USING SOCIAL MEDIA INFORMATION IN TRANSPORT- AND URBAN PLANNING IN SOUTH AFRICA

Quintin VAN HEERDEN [1]

[1] Built Environment, Council for Scientific and Industrial Research, Email: QvHeerden@csir.co.za

Keywords: data-mining, social media, transport planning, urban planning

Abstract

Transport- and urban planning often involves the use of large-scale simulation modelling with data sources that include, among others, the national census, the National Household and Travel Survey (NHTS) as well as trip diaries. Many of these sources of data are costly and time consuming to obtain, clean, and analyse. This paper explores the significance of social media platforms as a source of open, more accessible data to infer people's movement and land-use patterns, by focussing on the City of Tshwane Metropolitan Municipality in the Gauteng Province of South Africa. Users of social media platforms, such as *Twitter*, *Foursquare* and *Flickr*, post millions of publicly available updates each day. Some of these postings have spatial-temporal characteristics in the form of geo-tagged, time-stamped metadata fields. In this paper, data-mining techniques are used to obtain, extract, and analyse data from social media to illustrate the usefulness of such open data in the transport- and urban planning domains in South Africa.

1. Introduction

Transport- and urban planning are vital to ensure appropriate spending of funds for infrastructure development and maintenance, to ensure sustainability in cities, and to enhance a country's economic competitiveness. To do effective planning for enhanced service delivery and to influence urban policy, transport- and urban planners endeavour to understand how different parts of the transport- and urban systems affect one another. A common tool used in these environments is that of simulation modelling, which allows one to evaluate the effect of changes in the transport- and urban systems on, for instance, expected travel times and land-use patterns in the future.

In both transport- and urban simulation models, the required data are vast and often very difficult to obtain. Data which might be more easily obtainable through, for instance, surveys, tend to be very expensive. The time and effort involved in collecting data on travel behaviour from trip diaries is also cumbersome. Furthermore, these large-scale simulation models are notorious for the long time needed to develop it and due to its complexity often has a simulation run-time of several days. The time it takes to develop and run these models is, however, justified seeing that these models should inform multi-billion rand infrastructure investment decisions. Yet, we need to find alternative, quicker, cheaper and reusable ways to gain insight into people's movement and land-use, which could complement existing methodologies that produce the input to these simulation models as well as provide other insights relating to the transport- and urban planning domains.

It is possible to exploit publicly available data in the form of social media. Millions of postings are made every day on social media networks such as *Twitter*, *Foursquare*, *Flickr*, and *Instagram*. The metadata linked to these postings contain very detailed information that could be utilised once it is reworked into a usable format. While many studies have been conducted abroad using social media as a source of information for research purposes, South Africa has yet to catch up with this trend, especially in the transport- and urban planning domains. One of the major reasons often cited for this phenomenon is that a large portion of the South African population is very poor (StatsSA, 2011), hence chances are that cellular phone usage, and more specifically the use of social media on cellular phones, would not be high enough, or else biased towards the higher income groups. According to the *Social Media Landscape 2015* report by World Wide Worx (2015), a large portion of the South African population still accesses *Facebook* by means of a basic cellular device (not a smartphone). However, in the same report it is stated that *Twitter* had an increase of 20% in its user base from South Africa, with a total of 6.6 million users at the end of 2014. For this reason, this paper explores whether social media, specifically *Twitter*, can be used to obtain insights into the transport- and urban planning domains in South Africa through the development of smart and sustainable procedures.

The main focus of this paper is to evaluate how publicly available data from social media can be extracted and used with other datasets to gain insight into people-movement and land-use patterns in both the transport- and urban planning domains. A secondary focus is to evaluate the usefulness of such data in a developing country, such as South Africa, where smartphone user growth is slower than in developed countries. The next section gives an overview of the literature relating to social media data mining in the transport- and urban domains. Thereafter the methodology is described of how the datasets were obtained and prepared for use in this paper. Section 3 sheds light on the usefulness of the data obtained from *Twitter* by means of analyses, illustrations, and a machine learning model, by using the City of Tshwane

Metropolitan Municipality as the study area. Concluding remarks are made with reference to sustainability and further possible avenues of research are discussed in Section 4.

2. Literature Review

Transport- and urban planners are concerned with, among others, knowing how people make choices as to where to live, where to work, and how to move between these (mostly) different locations. Insights into these choices could potentially assist in improving service offerings to these individuals. While many insights into travel patterns and transport needs could be gained from surveys, such as the National Household Travel Survey (StatsSA, 2013), these surveys are extremely costly and time-consuming to conduct. Social media provides an avenue to obtain some insights into travel and land-use patterns from publicly available data.

