Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 48 (2012) 775 – 785

Procedia
Social and Behavioral Sciences

Transport Research Arena– Europe 2012

Use of social media for transport data collection

Dimitrios Efthymiou[a*], Constantinos Antoniou[b]

[a]PhD Candidate, National Technical University of Athens, Iroon Polytechniou 9, Athens 15780, Greece
[b]Assistant Professor, National Technical University of Athens, Iroon Polytechniou 9, Athens 15780, Greece

Abstract

The multi-characteristic synthesis of internet and social network users (different nationality, age, education level, interests) renders these platforms powerful tools, suitable for many purposes. Until now, businesses use them for marketing, political candidates for their election campaigns, information networks for news updates, companies for recruitment and, most recently, nations for revolutions. In this paper, the use of social networks for conducting transport surveys is presented. The integration with e-mail providers broadens their use and makes them more suitable for data collection. In addition, statistics regarding discussions (tweets) with words related to the survey's subject were extracted from Twitter and evaluated. Since the applications of social and other internet networks are always developed, their use for internet surveys should be further examined in the future.

Keywords: Social media, carsharing, bikesharing, electric vehicles, questionnaire

[*] Corresponding author. Tel.: +30-210-772-1675; fax: +30-210-772-2629.
E-mail address: defthym@mail.ntua.gr

doi:10.1016/j.sbspro.2012.06.1055

1. Introduction

The use of social media increases significantly day by day. According to statistics published by facebook (http://www.facebook.com/press/info.php?statistics), on January 2011 the most popular social network had more than 500 million active users, 50% of whom log on every day. On average, each member has 130 friends and in total, they spend more than 500 billion minutes every month logged on. Facebook is available on more than 70 languages and 70% of its users are outside the USA. A recent study in Greece shows that 36% of those that connect to the internet use social networks, especially those up to 34 years old (Observatory for Digital Greece, 2011).

The variety of their members' characteristics (people of different nations, age, level of education, with different interests) renders them powerful tools, suitable for many different applications. In the past, social media have been used for marketing and political campaigns, for news updates by information networks or even from people, for professional advertisement and recruitment and, most recently, for organization of protests revolutions.

Social Media are Internet-based applications that allow users to "post", "tweet" and so, allowing them to externalize their personal experiences and thoughts and share them with others. Social Media is the relatively new, technological way of social networking. New social media are always being developed but some of them never become popular. There are platforms that are addressed to users with either common interests, or qualifications, or to the general public. For example, academia.org is addressed to the academic community; LinkedIn.com is addressed to professionals and is focused to recruiting and job-hunting, while twitter.com and facebook.com are addressed to everyone. The recently launched Google's social platform (Google+) seems to integrate more than one characteristic of its predecessors. Social media users are of different nationality, age, education level, employment status and they have different interests. This multi-variation of its members composes a rich sample, which renders these platforms powerful tools, suitable for collection of transport data, either by surveys or other ways. Social media have already been used for transportation-related applications (e.g. Amey et al., 2011, Bregman, 2011, Grigolon et al., 2011, Carvalho et al., 2010).

The motivation for tis research came through the need for various data collection applications to support ongoing research activities. The authors are interested in the people's thoughts about carsharing, bikesharing, electric vehicles, but also about the travel patterns of young Greeks and the potential of joining these schemes. As many of these initiatives are not under the umbrella of a funded research program, the method needed to be made with the minimum cost but to achieve results of high quality. The use of social media was evaluated from an early stage of the research.

1.1. Background and motivation

Sample survey is for more than 75 years a useful tool to learn people's behaviours and opinions. The means of surveying, the interaction, time involvement and trust of human at them change continuously. Many limitations from the pen and pencil surveys, to the mail, telephone, internet and mixed mode surveys have been overcome in recent years. Dillman (2009) introduced the tailored design of the survey, a theory that was based on the reduction of the four error sources: coverage, nonresponse, measurement, sampling. He suggests the integration of survey procedures to encourage people to respond.

