LIFE AND DEATH OF VOLUNTEERED GEOGRAPHIC INFORMATION CONTRIBUTORS IN A LARGE ONLINE COMMUNITY—THE CASE OF OPENSTREETMAP

By © Daniel Bégin

A Thesis submitted to the School of Graduate Studies

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Geography

Memorial University of Newfoundland

May 2018

St. John's, Newfoundland and Labrador

ABSTRACT

The advent of the Web 2.0 has democratized both the production and dissemination of knowledge by enabling communities of online contributors to generate content collectively. This thesis focuses on "Volunteered Geographic Information" (VGI), a type of user-generated content (UGC) oriented toward geographic information. The provided content is known to be highly heterogeneous in coverage, nature and quality, reflecting a patchwork of motivations, interests, knowledge and skills of individual contributors. Characterizing VGI data requires understanding contributors' behaviour. Typologies of contributors are proposed in an attempt to link VGI contributors with the nature of the data they provide. Those typologies are directly or indirectly related to the time spent by the contributors in a project, but they do not use a formal temporal perspective to understand their behaviour. We considered the time spent by contributors in a given VGI project as an essential component for understanding their contribution patterns (e.g. volume, content, quality).

In order to fill this knowledge gap regarding how the time in the project may have impacted contributors' behaviors, I analyzed the behaviour of the OpenStreetMap (OSM) contributors, of a large VGI community. I identified different events that affected enrollments and withdrawals over a project's history using time series analyses. I established the phases of contributors' life cycle using survival analyses and linked their contributions to the different phases.

Six distinct phases were identified in the life cycle of OSM contributors. Analyses

revealed that these phases were grouped into three major stages: An “Assessment” stage that last a few months, followed by an “Engagement” stage that can extend over more than a decade, to eventually move to a “Detachment” stage over which the contributors leave the project. Analysis of contributions at each phase revealed that contributors’ behaviour is dominated by two distinct processes. When contributors enroll in a project, they seem to be driven by a learning-adaptation-dominated process before switching to a cumulative-damage-dominated process followed by a withdrawal from the project. In parallel, I found that the diffusion of innovation theory (DoIT) had an important impact all along the project’s history. This research not only shed light on online contributions but also reveals different aspects of human behaviours.

ACKNOWLEDGMENTS

This thesis began as a retirement project that quickly turned out to be a personal, intellectual and academic challenge that lasted for more than five years.

It is therefore important to start by expressing gratitude those who supported me and who have coped daily with my doubts and discouragements, shared my successes from time to time, and never doubted about my capacity to go through that project. First, I want to thank my wife Danielle who supported me throughout this long journey. She who accepted to discover St. John's, NL, and its vicinity, the time for me to get through a year of academic requirements. She, who also took time to manage day-to-day chores while I was debugging obscure recursive SQL queries or debunking science with Latour[1]. I would like to thank my brother Christian who has put all his science to help me move forward on the graduate studies path and to understand its requirements; to my father Claude who inquired each time we met about the progress of my researches; and my friends Mario, Martin, Jean-Marc, Bernard and Louise for their support.

But this project could not have been achieved without the interest and involvement of my co-supervisors. Dr. Rodolphe Devillers who has kindly agreed to take charge of my supervision all this time. For a first year in St. John's, and then from two thousand kilometres away, while I was working on my research in Sherbrooke. Thank you for your countless hours of reading, trying to make sense of my words, in (too long) manuscripts, usually written in approximate English. Thank for the countless corrections you made to

[1] https://en.wikipedia.org/wiki/Bruno_Latour

these manuscripts without which I would not be currently writing this text. Dr. Stéphane Roche who wished to supervise me despite his schedule of 36 hours a day! The few times we met, enlightened and stimulated me. Thanks for having taken the time to read all the material I sent you over all these years.

I would like to thank Dr. Alvin Simms who very kindly helped me over first my first analyses, and Emilie Novaczek who gracefully accepted to proofread latest manuscripts. Finally, I would like to thank Dr. Jeffrey Parson, Dr. David Coleman and Dr. Robert Feick who examined the thesis thoroughly.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

API	Application Programming Interface
BBC	British Broadcasting Corporation
BCS	British Cartographic Society
CT/ODbL	Contributor Terms/Open database Licence
FME	Feature Manipulation Engine
FOSS	Free Open-Source Software
GMT	Greenwich Mean Time
JOSM	Java OpenStreetMap editor
LWG	OSM legal working group
NMA	National Mapping Agency
ODbL	Open database Licence
OSC	Open Source Conference
OSM	OpenStreetMap project and community
OSMF	OpenStreetMap Foundation
OSS	Open-Source Software
SEO	Search Engine Optimization procedures
SOTM	State Of The Map conferences
UGC	User Generated Content
UTC	Coordinated Universal Time
VGI	Volunteered Geographic Information

List of Appendices

Co-authorship Statement

The candidate formulated the research questions, designed and conducted the analyses, and wrote a first draft of each of the chapters presented herein.

Drs. Rodolphe Devillers and Stéphane Roche supported the candidate throughout the research process, providing critical feedback on the research design and helping with the preparation of the thesis, specifically with the writing of Chapters 2, 3 and 4 that were published or submitted to peer-reviewed journals.

Chapter 2: Manuscript submitted on 5 Jun 2017, Accepted 1 Aug 2017, Published online: 11 Sep 2017

Bégin, D., Devillers, R. and Roche, S., 2017, Contributors' Enrollment in Collaborative Online Communities: The Case of OpenStreetMap. *Geo-spatial Information Science*, 19 (3), 282-295.

Chapter 3: Manuscript submitted on 1 Sep 2017, Accepted 2 Nov 2017, Published online: 4 Nov 2017

Bégin, D., Devillers, R. and Roche, S., 2017, Contributors' Withdrawal from Online Collaborative Communities, the Case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 6 (11), 340.1-340.20.

Chapter 4: Manuscript submitted on 29 Nov 2017, Accepted 24 Mar 2018, Published online: 16 Apr 2018

Bégin, D., Devillers, R. and Roche, S., 2018, Contributors' Life Cycle in Online Collaborative Communities, the Case of OpenStreetMap. *International Journal of Geographical Information Science.* pp. 1-20. Available from: https://www.tandfonline.com/doi/full/10.1080/13658816.2018.1458312.

Finally, the candidate revised the manuscripts in response to the different journals' reviewers.

Chapter 1: Introduction

Thousands of years ago, people willing to record and share the location of important places engraved information on various materials, creating the first known maps (Clark 2005, Utrilla *et al.* 2009). Over the centuries, with the rise of large empires and the ability to travel further, maps became a priceless source of information.

Despite considerable improvements in surveying and mapping practices, for millennia features of interest had to be walked or sailed to be mapped. The advent of aerial photography allowed national mapping agencies (NMAs) to repeatedly create detailed maps of entire countries, leading to highly standardized mapping processes and products. However, over the last decades, most NMAs activities have been challenged by an increasing difficulty to absorb their operating costs under growing budget constraints (Estes and Mooneyhan 1994, Goodchild 2007b).

The increasing affordability of location technologies like Global Positioning Systems (GPS), and their ubiquitous use in our daily life, has led to a surge in demand for up-to-date digital geographic information. This occurred so quickly that while NMAs struggled to adapt, large multinational corporations (e.g. Microsoft, Google) developed their own products. However, most of these products, public or private, were considered very restrictive in terms of use, cost or access (Goodchild 2007a, Coast 2007, Coast 2011).

In parallel, the development of the Web 2.0 (O'Reilly 2005) resulted in the democratization of both production and dissemination of knowledge by enabling online

communities to produce content. The literature has suggested multiple labels to describe this user-generated content (UGC), including collective intelligence (O'Reilly and Battelle 2009), crowdsourcing (Howe 2006), peer production (Benkler 2002), and "produsage" (Bruns 2006). The concept of user-led content production (Benkler 2002, Bruns 2006) is of a particular interest because it describes the collaborative and iterative work of a large number of users toward a shared goal defined by the community. It differs from crowdsourcing (Howe 2006, O'Reilly and Battelle 2009) in which contributors are not necessarily the main users of the content or involved in decisions about the nature of the content. According to Bruns (Bruns 2008), user-led content production communities have common characteristics that define the nature of both the contributors' behaviours and the content they generate:

Open participation: Anyone with an interest in a given content can contribute according to his/her knowledge and skill. Some communities may require registering before contributing, but it is usually not used as a barrier to contribute. The quality of the content is assessed and improved by the community as they browse and use it. Consequently the quality control is said to be probabilistic (Bruns 2008), depending on the volume of contributors and the frequency at which a given content is examined. This quality control strategy has been called the "Linus's Law" (named after Linus Torvalds, an open-source software developer), stating that "given enough eyeballs, all bugs are shallow" (Raymond 1999).

Fluid heterarchy: Everyone is considered as being able to provide valuable contributions, even though knowledge and skills can differ greatly between contributors. As a project evolves, problems are solved, discussions are held, some contributors will emerge from a community as *ad hoc* leaders, according to the merit the community assigns to their contributions at that time (Bruns 2008, Preece and Shneiderman 2009). Similarly, those whose contributions fail to meet implicit or explicit community standard will be subject to a *de facto* exclusion. Due to the open nature of these communities, enrollment and the withdrawal of participants may generate shifts in the group culture and contribution

assessment criteria over time.

Unfinished outcome: UGC is a perpetual work in progress, as both the contributors' needs and the community evolve over time. Content may result from the contributions of many participants, and will continue to be modified as the needs, the rules and the participants are changing. Furthermore, online collaborative communities usually enable participants to freely choose when and what content they will contribute. Communities that require simple tasks as the minimum contribution will be the more inclusive to potential participants of various skills and knowledge levels and, as a result, these communities may grow to be larger than counterparts with more complex tasks that raise boundaries to participation (Bruns 2008). The broader the number of components a given content has, the higher the odds is that the result may be incomplete if the needs of the contributor can be fulfilled with a partial result.

Common property: In order to work on a shared outcome in which contributors build on content provided by others, intellectual property rights must be adapted to ease content reuse. Consequently, different licensing frameworks have been developed to facilitate such usage in which individuals' intellectual rights yield to the community. These licences usually require an appropriate acknowledgment (attribution) when the content is made public by an external entity. Furthermore, licences often require that external entities who build upon the work of the community make these improvements available through share-alike clauses.

This field of research, when related to user-generated geographic data, is described as "Volunteered Geographic Information" (VGI) by Goodchild (2007a). This thesis focuses more specifically on user-led content VGI communities in which people can share spatial information without the constraints of authoritative external organizations (Goodchild 2007a, Goodchild 2007b). We view this as a return to the origins of the mapping made possible by modern technologies.