In the transport planning domain, Gal-Tzur *et al*. (2014) investigate the potential to data-mine social media for valuable data to inform transport policy makers. They report that this type of data can assist in overcoming transport challenges such as parking needs or bottlenecks in the transport system. In another study, Gao *et al*. (2014) investigate the possibility of using large-scale social media data for the estimation of origin-destination trips. They show favourable estimates of mobility flows and compare it with survey data.

In the urban planning domain, Noulas *et al*. (2013) combine telecommunication data with locations obtained from Foursquare, a search and discovery service of points-of-interest. The authors infer the types of activities in neighbourhoods by data-mining these points-of-interest from *Foursquare* and linking it to a dataset of cellular signals from a telecommunications provider. Using a machine learning algorithm, they are able to predict the land-use of zones with favourable confidence.

Ferrari *et al*. (2011a) identify hotspots in the city by analysing urban patterns in a *Twitter* dataset. They were able to determine recurring crowd behaviour and link this behaviour to common locations. Ferrari *et al*. (2011b) further discover routine behaviours and patterns from a dataset obtained from *Google Latitude*. Being able to predict routine behaviour of users would be beneficial in multiple environments, for instance targeted advertising and marketing.

Frias-Martinez *et al*. (2012) characterise urban land-use by analysing geotagged tweets in the Manhattan area of New York City. They infer points-of-interest as well as land-use patterns from geotagged tweets with favourable accuracy. Furthermore, Crandall *et al*. (2009) detect landmarks from geotagged photos from *Flickr*, a photo management and sharing application. The authors plotted the GPS locations of photos to create maps of tourists' movements.

Zheng *et al*. (2012) analyse the movements of tourists through a city by also data-mining *Flickr* for geotagged photos. They specifically look at tourists' locations relative to points-of-interest as well as the travel paths taken by tourists between these points. Such information could be useful, for instance, in the design of public spaces to allow for better walkways or means of transport between these points-of-interest, based on the frequency of visits to these places.

Zagheni *et al*. (2014) use *Twitter* data to infer both intra-zonal and international migration patterns. Schneider *et al*. (2013) mention that daily mobility patterns can be attributed to a set of 17 motifs. They use spatial-temporal trajectories in combination with network theory to determine these motifs and claim that individuals exhibit a very specific motif that influences his or her travel behaviour, which could be inferred from data.

Some of the methodologies in this literature review are used in this paper. The purpose of this paper is not to develop perfect-fitting models, but rather to illustrate by means of practical examples and further discussions how these techniques can be used in a smart and sustainable manner on social media data for developing countries as well.

3. Research Methodology

The data that were used for this paper were obtained and prepared from various sources. Each dataset and the steps involved to prepare the data for use for this paper are described in the following subsections.

3.1 Study area

The study area used in this paper is the City of Tshwane Metropolitan Municipality (hereafter also referred to as the City of Tshwane) in the Gauteng Province of South Africa. Gauteng is considered to be the economic hub of the country, which should result in higher cellular phone usage due to better affordability, while the City of Tshwane, an area that includes Pretoria but extends further to the east, exhibits attributes of both the developed and developing world. The area was subdivided into *sub-places,* which are geographical areas determined and used by Statistics South Africa, which generally demarcates suburbs or villages. Figure 1 depicts the sub-places and where the City of Tshwane is located in South Africa.