According to Dillman (2009) internet surveys are advantageous when compared with telephone surveys because of the lower non-response rate. However, they are vulnerable to sampling error. They address a specific number of people, those that use internet, who usually are college students or certain professionals. A survey in USA by Horrigan & Smith, (2007) has shown that 29% of people do not have internet access, which leaves them out of the survey. Only 2% of those with internet access use it only at work, which means that few of them are restricted in the way that they can use it. In addition, another significant restriction is that less educated, people with lower income, people more than 65 years old and non-whites have less access to the internet. Dillman also suggests that sampling error, nonresponse error, coverage error and measurement error should all be considered in order to achieve better results from a survey.

The first e-mail surveys were similar to paper surveys. They were text-based, either attached at the e-mail or directly written at it, and limited in length. Their advantage was that they had a faster delivery and response time and decreased cost. The development and the expansion of the Web in the early 1990s, offered new interactive features for internet based surveys which could now be enhanced with images, audio and other compelling features. Schonlau (2002) states that response rates of web surveys range from 7 to 44 per cent, while the respective response rates of e-mail surveys fall between 6 and 68 per cent. In addition Schonlau is making an extensive analysis of the factors that should be considered during the design and the implementation of an internet survey, the factors that were considered in that research.

The structure of this paper is as follows. The survey design and dissemination via various media is described in Section 2, followed by the analysis of the survey response data in Section 3. A preliminary investigation of the potential of mining the social network twitter for transport data collection applications is presented in Section 4. Section 5 includes a discussion and concluding remarks.

2. Survey design and dissemination

For the first part of the presented research, a questionnaire about carsharing and bikesharing needed to be structured electronically. Its design was based on two aspects. The first aspect was to create a questionnaire that will be addressed to everyone at an international level, and for that reason it was internet-based and composed in the English language. The second aspect was to achieve more accurate (less biased) results with the minimum cost and time. Publicly available and free research tools were selected. As every internet-based survey, this is vulnerable to biases, but being concerned about that an effort to eliminate them using various methods was made. Different survey software were considered before the design of the questionnaire: The commercial SurveyMonkey (www.suveymonkey.com), the open source LimeSurvey (www.limesurvey.org) and the Google Forms capability within Google Documents (google docs) made it to the short list. Despite the number of features that SurveyMonkey offers, it was not used because of the high cost of its use. LimeSurvey and Google docs offer integrated platforms for design and analysis of the results. Furthermore, Google docs offers more than 7GB of storage space for free. The only disadvantage of the Google offering is the restrictions in the number of the question types; however, most of these restrictions can be overcome through creative use of the available types. The questionnaire was finally designed using the tools offered by Google, because it offers an integrated, simple, flexible and manageable environment.