1.1. Conceptual Framework

Budhathoki et al. (2010) have proposed a contributor-centric conceptual framework to

understand the dynamic of these user-led content VGI communities. The framework is articulated around three components: "motivation"—"action & interaction"—"outcome." A modified version of this conceptual framework is presented in Figure 1-1.

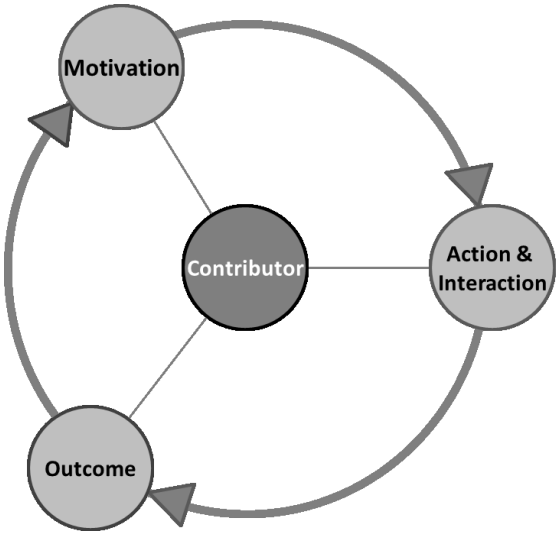


Figure 1-1 Modified view of Budhathoki's contributor-centric conceptual framework.

In this model, a contributor may find a variety of reasons to contribute to a project ("motivation"). A contribution ("action & interaction") is usually intended to improve a project (e.g. product, infrastructure, rules and norms). The effect the contribution has on a project ("outcome") is then assessed. The evaluation of whether or not the outcome meets the contributor's needs, desires or aspirations may impact their motivation. This potentially transformed motivation may affect the contributor's decision to continue or to stop contributing to the project.

The **"Motivation"** component is mostly based on Self-determination Theory (Ryan and Deci 2000). This theory suggests that motivational factors behind an action are twofold: intrinsic and extrinsic. Intrinsic motivation is what drives people to fulfill their inner

potential and interests; it relates to the joy of performing an action. Extrinsic motivation is mostly driven by rewards or other compensations resulting from that action. The literature has found that most of the time, participants get involved and keep contributing to a collaborative project because of project's objectives (Nov *et al.* 2011, Aknouche and Shoan 2013), the nature of the tasks (Houle 2005, Borst 2010, Hemetsberger and Pieters 2003) or simply because contributing is enjoyable (Budhathoki *et al.* 2010, Aknouche and Shoan 2013); all these factors are linked to intrinsic motivations (Budhathoki 2010, Budhathoki *et al.* 2010).

The **"Action & Interaction"** component is twofold and describes the different activities in which VGI contributors might be involved. "Action" mostly refers to the operations required to create and improve the product (Budhathoki 2010, Rehrl *et al.* 2013, Rehrl and Gröchenig 2016), but it also includes maintaining the project's infrastructure, developing new applications, and establishing norms and rules (Haklay and Weber 2008, Budhathoki *et al.* 2010). "Interaction" refers to the iterative process by which contributors collaborate to improve the product (Mooney and Corcoran 2013, Mooney and Corcoran 2012), maintain the project, and keep the community healthy. However, actions and interactions performed in a project are not distributed evenly between contributors. This inequality is an important feature of participation in online communities. Nielsen (Nielsen 2006) has proposed a rule of thumb to describe this behaviour. The "90-9-1 rule," states that 90% of participants do not contribute (or contribute little), 9% contribute occasionally, and the remaining 1% contributes seriously, providing most of the content.

The **"Outcome"** is the result of collaborative efforts from all participants and it is constantly evolving as contributions are added to the project. Contribution decisions (i.e. whether to provide particular information for a given location and at a given time) are driven by participants' evaluation of likely outcomes, individual motivations, knowledge and skill (Heckhausen and Heckhausen 2008, Heckhausen *et al.* 2010). Consequently, VGI products tend to be highly heterogeneous (Ma *et al.* 2015), reflecting the diversity of their contributors.

The literature has described VGI as a global patchwork of geographical data (Goodchild 2007b), or even as a collection of "cupcakes" when compared to the "layer cakes" produced by NMAs (Roche 2012). When VGI was introduced, the early literature questioned the validity of VGI because of contributors' credibility and motivations (Flanagin and Metzger 2008, Coleman *et al.* 2009, Coleman 2010). Some of these concerns became secondary with the publication of data quality assessment studies on VGI (Haklay 2010, Zielstra and Zipf 2010, Mooney *et al.* 2010, Girres and Touya 2010). However, regardless of how VGI data is studied, the results (i.e. content and quality) are always a function of contributors' interests and motivations (Bégin *et al.* 2013).

In an attempt to link VGI contributors and the nature of their contributions, the literature has proposed different typologies based either on contributors' knowledge and skills (Coleman *et al.* 2009), the volume of their contributions (Neis and Zipf 2012) or the quality of these contributions (Arsanjani *et al.* 2013). Similar studies were made in other types of online communities, revealing that the nature of contributions seems to evolve

over contributors' lifetime (Bryant *et al.* 2005, Preece and Shneiderman 2009). Although all proposed typologies were directly or indirectly related to the time spent by the contributors in a project, none formally used this temporal perspective to understand behaviour. For instance, Bryant et al. (Bryant *et al.* 2005) proposed a binary typology (novice-expert) without specifying the time span of each phase. Similarly, Preece and Shneiderman (2009) proposed a more complex typology but again without determining any time scale. The other typologies refer to time more implicitly, using time correlated metrics such as the volume of contributions or changes in the quality of the data linked to increasing knowledge and skills over time.

We consider the time spent by contributors in a given VGI project as being an essential component for understanding contribution patterns (e.g. volume, content, quality). However, knowledge about this essential component is still lacking both the VGI and the other online communities. It is important for these communities and their managers to identify when and which retention techniques should be used to have the most impact when trying to retain contributors. Identifying when these techniques may have the greatest impact has not been formally studied in the literature. Similarly, some of the many events that mark the history of a project can have a significant impact on the life cycle of its contributors. Identifying the nature of the events that have a positive or negative impact on the lifetime of the contributors is also of paramount importance but is poorly discussed in the literature. The nature of the retention techniques to apply is however not in the scope of this study.

1.2. Research Scope and Objectives

In order to fill the knowledge gaps discussed above, I quantified key parameters of the Budhathoki's conceptual model (Figure 1-2) by linking them to measurable shifts in contributors' behaviour over time. Important measures include the frequency of contributions (i.e. the number of cycles performed in the model) as well as the time a participant spent in a project (i.e. the time during which these cycles were performed). In the same way, I needed to understand the context of participant enrollment, contribution, and withdrawal by taking into account the development of the project and important events in its history.

Parallels with demographic studies became evident. For instance, a first contribution to a project might be seen as the birth of a contributor and a withdrawal could be seen as the death within the project environment. Similarly, the events that dot the history of a project may have an effect on contributors' lifespan, as did epidemics, wars and technology in human history. This comparison led us to consider concepts such as birth and death rates, life expectancy and life cycle to deepen our knowledge about VGI contributors.

Temporal factors that could affect contributors' behaviour include lifespan and life cycle phase. This framework supports analysis of the evolution of contributors' behaviour over time, knowledge and skills increase and/or interests broaden. Contributors' life cycles lies between two self-determined events: the enrollment and withdrawal from the community. These decisions are not made in a vacuum, they integrate the information and

the perceptions the contributors have about the ability of a project to meet their needs and aspirations. Consequently, these decisions are closely related to the history of the project. This is of paramount importance because the history of VGI communities is very recent. The environment in which VGI contributions are made has undergone profound transformations, which might have affected contributors' life expectancy and the phases of their life cycle.

1.2.1 Research hypothesis

The time spent in a VGI project by its contributors can be described as a life cycle composed of different phases which should affect the nature (e.g. volume, frequency) of their contributions.

The purpose of this research is then to identify and characterize the life cycle phases of a population of VGI contributors from a temporal perspective, to identify events or environmental changes that affected contributors' lifespan or phases' duration, and to assess if the nature of contributions can be related to these phases. In order to do so, I needed to answer the following questions.

1.2.2 Research questions

1- What are the different phases of the life cycle of VGI contributors?

2- Is there a relationship between the different phases and the nature of provided contributions and if so, what are these relationships?

3- Are there any events, or factors that have changed contributors' lifespan throughout the development of VGI communities?

Contributors' lifespan lies between enrollment and withdrawal from a community. Consequently, I needed to understand how both evolved over time as proxy measures of project's capability to meet contributors' needs, desires or aspiration (i.e. the motivation component). Identifying the changes in contributory environment that affected project attractiveness over time determined my last two questions.

4- What events or factors affected enrollments in a community over years?

5- What events or factors affected withdrawals from a community over the same period?

1.3. Selection of a VGI community

The most successful VGI community has formed around the OpenStreetMap (OSM) project (Haklay and Weber 2008). Founded by Steve Coast in 2004 (OpenStreetMap contributors 2017), the project aims to create and distribute free geographic data around the world because "most maps you think of as free actually have legal or technical restrictions on their use, holding back people from using them in creative, productive, or unexpected ways" (OpenStreetMap contributors 2014). The project was built and is still maintained by its community. The project infrastructure and the applications used to edit it are deployed in an open-source software (OSS) environment. The entire project documentation is maintained in a wiki where standards and specifications are elaborated and discussed by the community. Editing interfaces allow contributors to map features of

interest. Features' identification and location are made possible via satellite imagery from the public domain and community-provided GPS tracks. Mapped features are described using tags (key = value) determined by the contributors. Finally, and maybe most importantly, each time an edit is provided by a contributor, the product is updated almost in real time and made freely available through the OSM web site (i.e., the outcome).

The OSM project has become an important source of geographic information all over the world. The quality of its data has often found to be comparable to, if not better than, available authoritative data sources (Dorn *et al.* 2015, See *et al.* 2013). This situation even led some NMAs to look at VGI as a source of data when updating their own products (Sabone 2009, Beaulieu *et al.* 2010, Bégin 2012). However, as a VGI project, the quality of OSM data relies on contributors' behaviour. The literature has found that data quality is linked to, among other things, the number of contributors (Haklay *et al.* 2010) interested in a given area (Napolitano and Mooney 2012, Neis and Zipf 2012), their interest about map features found in the area (Bégin *et al.* 2013), and the care they took when delineating and tagging these features (Girres and Touya 2010, Mooney *et al.* 2010).