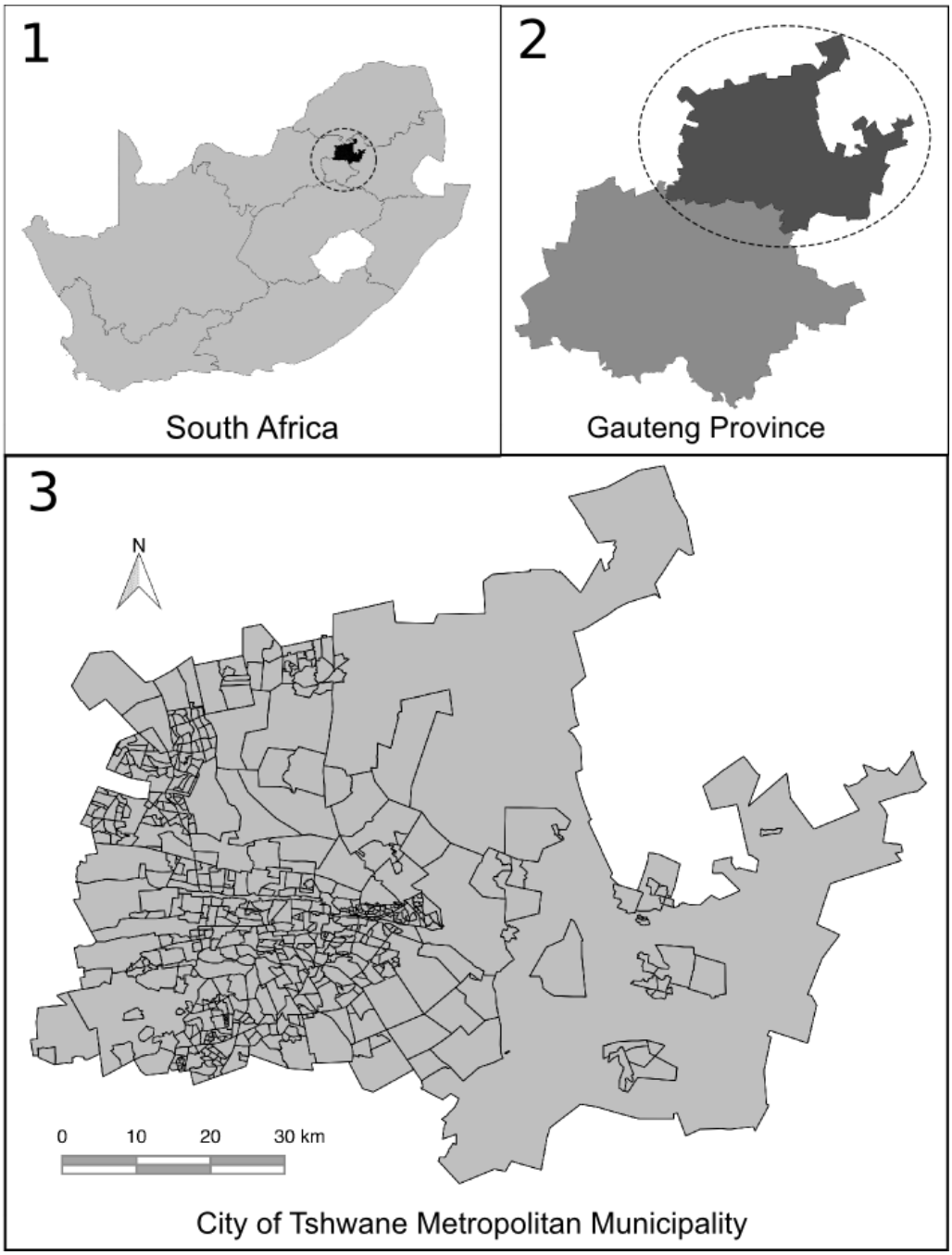

Figure 1 Sub-places in the City of Tshwane, Gauteng, South Africa

3.2 Twitter dataset

The process to obtain the *Twitter* dataset involved an implementation of the *Streaming* application program interface (API) of *Twitter*, which is essentially a means to obtain data from a live *Twitter* feed. This API was used to capture tweets in near real-time during the period of 24 April 2015 – 25 June 2015, which equalled 63 days. A total of 2 821 338 tweets was obtained. Each tweet contains a large amount of metadata in JSON format, but for the purposes of this paper only four parts from each tweet were stored for further analysis: the unique tweet identification number, the username (anonymised on receipt) of the user that posted the tweet, the longitude and latitude of the tweet (if the user enabled the geotag option), and the date stamp. For the tweets that contained a coordinate, the location of the tweet was used to determine within which sub-place the tweet originated from. The sub-place ID was subsequently also stored with the tweet. Table 1 contains a summary of the fields that were stored for each tweet.

Table 1 Summary of data stored for each tweet

Field	Description
Tweet ID	Unique ID used by *Twitter* to distinguish between tweets
Username	An anonymised ID string (originally the username of the person who posted the tweet, which begins with a "@")
Coordinate	The longitude and latitude of the tweet, if it was available
Date stamp	Precise date and time of the tweet
Sub-place	If the tweet had a coordinate, the ID string that corresponds to the Sub Place ID in the Shapefile was added to the dataset

From the original dataset, only 1 151 133 of the tweets were geo-tagged (contained a coordinate in its metadata), and finally only 146 586 tweets were within the confines of the City of Tshwane (the study area). Only sub-places that had at least 150 tweets were considered for further analysis. For each sub-place, tweets were grouped into 72 bins, which correspond with 72 equally spaced twenty-minute intervals in the day. The frequency of tweets in each bin was determined and normalised, which represent proportions of tweets throughout the day.

3.3 Transport network

For the purposes of this paper, only public transport networks were considered. The routes and itineraries of buses, trains, and mini-bus taxis that operate in the City of Tshwane were obtained from the City of Tshwane Metropolitan Municipality.