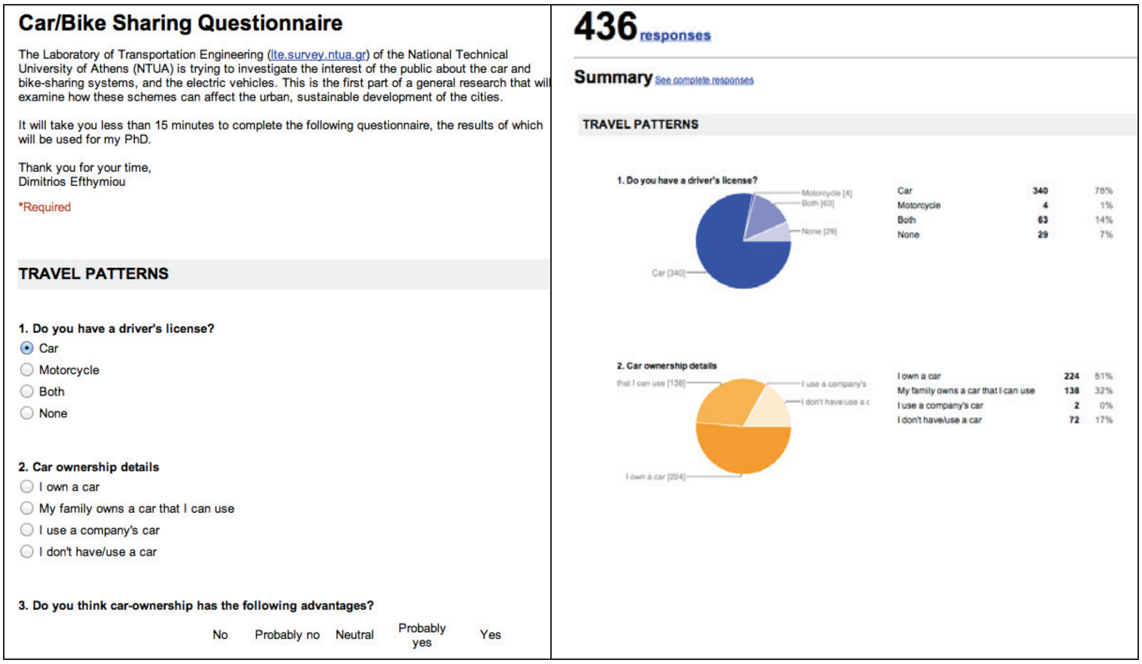
Before making the questionnaire's link publicly available, a few thinks were considered:

- The number of recipients should be known, because this would support the calculation (or approximation) of the response rate.
- The sample must be not only from Greece, but also from other countries, to the degree that it is possible.

- The sample must be random, theoretically it should cover all the ranges of age, income, education level. Of course this is the most difficult requirement to achieve. For example, large professional and scientific lists were used (such as TMIP, http://tmip.fhwa.dot.gov/ and UTSG, www.jiscmail.ac.uk), but the membership of these lists is not necessarily representative of the entire population.

The questionnaire begins with a descriptive paragraph on how the results will be used to benefit the users and the public in general, so as to encourage people to reply (Groves et. al, 1992). Then there were questions about the travel patterns, such as the type of driving licenses owned, car ownership, bicycle ownership, main mode they use to go to different destinations (work, school, grocery shopping, general shopping, etc.), trips per week etc. Questions of special interest about the survey subject followed (the subject was carsharing and bikesharing, but will not be analysed further in this paper). The last part included questions about the demographic characteristics of the respondent.

Figure 1a presents a few indicative questions and the full questionnaire is available online at the following link: https://docs.google.com/spreadsheet/viewform?formkey=dEEwLVRoSElXMWhqSjlVT2FQUVFabnc6MQ. Figure 1b shows the way that the summary response results are presented by the system.



Car/Bike Sharing Questionnaire

The Laboratory of Transportation Engineering (lte.survey.ntua.gr) of the National Technical University of Athens (NTUA) is trying to investigate the interest of the public about the car and bike-sharing systems, and the electric vehicles. This is the first part of a general research that will examine how these schemes can affect the urban, sustainable development of the cities.

It will take you less than 15 minutes to complete the following questionnaire, the results of which will be used for my PhD.

Thank you for your time,
Dimitrios Efthymiou

*Required

TRAVEL PATTERNS

1. Do you have a driver's license?
● Car
○ Motorcycle
○ Both
○ None

2. Car ownership details
○ I own a car
○ My family owns a car that I can use
○ I use a company's car
○ I don't have/use a car

3. Do you think car-ownership has the following advantages?

No	Probably no	Neutral	Probably yes	Yes

436 responses

Summary See complete responses

TRAVEL PATTERNS

1. Do you have a driver's license?

Car	340	78%
Motorcycle	4	1%
Both	63	14%
None	29	7%

2. Car ownership details

I own a car	224	51%
My family owns a car that I can use	138	32%
I use a company's car	2	0%
I don't have/use a car	72	17%

Fig. 1 (a) Part of the questionnaire, (b) summary of responses (application available by Google Docs)

The stages of dissemination via the social media are described in Table 1. An "event" was created on facebook and the survey was hosted there for two weeks. With this "event", a group of candidate respondents is created, to whom the creator can send mass messages. This helped to keep the interest in high levels. The strategy for reminders was such that a reminder message was sent every (about) 100 completed questionnaires, so as to remind it to the others and achieve a higher response rate. The results showed that a high reply level was achieved in a few days. It is noted that the use of twitter and Linkedin were also examined. However, they were excluded from the analysis as the former requires to have many followers and the latter does not offer a similar tool for dissemination.