Retrieving, manipulating and analyzing data from a large online community like OpenStreetMap fell in the realm of the "Big Data". Such context requires special data manipulation techniques at all stages and a suitable hardware. A high-end Dell desktop computer (8 CPU, 16 GB RAM) was used for all the processing. Tasks segmentation and aggregation was used for parallel processing. In this case, the resulting PostgreSQL database tables and indexes required more than 2 TB of disk space. Twenty-five million

contributions from about 450,000 participants were analyzed.

1.4. Organization of Dissertation

This dissertation contains three papers, in addition to the introduction and discussion chapters. The three papers are articulated according to the VGI conceptual framework and their relationship with that framework is illustrated below (Figure 1-2).

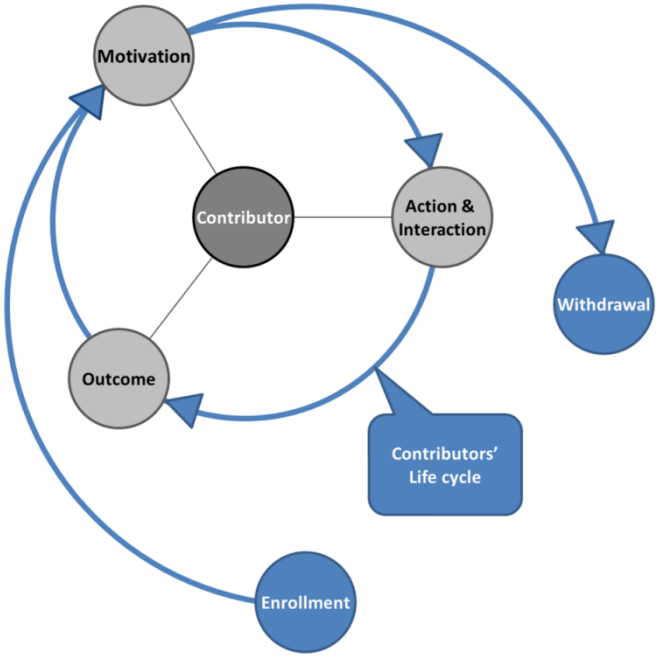


Figure 1-3 Conceptual framework and papers relationship (blue).

Chapter 2: By analyzing **Enrollment**, I aimed to understand the evolution of contributors' enrollment in OSM project and the events (e.g. application improvements, media reports) that may influence enrollment. We used time series analysis to characterize variations of both the daily rates of new registration (fertility rates) and first contributions (birth rates). Significant variations of these rates were compared with the project's history and the events that potentially affected people's motivation to enroll and

contribute were identified. Similarly, the time between registration and first contribution (length of gestation) was assessed to identify whether improvements made to the participatory environment have had an influence on its duration over time. We also assessed the proportion of registered members who never contribute (prenatal mortality rates) for the same reason. We expected these participants, referred to as "lurkers" in the literature (Nielsen 2006, Sun *et al.* 2014), to be very sensitive to changes to the OSM project and interface. This chapter was published in the following paper:

Bégin, D., Devillers, R. and Roche, S., 2017, Contributors' Enrollment in Collaborative Online Communities: The Case of OpenStreetMap. *Geo-spatial Information Science,* 19 (3), 282-295

Chapter 3: By analyzing **Withdrawal**, I aimed to understand the evolution of contributors' withdrawal and to identify events (e.g. changes to rules, internal conflicts) that may influence participants' decision to withdraw. However, the main challenge in assessing the number of withdrawals from the community was the distinction between participants who were waiting for the next opportunity to contribute from those who had permanently left the project. We developed a formal approach to statistically identify withdrawn contributors from the history of their contributions which incorporated contributors' circadian cycle to remove biases from the source data. Once withdrawn contributors were identified, survival analyses enabled us to characterize participants' average lifespan (life expectancy), the proportion who withdrew over time (death rates), and the probability they remained active (survival rate) over a given period of time. Finally, time series analysis was used to characterize variations in the daily rates of

withdrawals and compared them with project history to understand the nature of the events that potentially affected participant motivation and led them to withdraw from the project. This chapter was published in the following paper:

Bégin, D., Devillers, R. and Roche, S., 2017, Contributors' Withdrawal from Online Collaborative Communities, the Case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 6 (11), 340.1-340.20.

Chapter 4: By analyzing the **Contributors' life cycle** segment, I aimed to identify the phases in contributors' life cycle and to understand how the nature of their contributions may change according to these phases. We also tested whether the phases differ between participants according to both the volume of contributions they provided (i.e. Nielsen's 90-9-1 rule), and according to the epoch at which they registered to the project (i.e. Diffusion of Innovation Theory). The distribution of the time spent by the participants in the project enabled us to identify periods during which contributors seem to have homogeneous behaviours (i.e. phases in their life cycle). This chapter was published in the following paper:

Bégin, D., Devillers, R. and Roche, S., 2018, Contributors Life Cycle in Collaborative Online Communities, the Case of OpenStreetMap. *International Journal of Geographical Information Science.* pp.1-20. [Accessed 2018-04-16]. Available from: https://www.tandfonline.com/doi/full/10.1080/13658816.2018.1458312.

Finally, the complexity of operations associated with "big data" as part of this research has been described in Appendix A.

1.5. References

Aknouche, L. and Shoan, G., 2013, *Motivations for Open Source Project Entrance and Continued Participation.* Thesis (Master). Lund University.

Arsanjani, J.J., et al., 2013. Assessing the Quality of OpenStreetMap Contributors together with their Contributions. *The 15th AGILE International Conference on Geographic Information Science—Short Papers,* May 14-17 Leuven (BEL). Technische Universität Dresden, 1-4.

Beaulieu, A., Bégin, D. and Genest, D., 2010. Community Mapping and Government Mapping: Potential Collaboration? *Symposium of Commission I, ISPRS,* 16-18 June 2010 Calgary (CAN). 1-3.

Bégin, D., 2012. Towards Integrating VGI and National Mapping Agency Operations—A Canadian Case Study. *Role of Volunteer Geographic Information in Advancing Science: Quality and Credibility Workshop,* 18 September 2012 Columbus (USA). 1-2.

Bégin, D., Devillers, R. and Roche, S., 2013. Assessing Volunteered Geographic Information (VGI) Quality Based On Contributors' Mapping Behaviours. In: B. Wu, W.J. SHI and E. Gilbert, eds. *8th International Symposium on Spatial Data Quality,* May 30—June 1 Hong Kong (CHN). Hannover (GER): ISPRS, 149-154.

Benkler, Y., 2002, Coase's Penguin, or, Linux and" The Nature of the Firm". *Yale Law Journal,* 369-446.

Borst, W.A.M., 2010. *Understanding Crowdsourcing—Effects of motivation and rewards on participation and performance in voluntary online activities.* 1st ed. Rotterdam (NLD): Erasmus University of Rotterdam.

Bruns, A., 2006. Towards Produsage: Futures for user-led content production. In: F. Sudweeks, H. Hrachovec and C. Ess, eds. *Proceedings of Cultural Attitudes towards Communication and Technology,* June 28—July 01 Tartu (EST). 275-284.

Bruns, A., 2008. *Blogs, Wikipedia, Second life, and beyond: from production to produsage.* 1st ed. New York (USA): Peter Lang.

Bryant, S.L., Forte, A. and Bruckman, A., 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. *Proceedings of the 2005 international ACM SIGGROUP conference on supporting group work,* November 6-9 Sanibel Island (USA). New York (USA): ACM, 1-10.

Budhathoki, N.R., 2010, *Participants' motivations to contribute geographic information in an online community.* Thesis (PhD). Graduate College of the University of Illinois.

Budhathoki, N.R., Nedovic-Budic, Z. and Bruce, B., 2010, An interdisciplinary frame for understanding volunteered geographic information. *Geomatica,* 64 (1), 11-26.

Clark, J.O.E., 2005. *100 maps: The science, art and politics of cartography throughout history.* Sterling Publishing Company, Inc.

Coast, S., 2007. *The Pragmatic Mapper (part deux)* [online]. blog.openstreetmap.org. Available from: https://blog.openstreetmap.org/2007/03/27/the-pragmatic-mapper-part-deaux/ [Accessed 2017-05-21].

Coast, S., 2011. How OpenStreetMap Is Changing the World. In: K. Tanaka, P. Fröhlich and K. Kim, eds. *International Symposium on Web and Wireless Geographical Information Systems,* March 3-4 Kyoto (JPN). Berlin (GER): Springer-Verlag, 4-4.

Coleman, D.J., 2010. Volunteered geographic information in spatial data infrastructure: An early look at opportunities and constraints. *Proceedings of GSDI 12th conference,* 19-22 October Singapore (SGP). 1-18.

Coleman, D.J., Georgiadou, Y. and Labonté, J., 2009, Volunteered geographic information: The nature and motivation of producers. *International Journal of Spatial Data Infrastructures Research,* 4, 332-358.

Dorn, H., Törnros, T. and Zipf, A., 2015, Quality evaluation of VGI using authoritative data—A comparison with land use data in Southern Germany. *ISPRS International Journal of Geo-Information,* 4 (3), 1657-1671.

Estes, J.E. and Mooneyhan, D.W., 1994, Of maps and myths. *Photogrammetric Engineering and Remote Sensing,* 60 (5).

Flanagin, A.J. and Metzger, M.J., 2008, The credibility of volunteered geographic information. *GeoJournal,* 72 (3-4), 137-148.

Girres, J. and Touya, G.G., 2010, Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS,* 14 (4), 435-459.

Goodchild, M.F., 2007a, Citizens as sensors: the world of volunteered geography. *GeoJournal,* 69 (4), 211-221.

Goodchild, M.F., 2007b, Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0—Editorial. *International Journal of Spatial Data Infrastructures Research,* 2, 24-32.

Haklay, M., 2010, How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design,* 37 (4), 682-703.

Haklay, M., et al., 2010, How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal,* 47 (4), 315-322.

Haklay, M. and Weber, P., 2008, OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing,* 7 (4), 12-18.

Heckhausen, J. and Heckhausen, H., 2008. *Motivation and action.* 1st ed. New York (USA): Cambridge University Press.

Heckhausen, J., Wrosch, C. and Schulz, R., 2010, A motivational theory of life-span development. *Psychological review,* 117 (1), 32-60.

Hemetsberger, A. and Pieters, R., 2003. *When consumers produce on the internet: the relationship between cognitive-affective, socially-based, and behavioral involvement of prosumers* [online]. CiteSeerX. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.9299&rep=rep1&type=pdf.

Houle, B.B.J., 2005, A Functional Approach to Volunteerism: Do Volunteer Motives Predict Task Preference? *Basic and applied social psychology,* 27 (4), 337-344.

Howe, J. 2006, *The rise of crowdsourcing*, June 14th edn, Condé Nast Publications, New York (USA).