3.4 Land-cover dataset

The *GeoTerraImage* (GTI) 2013-2014 land-cover dataset was used as a starting point to aggregate the land-cover classes to sub-place level. This land-cover dataset comprises 72 classes that describe the man-made and natural landscape characteristics of 30mx30m raster cells for the whole of South Africa. For the purposes of this paper, only *built-up* classes (areas covered by houses or other buildings) were considered, which included 31 of the 72 classes. The *Commercial* and *Industrial* classes were used as-is. The other classes in the dataset are subdivided according to the total surface area being dominated by trees, grass, or bare surface. These secondary classes were aggregated to the primary class level, resulting in 8 primary classes, namely: Village, Urban Residential, Commercial, Township, Industrial, Urban Informal, Mining, and Other. For each sub-place, a percentage for each of the land-cover classes was assigned, which represents the proportion of land that belonged to that class. Each sub-place was assigned a land-cover class attribute based on the land-cover type that covers the highest proportion of land in the sub-place. Many of the sub-places have mixed land-cover, but for the purposes of this paper, only the dominant land-cover was considered. The dominant class was therefore assigned to the sub-place in the dataset. Figure 2 depicts the number of sub-places that were classified in each land-cover class, which shows that Urban Residential is by far the dominant land-cover class. The relevance of the distribution of the classes will be discussed further in the section relating to the use of social media data to predict land use.

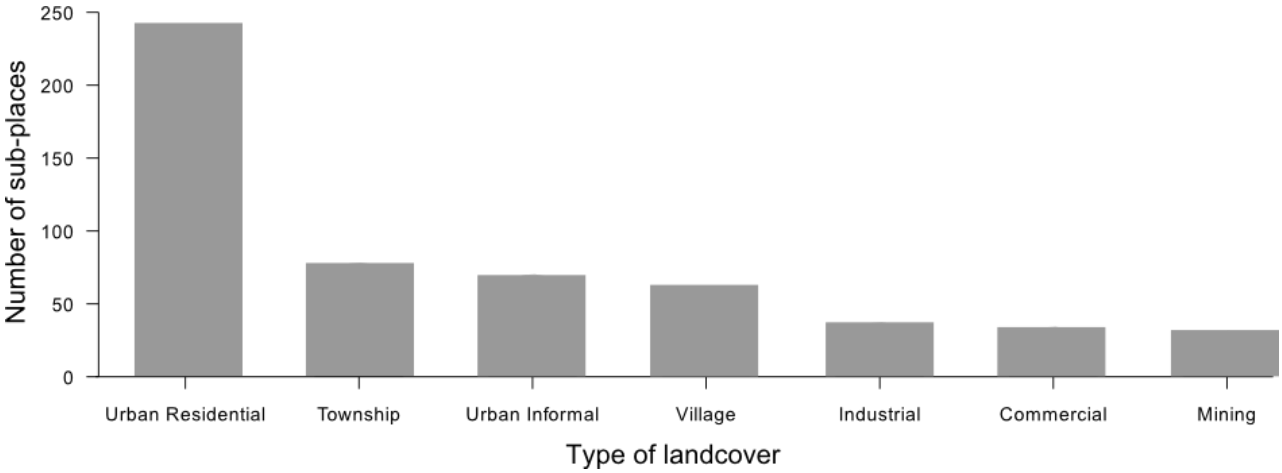

Figure 2 Count of the dominant land-cover type per sub-place

The *Twitter* dataset was accordingly updated to include the land-cover class. The final dataset therefore consisted of: a sub-place ID, 72 bins of normalised tweet activity, and a land-cover class.

4. Findings and discussion

As mentioned in the introduction, the study aims to derive two key information sets from the social media data, which are land-use and travel patterns. The discussion will focus on the evaluation of tweet patterns at different times of the day and how these relate to land use (the place from where tweets originate), and then testing whether land-use can be derived from the tweet data.

4.1 Tweet activity signatures and land-use patterns

It is possible to distinguish between different regions in a study area based on the tweet activity in the region, which can also be referred to as the tweet activity signature (Frias-Martinez, 2014). Regions with similar social network behaviour should have similar tweet activity signatures. Similarly, regions with differing social network behaviour should have different tweet activity signatures.

Two land-cover types, Commercial and Urban Residential, were analysed. For both these land-cover types, the tweet frequencies for all sub-places with the corresponding land-cover class were aggregated and normalised to obtain an overall view of daily tweet activity. **Error! Reference source not found.** depicts the social profiles/tweet activity signatures for these two land-cover classes.