In addition, the questionnaire was disseminated via e-mail to mailing lists and personal contacts. Two major lists were used in our case, University Transport Study Group (UTSG, www.utsg.net/) and Travel Model Improvement Program (TMIP, tmip.fhwa.dot.gov), to which members are transport professionals, as well as three student-lists of Imperial College and NTUA.

Table 1. Questionnaire dissemination strategy

	Way of dissemination	Number of sample per group	Day of research	Time of day
Imperial MSc 1	Mailing list	50	1	17:53
Group of Personal Contacts 1	e-mail	125	2	15:15
Group of Personal Contacts 2	e-mail	16	2	23:01
Group of Personal Contacts 3	e-mail	22	3	15:05
Facebook + Gmail	Social net. + e-mail	800 (700 common)	5	12:00
Reminder 1 (100)	Social network (Facebook)		5	
NTUA PhD Candidates	Mailing list	200	6	15:27
Reminder 2 (200)	Social network (Facebook)		7	
Imperial MSc 2	Mailing list	50	7	10:38
Reminder 3 (300)	Social network (Facebook)		13	
UTSG	Mailing list	1100	19	11:04
TMIP	Mailing list		19	20:01
Other	e-mail	67		

3. Analysis of survey responses

Figure 2 shows the number of responses over time, along with a density of responses for each of the main “streams” of dissemination. Table 2 presents an overview of the response rates by medium. People were able to answer many days after they receive the questionnaire. However, to be able to compare the response rates achieved by the different ways of dissemination, it is considered as the total number of responses three days after the dissemination. In the beginning, the questionnaire was sent to a list of 50 students and to 163 contacts. 38 answers were received, or 23% response rate. Five days later, it was disseminated to contacts both via the facebook event and e-mail, at the same people. Three days later, 28% of them had responded. The histogram shows that responses were received for many days after the dissemination, which increases the real response rate. Finally, 15 days later, it was sent to more than 1000 people via the Transport mailing lists, with a response rate of 1.7%. This way of dissemination is less effective, as the recipients have been accustomed to similar e-mails, and it was possibly ignored from the majority. In addition, their responses do not represent the responses of a random sample, as they are professionals in the field of Transport. However, if the objective of a survey is to focus on professionals in a specific field (rather than a random sample of the general population), such techniques can represent a very effective tool. An observation that was made based on the responses is that people not familiar with the topic exhibited an enthusiasm in responding that may not be shared with practitioners and researchers familiar with the topic.