Ma, D., Sandberg, M. and Jiang, B., 2015, Characterizing the Heterogeneity of the OpenStreetMap Data and Community. *ISPRS International Journal of Geo-Information,* 4, 535-550.

Mooney, P. and Corcoran, P., 2012. How social is OpenStreetMap? In: J. Gensel, D. Josselin and D. Vandenbroucke, eds. *The 15th AGILE International Conference on Geographic Information Science,* April 24-27 Avignon (FRA). Springer, 1-6.

Mooney, P. and Corcoran, P., 2013, Analysis of Interaction and Co-editing Patterns amongst OpenStreetMap Contributors. *Transactions in GIS,* 18 (5), 633-659.

Mooney, P., Corcoran, P. and Winstanley, A.C., 2010. A study of data representation of natural features in OpenStreetMap. *Proceedings of GIScience,* 14-17 September Zurich (CHE). 150-156.

Napolitano, M. and Mooney, P., 2012, MVP OSM: A Tool to identify Areas of High Quality Contributor Activity in OpenStreetMap. *The Bulletin of the Society of Cartographers,* 45 (1), 10-18.

Neis, P. and Zipf, A., 2012, Analyzing the Contributor Activity of a Volunteered Geographic Information Project—The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information,* 1 (2), 146-165.

Nielsen, J., 2006. *The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities* [online]. Nielsen Norman Group. Available from: http://www.useit.com/alertbox/participation_inequality.html [Accessed 2012-10-26].

Nov, O., Arazy, O. and Anderson, D., 2011. Technology-Mediated Citizen Science Participation: A Motivational Model. *Proceeding of the Fifth International AAAI Conference on Weblogs and Social Media,* July 17-21 Barcelona (ESP). Menlo Park (USA): The AAAI Press, 249-256.

OpenStreetMap contributors, 2014. *Main Page* [online]. OpenStreetMap Wiki. Available from: http://wiki.openstreetmap.org/wiki/Main_Page [Accessed 2017-06-19].

OpenStreetMap contributors, 2017. *History of OpenStreetMap* [online]. OpenStreetMap Wiki. Available from: http://wiki.openstreetmap.org/w/index.php?title=History_of_OpenStreetMap&oldid=1425869 [Accessed 2017-04-07].

O'Reilly, T., 2005. *What is web 2.0: Design patterns and business models for the next generations software* [online]. O'Reilly Media, inc. Available from: http://oreilly.com/web2/archive/what-is-web-20.html [Accessed 2017-12-04].

O'Reilly, T. and Battelle, J., 2009, Web squared: Web 2.0 five years on. *Web 2.0 Summit,* .

Preece, J. and Shneiderman, B., 2009, The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction,* 1 (1), 13-32.

Raymond, E., 1999, The cathedral and the bazaar. *Knowledge, Technology & Policy,* 12 (3), 23-49.

Rehrl, K. and Gröchenig, S., 2016, A Framework for Data-Centric Analysis of Mapping Activity in the Context of Volunteered Geographic Information. *ISPRS International Journal of Geo-Information,* 5 (3), 37.

Rehrl, K.*, et al.*, 2013. A conceptual model for analyzing contribution patterns in the context of VGI. In: J.M. Krisp, ed. Progress in Location-Based Services, Lecture Notes in Geoinformation and Cartography. Berlin (DEU): Springer-Verlag, 373-388.

Roche, S., 2012. *Should VGI map space or places? Geographic Information Science Workshop Presentation (GIScience 2012)* [online]. Oak Ridge National Laboratory. Available from: http://www.ornl.gov/sci/gist/workshops/2012/vgi_documents/Roche.pdf [Accessed 2012-11-23].

Ryan, R.M. and Deci, E.L., 2000, Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist,* 55 (1), 68.

Sabone, B. 2009, *Assessing Alternative Technologies for Use of Volunteered Geographic Information in Authoritative Databases*. Unpublished Master of Science in Engineering Thesis, Department of Geodesy and Geomatics Engineering, University of New Brunswick, Fredericton, N.B., Canada. Available online at https://unbscholar.lib.unb.ca/islandora/object/unbscholar%3A8421.

See, L., et al., 2013, Comparing the quality of crowdsourced data contributed by expert and non-experts. *PloS one,* 8 (7), e69958.

Sun, N., Rau, P.P. and Ma, L., 2014, Understanding lurkers in online communities: A literature review. *Computers in Human Behavior,* 38, 110-117.

Utrilla, P., et al., 2009, A palaeolithic map from 13,660 calBP: engraved stone blocks from the Late Magdalenian in Abauntz Cave (Navarra, Spain). *Journal of human evolution,* 57 (2), 99-111.

Zielstra, D. and Zipf, A., 2010. A comparative study of proprietary geodata and volunteered geographic information for Germany. *13th AGILE International Conference on Geographic Information Science,* June 4 Guimarães (PRT).

Chapter 2: Contributors' Enrollment in Collaborative Online Communities - the Case of OpenStreetMap

Abstract: The number of people registering in an online community depends on two main factors: interest in, and awareness of, the project. Registering to a project does not, however, imply contributing to it, as lacking the knowledge and skills can be a barrier to participation. In order to identify the nature of events that might have facilitated or hindered enrollments in the OpenStreetMap (OSM) project over time, we analyzed the correlations between the number of new participants and the events that dotted its history. Four different metrics were defined to characterize participants' behaviours: the daily number of registrations, the daily number of participants that made a first contribution, the delays between contributors' registration and their first edits, and a daily contribution ratio built from the number of new contributors and the number of new registered members. Time series analyses were used to identify trends, and outstanding variations of the number of participants. An inventory of events that took place along the OSM project's history was created and appreciable variations of the metrics have been linked to events that seemed to be meaningful. Although a correlation does not imply causality, many of the explanations these correlations suggest are supported by the results of other studies, for instance when we consider the time participants spend as "lurker," or the nature of the contribution from early participants. In other cases, they suggest new explanations for the origin of the spam accounts that affect registration statistics, or the decline in the proportion of registered members who actually become contributors

Keywords: OSM; contributors; lurkers; participation; knowledge; motivation

2.1. Introduction

With the advent of the Web 2.0, contributing to an online community of interest has never been easier and the improvement of Web applications removed most of the barriers linked to physical distance or volunteers' availability (Bryant, Forte, and Bruckman 2005). These

communities play an important role in today's society that increasingly value them as a credible source of information and the scientific community is increasingly referring to these communities as both a valuable work force and an important data source (Riesch and Potter 2014, Kimura and Kinchy 2016, Michelucci and Dickinson 2016). Volunteers' motivations for contributing to online projects have been well studied in the scientific literature (Ryan and Deci 2000, Stebbins 2015, Penner 2002, Clary 1998, Nov 2007, Budhathoki 2010, Borst 2010) and in summary; the number of people registering to a project depends on two main factors: interest and awareness.

First, the project must be of interest to potential contributors, which means that it must be perceived as being either relevant, appealing or both. These are considered as internal factors to the project. A project is relevant when people expect it to meet their needs, desires or aspirations, whether because of the nature of the task (Houle 2005, Borst 2010, Hemetsberger and Pieters 2003), or because of the project's objectives (Nov *et al.* 2011, Aknouche and Shoan 2013). People will find a project appealing if they foresee that their participation will be enjoyable or even fun (Budhathoki *et al.* 2010, Aknouche and Shoan 2013). Furthermore, registering to a project does not imply contributing to it, and the phenomenon of lurkers (i.e. members who do not immediately contribute) is well described in the literature (Preece *et al.* 2004, Schneider *et al.* 2013, Sun *et al.* 2014). These lurkers may be new members that have been confronted to a reality that differs from their expectations, preventing them from contributing for various reasons. In the context of volunteered geographic information (VGI), the knowledge and skills required to contribute can be certainly be an obstacle (DiBiase *et al.* 2006, Downs

and DeSouza 2006).

Second, potential contributors must be aware the project exists. These are identified as external factors to the project. In order to see the number of participants growing, most of them must like their experience and share it with others to slowly expand the circle of participants within friends, colleagues or groups of interests (Brown and Reingen 1987, Hemetsberger and Pieters 2003, Rogers 1983). If this process is successful, the community will eventually reach people on a much larger scale through blogs, conferences, or even mass media that can lead to an exponential growth (Tichenor *et al.* 1970, Rogers 1983) that is typical of most successful online communities.

The number of participants that enroll and contribute to an online project therefore depends on complex interactions between the project characteristics (e.g. objectives, infrastructure and community) and the participants' profile (e.g. motivation, expectations, knowledge and skills) as they evolve each other over time. Understanding these interactions and their relative impacts on an online collaborative project could help concerned people to decide what actions to take, or not to take, to allow these communities to grow and remain healthy. Unfortunately, little has been published about the actual effects such interactions have on the evolution of the number of contributors in an online project.

In order to apprehend the complex interactions between these factors, different metrics were used to assess the evolution of enrollments of a large VGI project. Since the factors that influenced the decision of individual participants to enroll are not known, the

correlations between their enrollment and the events that dotted the history of the project were used as proxy indicators. Although correlation does not imply causality, many of the correlations found suggested explanations that are supported by the literature while in other cases, they suggest new explanations that will need to be explored further.

This paper presents four metrics used to assess the enrollment of participants in the OpenStreetMap (OSM) project over time. It describes the procedures elaborated to prepare and analyze the data and discusses the variations that affected the metrics and their correlations with the events that dotted the history of the project.

2.2. Materials and methods

OpenStreetMap is a project of general interest that aims at mapping the world using a Wiki approach. Similarly to Wikipedia, participants decide what, when and where they contribute without any constraints, the respect of the community's guidelines being validated *a posteriori* by the other participants or by bots (OpenStreetMap contributors 2014: "Good practice" and "Editing Standards and Conventions" pages). With more than 4 million registered users, OSM has become the most successful VGI project on the web, even though the level of technical knowledge required to contribute is higher than the average collaborative community.

Furthermore, the project is very well documented and the data are freely available. The history of the project (e.g. technical improvements, normative changes, social activities) can, in parts, be reconstructed from the OSM blog (OpenStreetMap

Foundation 2017) and the OSM documentation wiki (OpenStreetMap contributors 2014). Information about individual OSM members is available through the OSM application programming interface (API) (OpenStreetMap contributors 2014: "API v0.6" page). Their personal profiles provide, among other things, the username, the registration timestamp, the number of contributions made and an optional free text field that can be used by the participants to present themselves. Contributions to the project are made available on a regular basis through history dump files (OpenStreetMap contributors 2014: "Complete OSM Data History" page). Those files contain all the edits made since the beginning of the project up to the release date of the dump files. In addition to the edits, the file also contains the virtual containers (changesets) that identify the content, the contributor and both the geographical and temporal extents of each editing session.