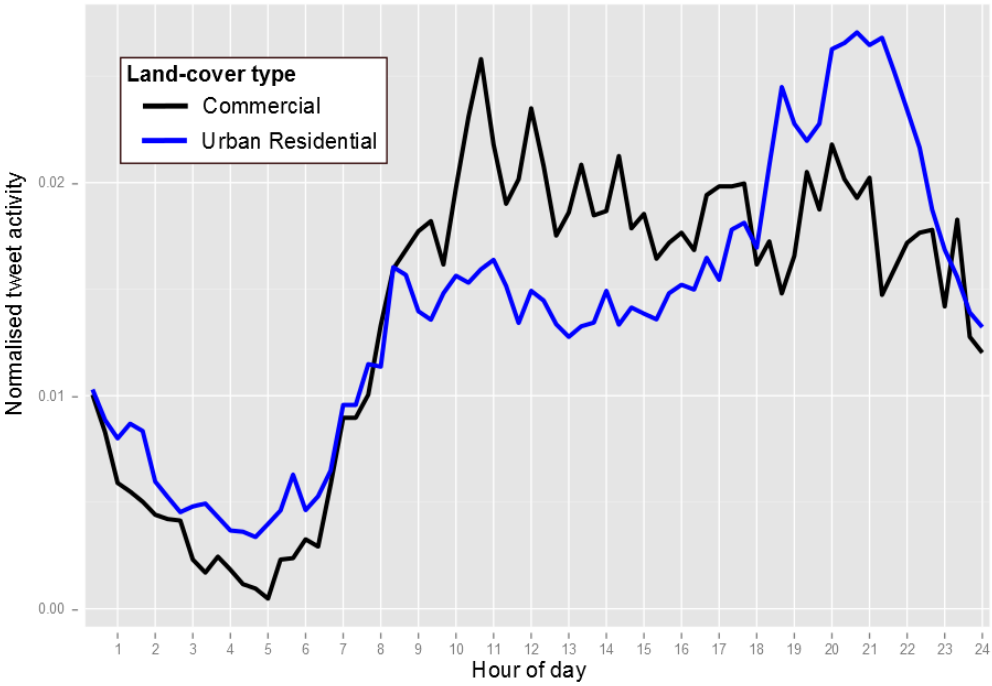

Figure 3 Tweet profiles of different land-cover types

From **Error! Reference source not found.**, it is evident that the tweet activity signatures for Commercial regions differ from that of Urban Residential regions. In Commercial regions, peak tweet activity is observed in the late morning hours between 09h00-11h00, which corresponds with the time that people arrive at work. Secondary peaks are observed between 11h00-13h00 and 17h00-18h00, which could relate to lunch time breaks and people leaving work to return to their homes. The Urban Residential tweet activity signature sees a peak activity in the early to mid-evening hours, which is when many of the prime-time series appear on television.

A simple text-mining procedure was followed to extract and clean the text from the tweets in these two regions. Two word-clouds were generated from the text to compare the content of the tweets, as is shown in **Error! Reference source not found.**. Words that occurred more often appear larger and darker. It is evident that the tweet content of these two land-cover classes is similar with many of the most frequent words present in both word-clouds. There are, however, some differences: place names such as “Pretoria” and “Hatfield” are more prominent in the Commercial area, whereas in the Residential area, more colloquial language is found, for example “lol” (which normally means “laugh out loud”). This information could be used for sentiment analysis and to track opinions of citizens that could in turn be used for enhanced service delivery.

Figure 4 Word-clouds of tweet content in Commercial (left) and Urban Residential (right) areas

4.2 Connectivity and information flow

Using the techniques from the previous section, it would be possible in future research to develop a complex network that describes the connectivity between different users in different regions. Furthermore, taking tweet content into account would allow researchers to gain a better understanding as to the subject matter of conversations between different users and areas. These insights could be used to better understand prominent role players in the social network as well as how information flows throughout the network and how individuals influence one another in the network.

4.3 Predicting certain land-use classes

In the previous subsections, a distinction was made graphically between tweet activity signatures of different land-cover classes. In the case where the land-cover class of an area is not available, there would be value in being able to predict the land-cover class given a certain tweet activity signature. To show how one could typically go about doing this, a classification model was developed.

The Twitter dataset containing the tweet patterns per sub-place as well as the dominant land-cover class was used as an input to the model. Stratified cross-validation was used with a 10-fold cross-validation in the WEKA software (Hall *et al*., 2009). That is, the dataset is randomly partitioned into 10 samples of equal size, of which 9 samples are used to train the model and the remaining 1 sample to test the model. Table 2 contains the confusion matrix, which shows the results of the model by comparing the predicted number of sub-places classified per land-cover type versus what the actual land-cover of the sub-places is.

From Table 2, it is evident that the Urban Residential land-cover class was predicted with almost 100% accuracy. Unfortunately, this model could not accurately predict other classes, except for the Commercial class that had an accuracy of 20%. This could be due to the fact that many of the sub-places have mixed land-use and when one land-cover type is specified as the dominant type, the second most prominent type may still cover a fairly large percentage of the sub-place. Furthermore, 213 of the sub-places have a dominant land-cover of Urban Residential, which equates to 62% of the total number of sub-places that were used in the model. Furthermore, many sub-places were omitted from the model, since they contained fewer than 150 tweets (the threshold that was used to include or omit sub-places). Frias-Martinez *et al*. (2012) similarly only focused on the main land use in each of the clusters in their study and obtained a classification accuracy of between 50%-70%. Future models would need to consider and predict mixed land-use.