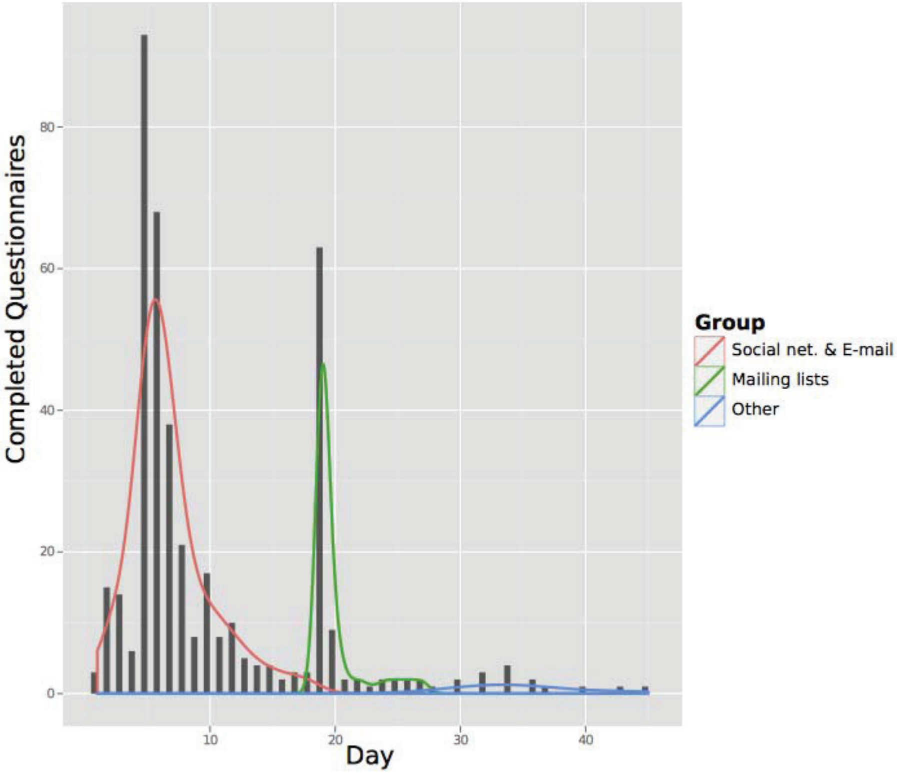


Fig. 2 Responses per day

Table 2. Response rates

Days		Response rate
1-3	Mail	23%
5-7	Facebook +Gmail	28%
22-24	Transport mailing lists	<1.7%

The results show a relevant dispersion concerning the countries of origin of the respondents. More responses are from Greece and there is a significant number from UK and USA (which is not surprising, considering that the mailing lists considered and the social media have a strong participation from these countries). For the purposes of the further research (the first indication of which can be found in Efthymiou et al., 2012), the responses were separated in two subgroups: Greeks and Others. As the group of the Others consisted of more than 25 countries, it cannot give a clear statement of people's behaviour because of the different local characteristics of each country (e.g. different income classes). However, they were used carefully for comparisons of the two groups in a few cases.