2.2.1 Metrics

The literature has proposed multiple metrics to study the OSM project, either from the nature of contributions (Neis and Zipf 2012, Steinmann *et al.* 2013, Corcoran *et al.* 2013, Rehrl *et al.* 2013), the quality of the data (Girres and Touya 2010, Keßler and de Groot 2013), the profiles of its contributors (Budhathoki *et al.* 2010) and the interactions they have between them (Mooney and Corcoran 2013, Arsanjani *et al.* 2015).

We used four metrics to characterize the participation to OSM on a temporal perspective. The first metric is the "daily number of new registered members" which aims at assessing variations in people's interest and awareness about the project. The second metric is the "daily number of new contributors" which provides both the number of

registered members who made a first contribution and therefore those who did not contribute yet. The third and fourth metrics are derived from the previous two. The "contribution ratio" results from dividing the number of new contributors by the number of new registered members on a daily basis, and the "contribution delay" which is the time spanned between contributors' registrations and their first contributions (i.e. the time spent as lurker).

2.2.2 Information retrieval

As a part of a larger project that started three years ago, a history dump file released on 1 September 2014, was downloaded from the OSM web site (OpenStreetMap contributors 2014: "Complete OSM Data History" page). FME (Safe software) workbenches were developed to extract and load to a PostgreSQL database the data from both the history dump file and from queries made to the OSM API. Statistical analyses and visualizations were carried out using R software. The observations used in this study were built from the timestamps of all contributors' first edits and an estimation of the registration's timestamps of all OSM members at that time. The dates of contributors' first edit were obtained from the creation timestamp of their first changeset, and the daily count of new contributors was based on these dates.

Obtaining the daily count of registrations would have required querying the OSM API for over 2.3 million individual profiles (as of 1 September 2014). Instead, only contributors' profiles were retrieved and their registration timestamps were used to approximate those of the remaining members (i.e. lurkers). These registration timestamps

were linearly interpolated using the R's "approx" procedure (R Core Team 2016) over the whole range of members' identifiers (ID) generated over that period (according to the IDs found in the history dump). The accuracy of the resulting timestamps was assessed over a sample of 3074 evenly distributed lurker profiles.

An inventory of the events that dotted the history of the OSM project was retrieved from the OSM Wiki pages (OpenStreetMap contributors 2014: "History of OpenStreetMap," "Past Events," "OpenStreetMap in the media," "Development activity" pages) and some OSM mailing lists were consulted (i.e. the general "talk," development ["dev"] and "legal" mailing lists). Since building an event classification was outside the scope of this research, we adopted the event categories developed by the OSM community (OpenStreetMap contributors 2014: "Current events" page) to include development milestones, media news and internal announcements (i.e. blogs and mailing lists). Categories were grouped under internal and external factors. Internal factors are categories of events that set or change the project's characteristics and determine whether the project is relevant or appealing to an individual, such as new rules or application improvements. External factors are categories of events that affect the number of people that may be aware of the project (i.e. project visibility), the perception they may have about the project, or both, such as media coverage or conferences. Within the different categories (presented later in Table 2-1), the "Mapping" category is a special case combining activities that are inherent to the project, but mostly impacted the visibility of the project (i.e. classified as external factors). Mapping parties (i.e. typical social gathering oriented toward a mapping task) have increased the visibility of the OSM project by bringing new

participants (Haklay and Weber 2008, Mashhadi *et al.* 2015). Similarly, the mapping efforts made by the OSM community after natural disasters have also increased the visibility of the project to international relief organizations (Horita *et al.* 2013, Zook *et al.* 2010).

2.2.3 Invalid account removal

Online collaborative projects often see user accounts removed by administrators, either because the users were banned or the accounts created to spam the project. A stratified random sample of members' profiles was performed to assess the proportion of these accounts over time and remove the accounts from the registration statistics we generated. Two random profiles were retrieved for every 1000 sequential ID. Instances for which the API did not return any profiles were invalid accounts removed by OSM admin and considered as such in our analysis. Three fields from users' profile were used to identify potential spam accounts: the username, the content of the free personal text field, and the number of contributions made.

The free text field of 4604 sampled profiles was first analyzed to identify possible spam content. Anticipating username patterns in spam accounts, all usernames were compared considering whether the accounts were flagged as spammed or not, contributed to the project or not, and the time at which they registered. Identified patterns were translated into a regular expression to identify most of invalid accounts from our sample, while minimizing erroneous identification of legitimate accounts. The proportion of invalid accounts was assessed over time using a moving average on a 101 samples

window (i.e. covering about 50,000 consecutive IDs) and was set to constant values on the edges.

2.2.4 Time series analysis

Standard time series analyses postulate the presence of a stochastic process, dividing the process into a centred random component and deterministic trend and seasonal components (McLeod *et al.* 2011, Hyndman and Athanasopoulos 2014). The trend component is used to assess the long-term variations in rates of registrations and initial contributions. Turning points in trend curves may result from changes in either the popularity of a project, the ease with which participants can contribute, or both. We expected these variations to correlate with events that had a long-term impact on the project. The seasonal component is expected to identify recurring events that modulate the rate of registrations and initial contributions. Finally, the random component should highlight outstanding variations of enrollment. The correlations with specific types of events may reveal clues about what affected participants' behaviour, such as some downtime from servers, or the coverage of the project by mass media.

The decomposition of the time series was performed using R's "decompose" procedure (R Core Team 2016). A yearly cycle was used as the time unit for seasonal variations, resulting in 365 observations (days) per unit. The determination of the "trend" components over a yearly cycle left 182 days without value on each side of the curve. The results are expressed as an average number of participants. The seasonal components are computed by averaging observations over each day of a year after the trends are removed.

In this case, trend components were removed by dividing observed values by the trends (creating a ratio) to take into account variances dependency on the means for both distributions (Hyndman and Athanasopoulos 2014). The seasonal components are then expressed as a proportion of the trend and the 365 resulting values are duplicated as necessary over the whole range of observations. The random components result from removing both the trend and the seasonal components from the observed values and are also expressed as a proportion of the trend.

2.2.5 Contribution delays and contribution ratios

Contribution ratios were obtained by dividing the trend component of initial contributions by the trend component of registrations. The resulting daily ratios provided the proportions of registered members that contributed to the project over time. The contribution delays were obtained by computing the time span between contributors' registration and their first edits. Daily averages and medians of computed delays were plotted to understand how they evolved over years.

2.2.6 Events Associations

Abrupt variations in the metrics were correlated to events that made the history of the project. A manual identification of major turning points was made on the trend component. Outstanding variations (outliers) found in seasonal and random components were identified using the R's "Boxplots" procedure (R Core Team 2016). Potential explanatory events were searched within a few days from identified variations and a qualitative analysis of the events was used to select the most relevant ones. The analysis

considered changes in the volume of participants prior and after each event and the potential number of people reached, or affected, by these events.

2.3 Results

The event repository counted more than 3560 events that dotted the history of the OSM project from 2005 to September 2014. Events were classified into seven categories and two factors that are shown in Table 2-1.

Table 2-1 Classification of events related to the OSM project Wiki.

Category	Factor	Category description
Meeting	Internal	Administrative, development and social activities
Upgrade	Internal	Infrastructure and software upgrade implementation
Forum	Internal	Mailing lists announcements and OSM Foundation blog
Licence	Internal	Contributor terms and ODbL[1] licence change milestones
Mapping	External	Mapping parties/efforts, including humanitarian activities
Conference	External	Conferences mentioning/discussing the OSM project
Media	External	Media coverage about OSM or related topics

[1] OSM switched to an Open Database Licence (ODbL) after a lengthy process that lasted for four years.

Internal factors regroup 1350 "Meeting," 135 "Upgrade," 52 "Forum" and 8 "Licence" events. External factors counted 725 "Mapping," 369 "Conference" and 939 "Media" events. With only a few exceptions, all potential explanatory events were found within a week or so from identified variations in participants' behaviour.

2.3.1 Invalid account removal

Interpolated registration timestamps (i.e. lurker registration) proved to be accurate, with a standard deviation of 37 minutes and 95% of observations being within one hour from

their actual timestamps. The resulting rates of registrations were compared to the rates of new contributors over time. While the number of contributors was increasing steadily, the number of registered members exploded on 8 July 2012, rising from an average of 704 to 2259 registrations a day, a volume that remained high over most of the period covered by the dataset (Figure 2-1).

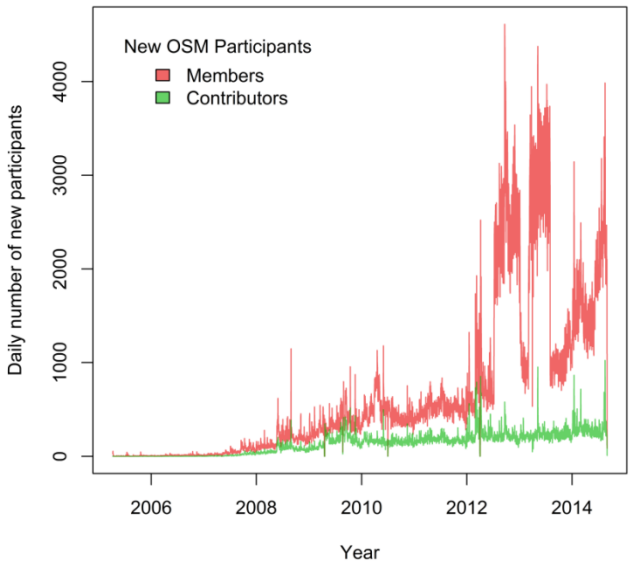


Figure 2-1 Distribution of new OSM members and new contributors over time.

Analysis of sampled profiles revealed that after 8 July 2012, large proportions of new accounts were created with spam contents in their text field. The examination of the text field revealed that on average 25% of all accounts created between July 2012 and September 2014 contained spams and, with only a few exceptions, spam contents affected only lurkers. Spamming contents were mostly random texts, without obvious purpose, that may result from search engine optimization (SEO) procedures.

Height discontinuities (i.e. rupture points) were identified in spam accounts distribution, generating nine segments (i.e. periods) over which the rate of bot accounts creation was relatively constant. These rupture points were compared to the list of events from OSM history to find potential relationships, and the results are presented in the table below (Table 2-2).

Table 2-2 Rupture points in spamming processes and potentially related events. "Date" refers to the rupture points, "Prior" and "After" show the proportion of accounts potentially derived from bot processes for each rupture points, "Day" is the number of days between the rupture point and the most relevant event found in the list within the surrounding days.