Table 2 Confusion matrix

Actual class	Classified as						
	Village	Urban Residential	Commercial	Township	Industrial	Urban Informal	Mining
Village		31					
Urban Residential	1	212					
Commercial		4	1				
Township		48					
Industrial		7					
Urban Informal		40					
Mining		1					

Another possible explanation for the error in classification of the Township, Urban Informal, and Village classes could be that they exhibit a similar tweet activity pattern to that of the Urban Residential class. While

a distinction is often made between these land-use classes in terms of demographics or household characteristics, the social media activity patterns tend to be quite similar.

Frias-Martinez *et al*. (2012) use an unsupervised learning approach to detect different classes from Twitter activity patterns. They show that traditional land-use classes can be extended to include new classes such as night life. Using such an approach could bring interesting new classes to the fore. An unsupervised learning approach could also assist in understanding whether there are indeed similarities in the tweet activity patterns between different classes, such as Urban Residential, Townships, Villages, and Urban Informal regions.

While the model only predicted the Urban Residential class with favourable accuracy, there is still merit in developing such classification models if more data are used and the dominant land-cover class is split into mixed land-use classes. In the urban modelling domain, multiple datasets are used as input into urban growth simulation models, of which one is a cadastre dataset that contains information on all parcels of land in an area. The cadastre dataset is sometimes incomplete. Being able to predict land-use from tweet activity patterns would be beneficial in such a case to complement an incomplete dataset. However, cadastres are typically very small, hence enough data would be needed that fall within the confines of the cadastre in question to be able to accurately predict the land-use from the tweet activity.

4.4 Travel patterns

The users who posted tweets were anonymised before further analyses were done on the dataset. One user was randomly chosen from the dataset and the locations of the tweets of this user were extracted into a subset with which to determine travel patterns.

Firstly, these tweets were subdivided into three parts based on the time-of-day that the tweets were posted. All tweets between 19h00-08h00 were grouped, pre-empting that these would typically be generated from home. This group will be referred to as the night-time tweets. Next, all tweets between 09h00-17h00 were grouped, pre-empting that these tweets would typically be generated from a place of work during normal working hours, if the person in fact is employed. These tweets will be referred to as day-time tweets. Finally all other tweets were grouped into a miscellaneous group.

For the night-time tweets, the locations were repeatedly found in the same vicinity. For this reason the locations were clustered and a heatmap was generated from these locations. This heatmap is shown in blue on **Error! Reference source not found.**. The darker the colour, the higher the frequency of tweets per square kilometre. This area is in fact classified as an Urban Residential area from the land-cover dataset, therefore one can infer that the user's place of residence ought to be in this area. Further investigation revealed that this is also a lower-income residential suburb.

The same methodology was used for the day-time tweets since recurring locations were found in the dataset of the user during the day as well. The cluster and corresponding heatmap is shown in red in **Error! Reference source not found.**. This area is classified as a Commercial area from the land-cover dataset and upon further investigation it was determined to be the Central Business District (CBD) in the City of Tshwane. From these two clusters, it can be inferred that the person resides in the residential area west of the CBD and is possibly employed in the CBD.

Next, the remaining tweets were analysed to ascertain how the person travels between these probable locations of residence and employment. A layer was added to the map that contains the minibus taxi and bus service routes. The locations of tweets close to the user's probable residence were in close proximity to public transport services. Subsequent tweets (indicated by the numbers 1 and 2 on Figure 5) were also in close proximity to public transport services that operate towards the CBD. The location of tweets close to the user's probable place of employment was also in close proximity to the public transport services. Due to the fact that the user resides in a lower-income residential area, the chances are higher that the user utilises public transport to commute between place of residence and employment and if this is the case, it is possible that the user makes use of a combination of taxi and bus services.