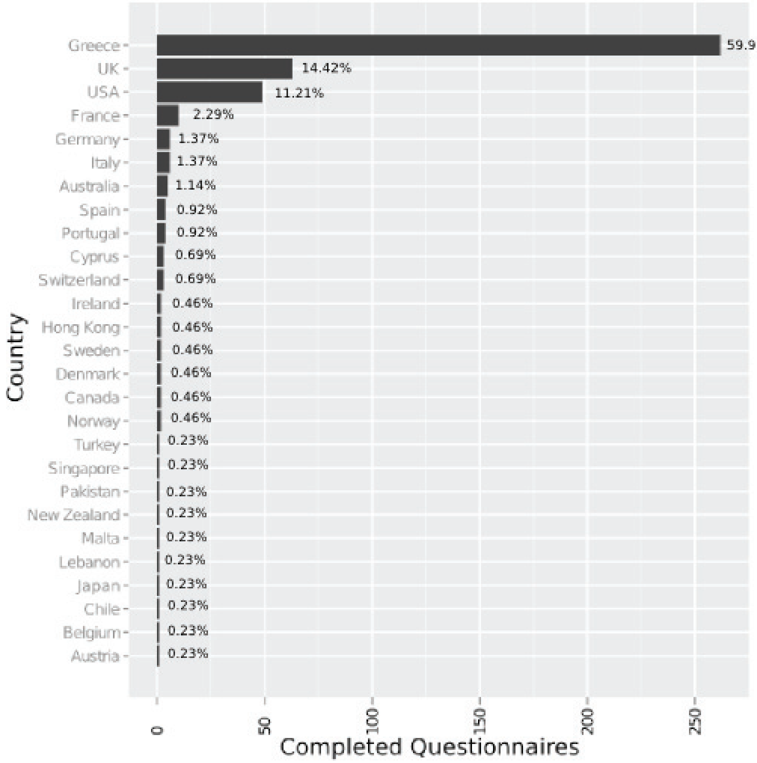


Fig. 3 Completed questionnaires by country

As expected, the demographic results (Figure 4) show that there are biases. The majority (88.9%) of the Greek sample is between 18 and 35 years old, which is correlated with the high percentage of singles (88.5%) while only 4.5% is more than 45 years old. This is because people of younger ages are more usual to have access to the Internet and be members of social media. A recent research in Greece (www.observatory.gr), found that only 40.7% of people 45 to 54 years old and 15.5% of those more than 55 have access to the internet, while 80.95% of those that have access belongs at the age class 18-35. Even though the percentage of those who have access has been doubled over the last four years (www.observatory.gr), it is still too low and approximately 1/3 below the average of the European Union. The percentage of those who use social media is higher at the ages between 18-34. Concerning the gender, an almost balanced 50/50 response sample was obtained (in Greece the distribution is 49.5% male and 50.5% female). Results for the income biases cannot be extracted, because a high percentage of households denotes (even in official statements) less money than they earn. Despite the biases, this methodology offers a practical way of estimating and analysing young people perception (about carsharing and bikesharing in that case). Different models were estimated for the age subgroup of 18-35 years old, where the sample is more compact (Efthymiou et al., 2012) as there has never been a survey of this subject in Greece before.

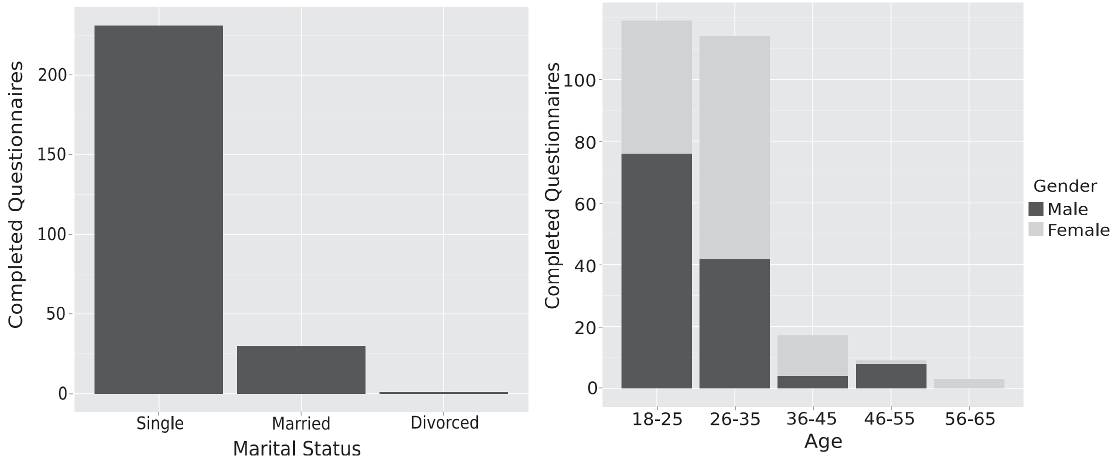


Fig. 4 (a) Marital Status, (b) Age and gender distribution in the responses

4. A preliminary Twitter application

Twitter is the second most popular social network, with similarities and differences from facebook. There are many cases in the past, in which news have been announced in twitter before being published by major news channels, while the simultaneous contribution of thousands of users limits the possibility of misinformation (www.observatory.gr). In twitter, users can "tweet" messages of maximum 140 letters and share them with their followers. According to the statistics published by its operators, 1 billion "tweets" are posted every week, while the average number of accounts created per day is about 460,000. These numbers render this platform a powerful tool where millions of people "socialize". Its role for this purpose was examined.