Date	Prior	After	Day	Potentially relevant events
2010-03-02	1%	31%	7	LWG meetings solved outstanding problems with ODbL
2010-05-06	31%	9%	6	New users must agree to ODbL to register
2010-09-17	9%	1%	7	OSC2010 conference in Tokyo/Fall[1]
2012-07-08	1%	69%	1	Data deletion of those who rejected ODbL about to begin
2013-01-05	69%	30%	1	OSM reached one million registered users
2013-03-05	30%	71%	10	SOTM France—2013 National OSM conference
2013-08-03	71%	14%	0	SOTM Baltic—2013 Baltic OSM conference
2014-06-09	14%	38%	4	SOTM Europe—2014 European OSM conference

[1] No obvious link except that a similar conference (OSC2010 Tokyo/Spring) was held a week prior spams began.

The results show two periods during which spam accounts were created on a larger scale. Both periods happened while the community was discussing a switch to a new license to better protect the data provided by OSM participants. The first one spans from March to September 2010, a six-month period after the OSM legal working group (LWG) resolved the remaining problems around the ODbL licence implementation. The second one started in July 2012, just before the data from those who did not agree to the ODbL licence were removed from the database. The creation of spam accounts has continued

over the period covered by the analysis with few rupture points that matched some State of the Map (SOTM) conferences in Europe.

According to our sample, the characteristics of lurkers' usernames changed significantly over the periods the spamming processes were active. During these periods, 88% of spammed accounts showed specific patterns of English words and digits in their usernames. As anticipated, such patterns were rarely seen for contributors (5%), or for lurkers outside these periods (7%). Three distinct patterns were identified and combined in a regular expression to estimate the proportion of accounts created by spamming processes. The regular expression was applied to our samples, identifying 530 of the 603 spam accounts, a detection rate of 88%. Only 47 of the 940 legitimate contributors were flagged as spam account resulting in 5% false positives. Equation (2-1) was used to estimate the proportion of OSM accounts generated by bots over time (Pbots):

$$Pbots = MAX(Pregex - 0.05,0) + Pspam, \quad (2\text{-}1)$$

where Pspam is the proportion of spam accounts and Pregex is the proportion of lurkers' usernames that matched the regular expression excluding spam accounts. The proportion was adjusted to compensate for the 5% false positives resulting from the regular expression.

The distribution of registration rates prior and after the correction is shown in Figure 2-2. Dark green segments indicate where both curves overlap (i.e. no bot accounts detected). Red segments illustrate removed bot accounts and the light-green segments

show the actual number of registered members after the correction. Our results suggest that the spamming processes succeeded to seed spams in 54% of the accounts they created during the period covered by the analysis. OSM administrators were able to close about one third of these accounts.

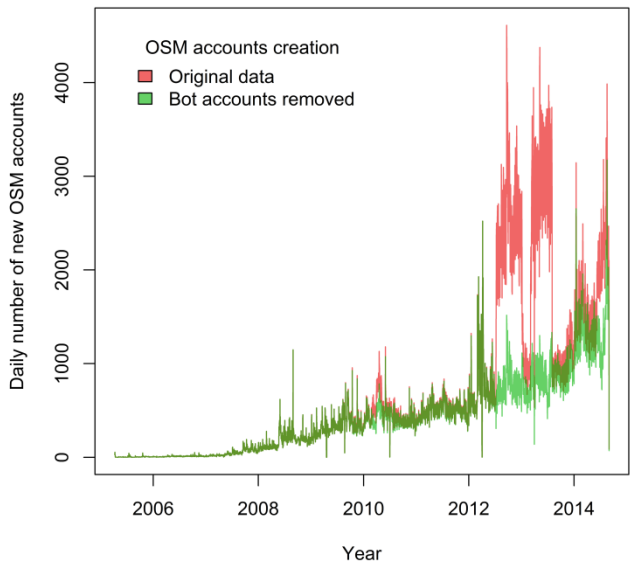


Figure 2-2 New OSM accounts prior (red) and after (green) bot account removal.

In order to assess if the spamming processes were still active, a sampling of the accounts created beyond this period until February 2017 was made. The result shows that spam accounts creation processes were still active with about half of newly registered members that may not be legitimate. The 3299 sampled profiles revealed that 10% of the accounts contained spams, another 10% had been closed by the OSM administrators, and more than 30% of the profiles were lurkers having a username pattern that match our regular expression.

2.3.2 Time series analysis

The data present two continuous sequences of discrete time-ordered observations that display increasing averages and variances with positive and negative peak events. Analyses results are presented on Figure 2-3.

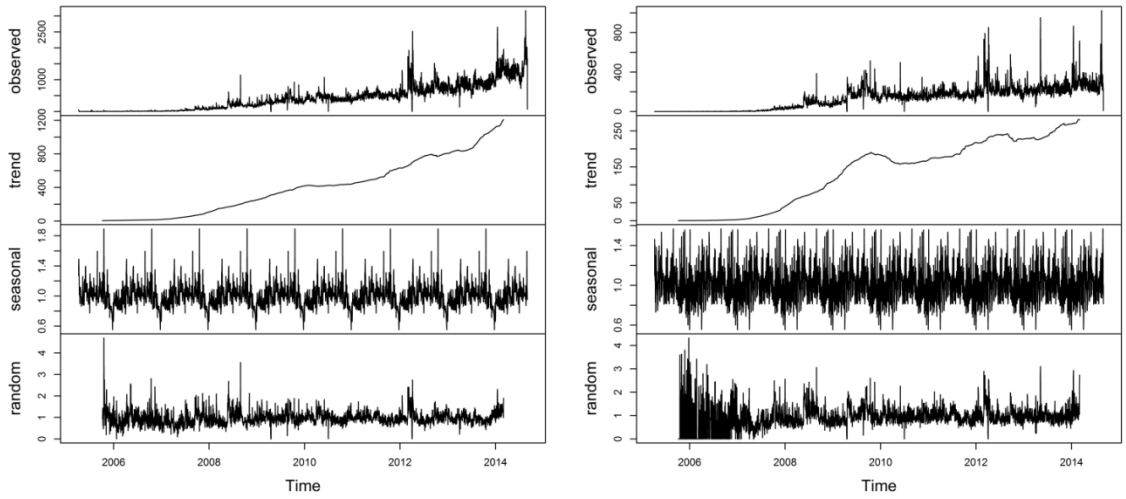


Figure 2-3 Compared time series analysis plots of (a) rates of new registered OSM members and (b) rates of new contributors, showing the observed values, the trend, the seasonal, and the random components Scales of Observed and trend values are the actual number of people, seasonal and random values are a proportion of the trend value.

2.3.2.1 Variations in seasonal components

Seasonal variations (Figure 2-3 seasonal) of registration rates follow an inverted U shape and are repeated annually over the studied period (Figure 2-3a). Average registrations are 10% above normal from April to October and 10% below normal from November to March, with a clear minimum in December (-30%). A similar pattern is seen for new contributors (Figure 2-3b). This could potentially reflect a higher interest of northern hemisphere participants to be involved in an outside activity during warmer months. No

relationship was found between seasonal peak events and known recurring statutory holiday or vacations, with the exception of Christmas Eves (minimums of both distributions). Most outstanding seasonal variations echoed large peaks of participation rather than recurring yearly variations because of the short history of the project. These peaks of participation influenced the average value of recurring variations per time unit because the number of cycle was too low.

2.3.2.2 Variations in random components

Random variations (Figure 2-3 random) show numerous peaks on both distributions. These peaks identify specific days when an unusual (i.e. small or large) volume of participants registered or made a first contribution to the project. Largest bursts of registration are expressed in Table 2-3.

Table 2-3 Outstanding random variations of new OSM members with explanatory events.

Outlier	Value[1]	Category	Associated explanatory event description
2005-10-20	4.70	Forum	OSM Promotional wallpapers and posters for sale
2006-05-14	2.54	Mapping	Mapping weekend at Manchester (GBR)
2006-10-16	2.81	Media	BBC reporting on Rutland's mapping party (GBR)
2008-05-30	2.69	Media	Der Spiegel (GER) compares OSM to Wikipedia
2008-08-29	3.56	Media	BBC quotes BCS[2] being positive about OSM
2012-02-29	2.41	Media	Report that Foursquare quits Google Map to join OSM project
2012-04-06	2.74	Media	Report that Wikipedia apps are now using OSM
2014-01-15	2.31	Media	Relay a blog about Why the World Needs OSM

[1] Value of the random component found in Figure 2-3 (a)
[2] British Cartographic Society

Regarding the daily number of new registered members, 123 outlier values were identified out of 3069 observations. Within these outliers, 22 days showed much smaller ratios while 101 days showed much higher ones. Low registration ratios happened mostly

at the beginning of the project without obvious related events except for connection problems with the servers, while they all occurred on planned servers' downtime later in the history of the project.

The nature of the events that correlate to high registration rates has evolved over time. At the beginning of the project, burst of registrations often followed technical threads in OSM forums (29%), upgrades (18%), or Open Source Software (OSS) conferences (16%), until the media (81%) took over after 2007. Largest bursts of registrations were mostly correlated to external factors (e.g. Media, Mapping).

Regarding the daily number of new contributors, 330 outliers were identified in which 199 days showed much smaller ratios, and 131 days much higher ones. The events that correlated with a small number of new contributions had similar explanations to the ones found for registrations. The events that correlated to high numbers of registration were dominated by upgrades (65%), mapping parties (12%) and forum threads (16%), the latest being mostly related to new editors and data importing tools. After 2007, the correlations shifted to media (43%), upgrades (34%) and mapping parties (11%). Largest bursts of new contributions are shown in Table 2-4.

Table 2-4 Outstanding random variations of new OSM contributors with associated explanatory events.

Outlier	Value[1]	Category	Associated explanatory event description
2005-12-27	4.33	Upgrade	Latest version of "osmeditor" is made available
2008-08-29	3.56	Media	BBC quotes the president of the BCS being positive about OSM
2012-02-29	2.89	Media	Report that Foursquare quits Google Map to join OSM project
2013-05-08	3.10	Upgrade	New ID editor is made available on the OSM web site
2014-01-15	2.31	Media	Relay a blog about "Why the World Needs OpenStreetMap."

[1] Value of the random component found in Figure 2-3 (b)

2.3.2.3 Variations in trend components

The trend components (Figure 2-3 trend) show the cumulative effect of all the events that dotted the history of the project. Some of these events may have played an important role in the way the project evolved over time. Over a dozen turning points were identified independently for each curve. With one exception, all these points found to be paired with each other over the same dates. The five largest turning points were selected for discussion and are presented in Figure 2-4.