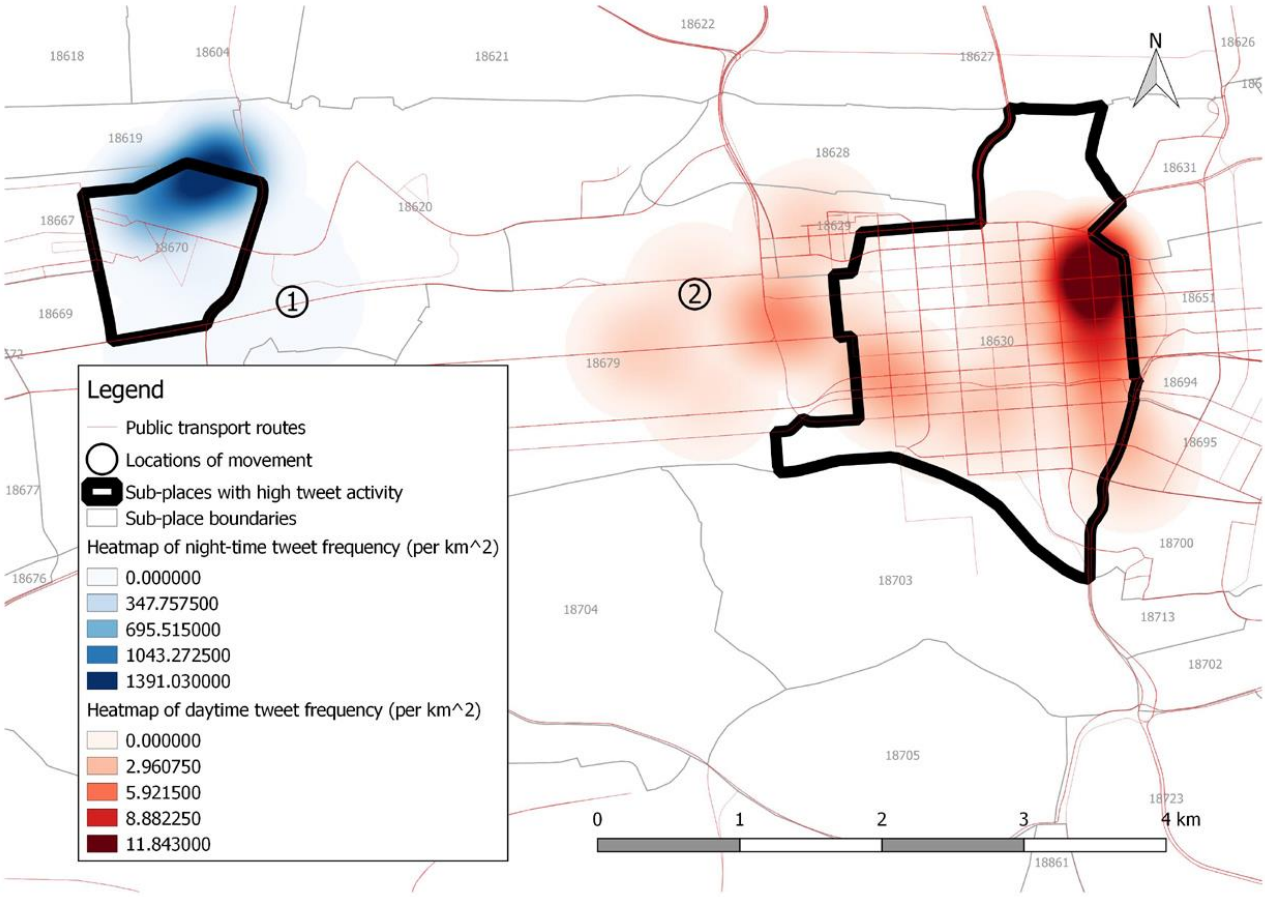

Figure 5 Travel pattern of 1 person

By analysing the tweet locations of this user, it was possible to construct a possible scenario of the mobility patterns of the user over a 2-month period. It might be that the user was merely seeking employment or that this was a temporary pattern. By observing these patterns over longer time-periods, one might be able to detect migration patterns if the locations of the users change. Zagheni *et al*. (2014) for instance, utilise such movement to infer migration patterns both internally and internationally. While the authors could not infer migration rates at specific points in time, they were indeed able to predict turning points in the migration trends.

These mobility patterns can further be analysed by distinguishing between typical weekdays and weekends such as in the study of Herder & Siehndel (2012). The authors show how activity patterns differ between these periods by using GPS logs. It is also shown that one of the users in the study moves mostly between 4 different locations of which the 2 most prominent locations would generally be the residence and employment locations. Similarly, this could be applied to distinguish between the mobility patterns of multiple Twitter users if enough data are available. Being able to predict routine behaviour of users would be beneficial in multiple environments, for instance, targeted advertising and marketing.

In the 2011 National Census in South Africa, locations of employment were not captured. Therefore, another useful application would be to determine people's places of employment from Twitter data, as was shown in this section.

5. Conclusion and possible further research

Proper transport- and urban planning are required to ensure urban growth, economic competitiveness, and sustainability in cities. To test different policy scenarios, large-scale transport and urban simulation models are often used, which require vast amounts of data and intricate modelling skills. Datasets are sometimes difficult to source, expensive to obtain through surveys, or incomplete. For these reasons alternative sources of data should be explored. Data-mining of social media is a common avenue to extract such datasets from publicly available data.

This paper gave an overview of some of the techniques that exist in the literature to obtain data from social media and to gain insights from such data. The techniques included supervised and unsupervised machine learning approaches, which range from classification of land-use classes to gaining insights into people's mobility between points of interest. This paper aimed at illustrating how some of these techniques, and combinations thereof, can be used in a developing country such as South Africa as well. The focus was on the transport- and urban planning domains with a case study of the City of Tshwane Metropolitan Municipality in Gauteng, South Africa.