The wealth of data hidden in this information can offer valuable information for data collection regarding attitudes and perceptions of the general public, as well as information on emerging incidents as they happen. Different twitter applications have been developed until now, some of which create a connection between this and the statistical software R (R Development Core Team, 2011). The integration of these platforms offers a powerful tool for statistical analysis. The thoughts or actions of twitter's users that are publicly available via the "tweets" can be scrapped, parsed and analyzed in R using the appropriate packages such as XML (Lang, 2011) or twitteR (Gentry, 2011). The integration and interaction between twitter and R were examined in this research. A script that can retrieve information about the number of the tweets containing the words carsharing (or car-sharing), bikesharing (or bike-sharing) and electric vehicles, and the geographical location of the users (in case the application was enabled by the users), was coded in R. The script then reads the time format from the html page, translates it in R and prints a graph using the ggplot2 (Wickham, 2009) package of R. The script also reads and stores the location of the users (in case that it is provided by them), which can then be plotted by the googleVis package (Gesmann and De Castillo, 2011).

Figures 5 and 6 demonstrate examples of temporal and spatial analysis of the twitter data mining results. Naturally, the ability to obtain location information on the participants raises privacy questions (Cottrill and Thakuriah, 2011).

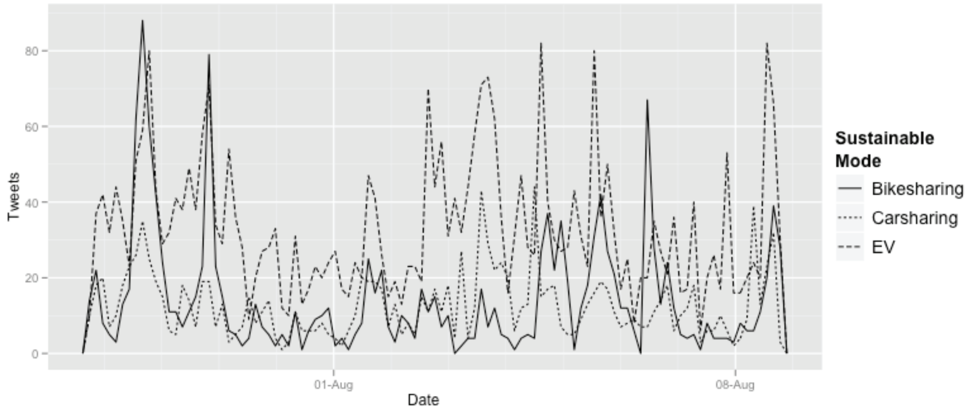


Fig. 5 Demonstration of temporal analysis of twitter mining results

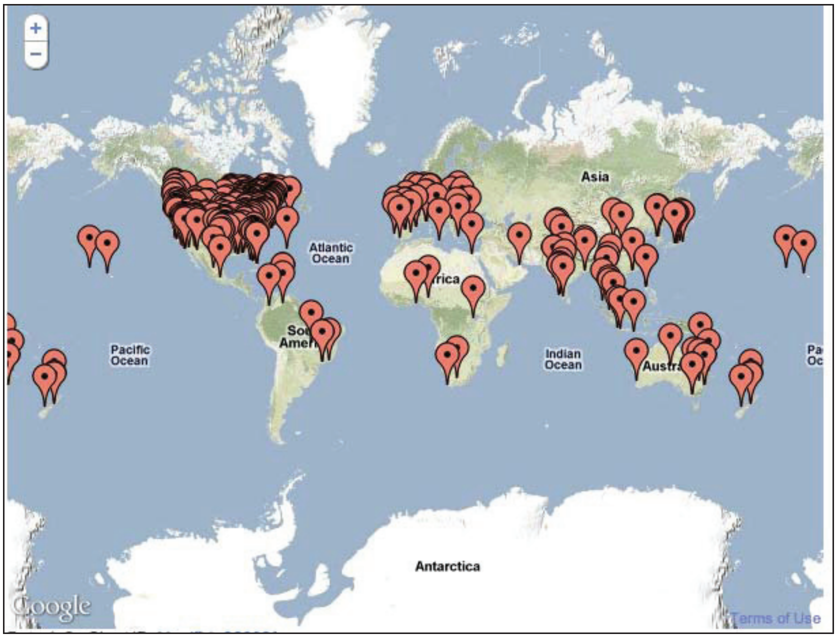


Fig. 6 Demonstration of spatial analysis of twitter mining results (tweets containing "EV")