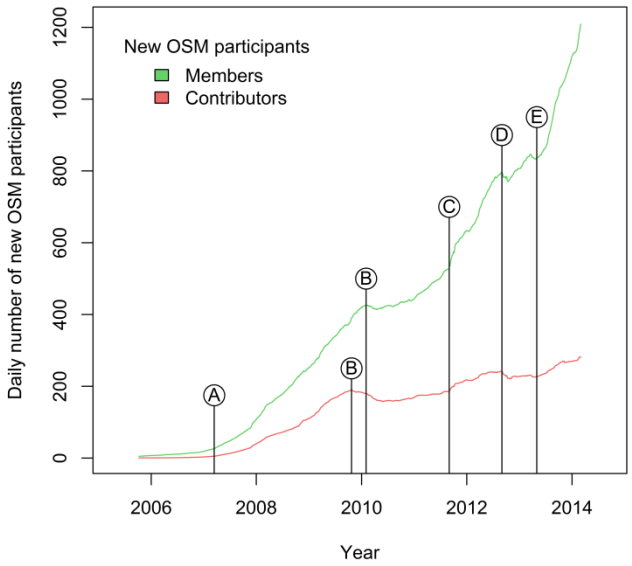


Figure 2-4 Trends in new OSM members and contributors with selected turning points events (A-E)

The OSM project really started attracting new participants after March 2007 (A); two years after the first contributions were made. Over the preceding months, high-resolution images from Yahoo! had become available to contributors[2], the project moved

[2] https://lists.openstreetmap.org/pipermail/talk/2006-December/009448.html

to API 0.4[3], and a user-friendly editor (i.e. Potlatch 1) was set up in the "Edit" tab of the project's web page[4]. Over the same time, the founder of the project, Steve Coast, published his thought about the need for the project in relation to the products offered by national mapping agencies (OpenStreetMap Foundation 2017).

The second points (B) show that both curves toppled around 2010. In October 2009, the OpenStreetMap Foundation (OSMF) board announced that its members (not OSM participants) were to vote for a licence change[5]. While the project had seen a steady increase of its participants (i.e. members and contributors) over the previous two years, the daily number of new contributors started declining at this time, followed by the number of registrations four months later, a month after the OSMF voted in support of the licence change.

During this period and over the following two years, harsh discussions were held on different forums about the licence change. Some of the members who did not agree with the licence change even copied the entire database and started a similar project under the previous terms (i.e. "forked" the project)[6]. Finally, since only a small proportion of OSM members declined the ODbL licence (Weait 2011), the OSMF called the case closed during a SOTM conference held in September 2011 (C) and decided to move ahead with the change.

The number of OSM participants then rapidly increased until the next turning point

[3] https://wiki.openstreetmap.org/wiki/API_v0.3
[4] https://lists.openstreetmap.org/pipermail/talk/2007-May/013920.html
[5] https://lists.openstreetmap.org/pipermail/talk/2009-December/045105.html
[6] https://groups.google.com/forum/#!topic/osm-fork/74iTTj6qXCk

(D) in September 2012, which correlates with the Tokyo SOTM 2012 conference, when the OSMF board passed a resolution to implement the new licence. The burst of new participants visible between points (C) and (D) correlated with the many online communities that changed their background maps from Google to OSM during this period[7]. However, since it ends with the change to the ODbL licence, it might also be related to calls made to the community to remap the data "tainted" by contributors who did not agree to the new license. The exercise aimed at remapping the data provided by those who did not agree with the ODbL licence, before and after a redaction process (bot) removed all their data from the database.

Over the following months (D to E), a short burst of registrations seemed to be related to the announcement of a commercial users' summit to be held later[8]. In May 2013 (E), a new OSM editor called ID was made available, leading to an increase in the number of participants[9].

2.3.3 Contribution delays

Generally, the average delay between users' registration and first contribution shortens gradually over time, while median delays shorten step by steps (Figure 2-5). The horizontal dotted line of Figure 2-5 represents a one-day delay.

The graph shows that both distributions mostly overlapped until December 2006 (A). During this period, contributors waited on average 604 days before making a first

[7] https://www.technologyreview.com/s/426481/wikipedia-of-maps-challenges-google/
[8] https://blog.openstreetmap.org/2012/10/22/weekly-osm-summary-54/
[9] https://blog.mapbox.com/24-hours-of-new-openstreetmap-users-fbf047d9ac11

edit. In April 2009 (E), the average delay dropped below 31 days and the median delays stabilized below 20 min which may correspond to the implementation of the API 0.6.

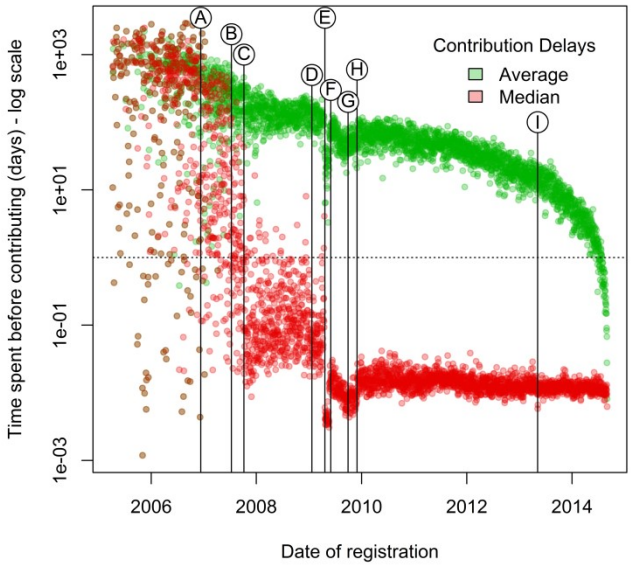


Figure 2-5 Delays between a user registration and contributions, and key turning point events.

The first drops in the median distribution (A and B) correlates with the arrival of the Yahoo! aerial imagery in OSM applet in December 2006[10], and then in JOSM (i.e. a popular OSM editor) eight months later[11]. A similar effect was found in October 2007 (C) after which most median delays dropped below one day which corresponds to the time at which the API 0.5 was implemented[12], a move that, according to the forums content, made queries and edits easier. Another drop in on both distributions (D) appears in January 2009, with an important compression of the range of values. This change co-occurs with a

[10] https://lists.openstreetmap.org/pipermail/talk/2006-December/009448.html
[11] https://lists.openstreetmap.org/pipermail/talk/2007-July/015682.html
[12] https://lists.openstreetmap.org/pipermail/dev/2007-September/006808.html

new version of the Potlatch editor[13] that brought more presets and a detailed coverage of England & Wales with public domain maps (i.e. to copy at will).

The release of the API 0.6 in mid-April 2009 (E) fitted with a significant break in both distributions[14]. Afterward, the delays reached a minimum for a month and a half, until the distributions broke again toward higher values at the beginning of June 2009 (F). This jump to higher delays matches the announcement on the OSM blog that the OSM web site was now available in German and partly in French[15]. Prior to that, the site was available only in English. Delays dropped again until late September 2009 (G) when the delays increased suddenly. This corresponds to the time at which the web site went available in 26 more languages[16]. The effects on both distributions were similar to what happened at the time the web site was translated into German and French (F).

In December 2009 (H), the delays stabilized at the time the Potlatch 2 editor was released[17]. After the advent of the Potlatch 2 editor, both distributions remained generally stable with trends toward shorter delays. However, this trend became stronger for average delays in 2013. At the same time, the mean and the variance of the median delays slightly dropped around May 2013 (I). Both changes occurred around the time of the arrival of a new OSM editor. The ID editor was made available on the web site in May 2013[18] and became the default OSM editor in August of the same year. Past this point, the average

[13] https://potlatchosm.wordpress.com/2009/01/21/more-presets/
[14] https://wiki.openstreetmap.org/wiki/API_v0.6
[15] http://www.h-online.com/open/news/item/OpenStreetMap-adds-new-translations-741829.html
[16] https://blog.openstreetmap.org/2009/09/29/osm-now-in-26-more-languages/
[17] https://lists.openstreetmap.org/pipermail/talk/2009-November/044783.html
[18] https://blog.openstreetmap.org/2013/05/07/openstreetmap-launches-all-new-easy-map-editor-and-announces-funding-appeal/

delay dropped rapidly to a point until it joined the median distribution. The latest measurements are affected by their proximity with the closing date of the history dump file used in this research. All those who signed up to the project prior that date and contributed after are not included in the graph. Only the quickest ones will have made a first contribution, which artificially shortens the delays as we move closer to the end of the data.

2.3.4 Contribution ratios

Figure 2-6 reveals that the contribution ratios increased from 5% in 2005 to 50% in 2009, before they declined to reach 27% at the end of the period covered by the analysis.

The first turning point appears in May 2006 (A). According to the list of events, a first collaborative mapping weekend was held in Manchester (GBR), attracting new volunteers that were initiated to GPS and mapping operations[19]. The events that matched the second (B) and third (C) turning points are likely linked with each other. In October 2008 (C), the OSM administrators opted out from a web site called BugMeNot.com[20] and blocked the related OSM accounts. This site allows people to connect to web sites requiring personal accounts by making public the logins (i.e. username and password) of a few accounts. In other words, it was enabling people to contribute anonymously. Using this clue, we found that just before the contribution ratios started lowering in May 2008 (B), someone named "bugmenot" inquired about mapping in an OSM blog[21].

[19] https://wiki.openstreetmap.org/wiki/Mapchester_2006-05-13
[20] https://lists.openstreetmap.org/pipermail/talk/2008-October/030122.html
[21] https://blogs.kde.org/2008/04/22/open-whitewater-maps

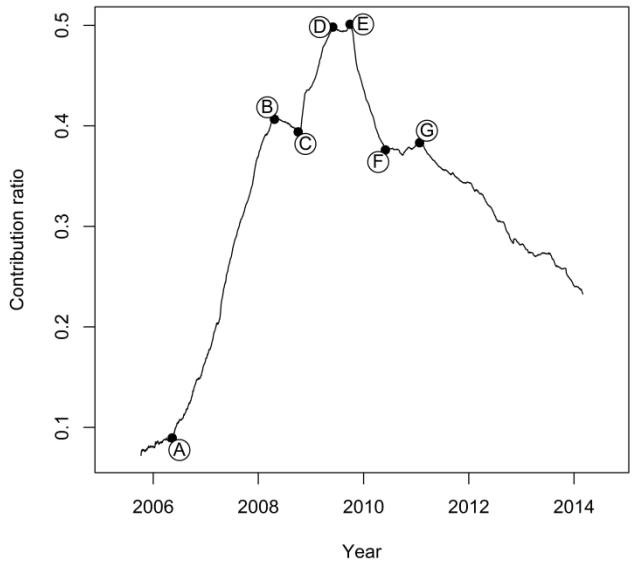


Figure 2-6 Evolution of contribution ratios over time and key turning point events.