It was shown that different land-use classes have different tweet activity signatures that distinguish them from one another. Furthermore, a classification model could predict the Urban Residential land-cover class

with almost 100% accuracy, but due to limited Twitter data in the other land-cover classes, the prediction did not perform that well in the other cases. This work can be extended firstly with a larger dataset, and secondly with an unsupervised approach, such as in the study of Frias-Martinez et al. (2014), to obtain areas with similar or different tweet activity signatures, purely based on the activity signatures of the areas. These classes can then be compared with actual land-cover classes to determine which classes share similar patterns.

A complex network approach could also be followed in future research to understand the connectivity between both individuals in the social network and areas that are geographically separated. It would further be possible to determine the influence and roles in the network by using graph theory.

Similar techniques could be used to mine data from Strava, a social network that allows cyclists and joggers to upload their geo-enabled exercise activities to a platform for further analysis. It is possible to obtain the GPS logs of athletes from Strava and infer movements from these in an area. Such analyses would be useful, for instance, for Non-Motorised Transport (NMT) projects. One drawback might be that developing countries would not have enough cyclists or joggers that track and upload their activities to such a service due to affordability constraints.

The techniques that were discussed as well as those that were demonstrated all use data that were obtained from social media through data-mining. These procedures are *smart* since they are innovative means to obtain publicly available data and rework these to a usable format, which in turn could be used in models that inform transport- and urban planning. Since many social media services provide APIs with which to connect to the service and obtain data from it, standard and automated procedures could be developed to obtain, reformat, and use the data. These procedures would be *sustainable* since they can be reused with minimal intervention. The derived information from these procedures could then be used to inform and influence management of services in the built environment.

6. Acknowledgement

The author wishes to thank Pierre du Plessis and Andre Breytenbach for their assistance with the land-cover dataset and their inputs as well as Gerbrand Mans for his input.

7. References

Crandall, D.J., Backstrom, L., Huttenlocher, D., & Kleinberg, J. 2009. Mapping the world's photos. In: *Proceedings of the 18th international conference on World wide web* (WWW '09). ACM, New York, NY, USA, 761-770. DOI=10.1145/1526709.1526812

Ferrari, L., Rosi, A., Mamei, M. & Zambonelli, F. 2011a. Extracting urban patterns from location-based social networks. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, pages 9–16. ACM, 2011.

Ferrari, L. & Mamei, M. 2011b. Discovering daily routines from google latitude with topic models. In Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on, pages 432–437. IEEE, 2011.

Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-martinez, E. 2012. Characterizing Urban Landscapes Using Geolocated Tweets. In: International Conference on Social Computing (SocialCom) and Privacy, Security, Risk and Trust (PASSAT), pp 239-248, ISBN: 978-1-4673-5638-1.

Gal-Tzur, A., Grant-Muller, S.M., Minkov, A., and Nocera, S. 2014. The Impact of Social Media Usage on Transport Policy: Issues, Challenges and Recommendations, Procedia - Social and Behavioral Sciences, Volume 111, 5 February 2014, Pages 937-946, ISSN 1877-0428, http://dx.doi.org/10.1016/j.sbspro.2014.01.128.

Gao, S., Yang, J-A., Yan, B., Hu, Y., Janowicz, .K. and McKenzie, G. 2014. Detecting Origin-Destination Mobility Flows From Geotagged Tweets in Greater Los Angeles Area. Working Paper.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H. 2009. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Herder, E. & Siehndel, P. 2012. Augmented User Modeling UMAP 2012, Montreal, Canada

Noulas, A., Mascolo, C., & Frias-Martinez, E. 2013. Exploiting Foursquare and Cellular Data to Infer User Activity in Urban Environments. 14th International Conference on Mobile Data Management (MDM), Vol 1, pp 167-176, ISBN: 978-1-4673-6068-5.

Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C. 2013. Unravelling daily human mobility motifs. Journal of the Royal Society: Interface 2013. DOI: 10.1098/rsif.2013.0246.

StatsSA (2011). Poverty Trends in South Africa: An examination of absolute poverty between 2006 and 2011. Available online: <http://beta2.statssa.gov.za/publications/Report-03-10-06/Report-03-10-06March2014.pdf> [Accessed: September, 2015].

StatsSA (2013). National Household Travel Survey. Available online: <http://www.statssa.gov.za/publications/P0320/P03202013.pdf> [Accessed: September, 2015].

World Wide Worx. 2015. Executive Summary of the South African Social Media Landscape 2015.

Zagheni, E., Garimella, V.R.K., Weber, I., State, B.: Inferring international and internal migration patterns from twitter data. In: WWW (Companion Volume), pp. 439–444 (2014).

Zheng, Y-T., Zha, Z-J., & Chua, T-S. 2011. Mining travel patterns from geo-tagged photos. ACM Transactions on Intelligent Systems and Technology. Vol 3(3): pp 1—18.