5. Discussion and conclusions

The use of social media becomes more and more popular year by year. The result of the continuous sharing of information by their users is the creation of a huge database of social content that is publicly available in some extend. This amount of information, but also the generation of new for the purpose of specific research, could be used by transport researchers to support their studies. Questionnaires can be structured for free Survey-groups can be easily created in these platforms, but attention should be paid to the range of demographic classes that they cover (social media are mainly used by ages between 18-35 at this time). Software for statistic analysis such as R (R Development Core Team, 2011) offer packages that enable the creation of applications for web data collection and their illustration (e.g. in Google maps using the geographic location) if be integrated with social media. The possibilities offered by the wealth of information in social networks are endless. On the other hand, there are many practical limitations and concerns that will need to be further examined, before these data become more credible.

Acknowledgements

This research has been partly funded through the SustainCity project, co-financed by the European Commission within the FP7 program.

References

Amey, A., J. Attanucci and R. Mishalani (2011). "Real-Time" Ridesharing – The Opportunities and Challenges of Utilizing Mobile Phone Technology to Improve Rideshare Services. Proceedings of the 90th Annual Meeting of the Transportation Research Board, January, Washington, D.C.

Bregman, S. (2011). What's the worst that can happen? How to stop worrying and love social media. Proceedings of the 90th Annual Meeting of the Transportation Research Board, January, Washington, D.C.

Carvalho, S., L. Sarmento and R. Rossetti (2010). Real-time sensing of traffic information in twitter messages. Proceedings of the 4th Workshop on Artificial Transportation Systems and Simulation ATSS at IEEE ITSC 2010, Madeira, Portugal, 19-22 September, 2010

Cotttrill, C. and P. Thakuriah (2011). Protecting Location Privacy: A Policy Evaluation. Proceedings of the 91st Annual Meeting of the Transportation Research Board, January 2011, Washington, D.C.

Dillam D., Smyth J. D. & Christian M. (2009). *Internet, Mail and Mixed-Mode Surveys: The Tailired De-sign Method*. 3rd edition, Wiley

Efthymiou, D., C. Antoniou and P. Waddell (2012). Which factors affect the willingness to join vehicle sharing systems? Proceedings of the 91st Annual Meeting of the Transportation Research Board, January 22-26, Washington, D.C.

Gesmann M. & De Castillo D. (2011). Interface between R and the Google Visualisation API. GoogleVis package for R. Available on-line in: http://cran.r-project.org/web/packages/googleVis/googleVis.pdf

Gentry J. (2011). Package 'twitteR'. Available on line in http://cran.r-project.org/web/ packages/twitteR/ twitteR.pdf

Grigolon, A.B., A. D. A. M. Kemperman and H.J.P. Timmermans (2011). Using Web2.0 Social Network Technology for Sampling Framework Identification and Respondent Recruitment: Experiences with a Small-Scale Experiment. Proceedings of the 90th Annual Meeting of the Transportation Research Board, January, Washington, D.C.

Groves R. M., Cialdini R. & Cooper M., P. (1992). Understanding the decision to participate in a survey. Public Opinion Quartely (56, pp. 475-495)

Horringam J. & Smith A. (2007). Home Broadband Adoption. Pew Internet and American Life Project. URL: http://www.pewinternet.org/

Lang, D. T. (2011). Package 'XML'. Tools for parsing and generating XML within R and S-Plus. Available on-line in: http://cran.r-project.org/web/packages/XML/XML.pdf

Observatory for Digital Greece (2012). The use of Internet by the Greeks. Available online: www.observa tory.gr/meletes (in Greek)

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Schonlau M., Fricker R. & Elliot M. N. (2002). *Conducting Research Surveys via E-mail and the Web*. ISBN 0-8330-3110-4. Published by Rand

Wickham H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. Use R! Series. Springer, ISBN-10: 0387981403