of South Africa. (2008). Non-motorised policy. Statistics South Africa: Pretoria.

Turk Telekom International. (2015). Smart Mobility. Available from: http://www.turktelekomint.com (01 October 2015).

Todes, A. (2011). Reinventing planning: Critical reflections. *Urban Forum,* 22, pp. 115-133.

Republic of South Africa. (2011). Census 2011. Statistics South Africa: Pretoria.

Wechsler, D. (2014). Crowdsourcing as a method of transdisciplinary research-Tapping the full potential. *Futures*, 60, pp. 14-22.

Xu, J., Kang, Q., Song, Z. & Clarke, C., P. (2015). Applications of mobile social media: WeChat among academic libraries in China. *Journal of Academic Librarianship*, 41, pp.21-30.

Zhou, X. & Mahmassani, H., S. (2007). A structural space state model for real-time traffic origin-destination demand estimation and prediction in a day-to-day learning framework. *Transportation Research*, 41, pp.823-840.