In June 2009 (D), the contribution ratios stabilized until October 2009 (E) when it started dropping. The event repository provided similar potential explanations over both turning points. In June 2009 (D), the OSM web site was made available both in German and in French[22] and five months later (E) the site was available in 26 more languages[23], in an attempt to make the registration process more accessible by non-English-speaking contributors. Finally, the drop of contribution ratios paused between May 2010 (F) and January 2011 (G). In the first case (F), the event repository showed that - two weeks prior this turning point - the new OSM members had to accept the ODbL licence to register from then on[24]. The search of an explanatory event for the last point (G) was not successful. Neither the list of events nor the different mailing lists we consulted provided a

[22] http://www.h-online.com/open/news/item/OpenStreetMap-adds-new-translations-741829.html
[23] https://blog.openstreetmap.org/2009/09/29/osm-now-in-26-more-languages/
[24] http://wiki.openstreetmap.org/wiki/Open_Database_License/Contributor_Terms/Human_readable

meaningful event.

Analyzing the 3299 profiles sampled to assess current spamming processes, we estimated that, between September 2014 and February 2017, the average proportion of new members that made a contribution was about 25%. This estimation was obtained using the number of legitimate accounts found in our sample (i.e. not spammed and not matching our regular expression) and the number of these accounts that had at least one contribution. This proportion is similar to the latest values obtained in September 2014.

2.4 Discussion

The number of participants that register to a project depends on the number of people that are aware of the project and the interest it generates after they have discovered it. This interest is in turn determined by the perceptions individuals have about either project's relevance, attractiveness or both, depending on how they heard about it. The correlations we found between significant variations in participants' behaviour and some events that dotted the history of the project tells us a story about both the project's evolution and what affected participants' motivations. Although correlation does not imply causality, it seemed a first step to link participants' behaviour and the contexts in which they have enrolled in the project. Three distinct phases were identified according to participants' behaviours when enrolling in the project. These phases were found regarding both project's development and the sources that made participants aware of the project.

2.4.1 Enrollment and project's development

If the objectives of a project do not change much over time, its attractiveness may evolve according to, among other things, the complexity of the tasks and the knowledge and skills required to contribute. Both are usually mitigated by improving documentation and applications. The more adequate they are, the greater the number of participants can be. Our results suggest that participants may have had different behaviours according to the development phases of the project and, as the project evolved, the needs and tasks required may not have attracted the same type participants. Three distinct phases were identified.

The first such phase is one of an infrastructure development that extended from 2004 to 2007. The project was initiated in 2004, but participants were able to contribute data only from 2005. The correlation found during this period suggests that the project was under construction considering its unreliable infrastructure and missing or inadequate contribution tools. Trends in daily enrollments and initial contributions (Figure 2-4) show a limited increase in the number of participants until high-resolution images and user-friendly mapping applications were made available in 2007. Event association with the random components of time series analyses showed that most of lowest enrollments and contribution rates correlated with downtime and poor servers' performances. At the same time, most of their highest rates were correlated with highly technical threads on OSM forums or OSS conferences. Furthermore, the most surprising characteristic of the phase is that more than half of those who enlisted during this period waited on average two years before contributing data (Figure 2-5). Registering to a project under construction and not been able to contribute data on the short term suggest that the primary interest of these

participants may not have been to contribute data. The correlations found rather suggest that their primary interest were rather being to contribute as developers or to support project objectives. In this context, the volume of participants for whom the project may appear as either relevant, appealing or both, remained limited considering required knowledge and skills and the uncertainty of the project.

The second one is a consolidation phase that extended from 2007 to 2009. During this period, the daily enrollment rates grew by a factor of 10 and the initial contributions by a factor of 20. This high increase of initial contributions may result from the combination of contributions from new participants and from the older ones who waited until this period to contribute data. Lowest enrollments and contribution rates were now fitting with planned downtime periods, while highest rates correlated with external events or upgrades for initial contributions. As the infrastructure and contribution tools were improving, the time the majority of new participants took to contribute (Figure 2-5) dropped from years to hours, and the proportion of them who contributed (Figure 2-6) reached almost 50%. In other words, at the end of this phase, about half of the people who enroll in the project contributed data within an hour. These correlations suggest that the volume of participants exploded only after the infrastructure was properly settled and adequate contribution tools were provided. The project was now more appealing to those who were eager to fill blank areas on the map, even if contributing still brought some uncertainties.

The last one is an operational phase that started in 2009. By definition, such phase

consists in recurring maintenance operations, updates and tool improvements made on a continuous basis. Except during the licence change conflict, the daily rates of enrollment kept increasing faster, while initial contribution rates were growing at a constant pace (Figure 2-4). Delays before contributing were now short and constant (Figure 2-5). An unexpected behaviour that characterizes this phase is a drop of the contribution ratios (Figure 2-6) that reached 27% in 2014, before it stabilized around this value after this date. This drop correlated with an improvement to the registration process when the interface was made available in multiple languages while the contribution tools and wiki pages (i.e. the documentation) were not translated simultaneously. Such potential language barrier when it comes to contributing to the project must be evaluated to ensure that contributors from all over the world can share their local knowledge, especially if it results from an increase in the proportion of participants that come from developing countries. Nowadays, contributing to the project involves little uncertainties or risks. Those who wish to fill remaining blank areas on the map, or add details to their neighborhood can easily contribute.

2.4.2 Enrollment and the sources of participants' awareness

The detailed investigation of the time series analyzes revealed correlations between high rates of enrollment and external events (i.e. media, conference and mapping activities). By definition, external events reached people from outside the project and increased the number of participants when they triggered their interest. Less than 5% of registered external events correlated with bursts of enrollments, but a few of these events have had a

large impact on participants' behaviour. Furthermore, we observed that the nature of these significant events shifted over time from individuals to collective, authoritative to social.

The literature has investigated the effects of "important others" on people's motivation to enroll in volunteered activities, either because of emotional links or because of their credibility (Fishbein and Ajzen 1975, Metzger 2010, Rogers 1983). OSM participants certainly had an influence on friends or colleagues to enroll or not in the project, but the private nature of these events excluded them from our analysis. However, the public nature of the events that dotted the history of the project enabled us to identify other types of "important others" from the influence they seemed to have had on participants' motivations to enroll. Three phases were identified according to the nature of the events that appeared to have motivated people to enroll as the project was developing.

The first one would be described as a "close encounter" phase that was parallel to the project's infrastructure development phase. During this phase, mapping parties brought bursts of new participants. These gatherings (Hristova *et al.* 2013, Haklay and Weber 2008) provided OSM participants an opportunity to initiate friends and colleagues to the project by witnessing mapping operations and appreciating the outcome on the map. At the same time, treads from OSM forums also correlated with burst of new participants. These bursts happened either because these people, already aware of the project, enrolled after those treads, or because they used those threads to motivate friends and colleagues to enroll.

The second phase could be labelled as "seeking for authoritative approval," a

period that matched with the project's consolidation phase. Early media coverage appeared to have triggered multiple bursts of registration but not all these events had an impact. Well-established media (electronic and conventional) have triggered most of the largest bursts found during this period. For instance, when the BBC or *Der Spiegel* reported on OSM (Table 2-3), the burst of enrollments that followed may have resulted not only from media's popularity but also from their authoritativeness. Authoritative sources cited in these media were also correlated with some peaks of registrations. After the BBC reported positive comments from the president of the British Cartographic Society, the second-highest peak of registrations we found appeared the following day and lasted for a week. A similar effect was found when the web site "slashdotted.com" linked to a story citing an authoritative searcher of the domain (Goodchild 2007).

The latest one could be referred to as a "seeking for credibility" phase. Large peaks of enrollment happened after electronic media reported on large organizations that interacted with OSM. For instance, large bursts happened after electronic media reported that well-known online communities changed the Google's map background of their applications for OSM data. The origin of the bursts may be explained from two perspectives regarding the credibility it may have brought to the project. On the one hand, the concerned communities may have brought credibility to the project, at least from their members' perspective; on the other hand, the credibility of Google as a map provider may have been reassigned to OSM when it was chosen as an alternative. A similar effect could be considered for a large burst of enrollment after electronic media reported that Google's workers were caught vandalizing OSM (Garling 2012), sending the message that they may

have felt threatened by the project. The interaction of these organizations with OSM may not only have provided credibility to the project but, in some of these cases, it may also have brought large numbers of participants were eventually interested in freely enhancing the background map of their favourite applications.

The OSM's humanitarian contributions (Ahmouda and Hochmair 2017, Soden and Palen 2014, Poiani *et al.* 2016) are important activities that were widely used to publicize the project and give it credibility. Considering the large volume of media coverage reporting on these activities in the event repository, we were expecting these contributions to have triggered burst of new participants over years. However, no correlations were found that could be linked directly to these activities, except after few media reported OSM community's involvement in relief operations of Haiti earthquake. These activities probably brought large numbers of new participants to the project but, unless they have immediately triggered their enrollment, their effects could not be measured, or correlated.

2.4.3 Enrollment and project's internal conflicts

The correlations found with the licence change milestones may have revealed important characteristics of participants' behaviours during internal conflict: a potential retaliation of few offended participants and the postponement of enrollment and contributions from new members.

2.4.3.1 Retaliation of offended participant

Most of today's digital communications are impacted by unsolicited junk information

(Gyongyi and Garcia-Molina 2005, Chakraborty *et al.* 2016). Aggressive marketers use cheap SEO mechanism to advertise, sell their products, or make a web site appear as popular for search engines (Chakraborty *et al.* 2016). It is therefore not surprising that online collaboration sites, such as OpenStreetMap and Wikipedia are affected the creation of fake accounts containing spam contents (Yamak *et al.* 2016). However, according to our analyses, the fake accounts that spammed the OSM registrations for a couple of months in 2010 and since July 2012, looks like the retaliation of one or a few offended participants. Vandalizing the OSM registration with low-quality SEO could have made the site tagged as "spammy" by search engines such as Google (2011), lowering the odds that the site will appear in the first pages of the search results (Google 2012). Another consequence is that fake registrations required more resources (e.g. processing, disk space) and generated erroneous registration statistics. With about 50% of the accounts created since 2012 originating from a spamming process, more robust protection against registration bots (e.g. Captcha) should be implemented considering the current email confirmation used by OSM has proven not to be sufficient.

2.4.3.2 Postponement of enrollment and initial contributions

Internal conflicts, as the one triggered by the licence change, have the potential to throw a community apart. Communities develop around perceived shared goals, values and beliefs, a unique ethos that brings people to identify themselves with a community (Stebbins 2015, Budhathoki *et al.* 2010). The licence change process struck the values of many contributors who then vigorously opposed the change or its process. Even if the

number of opponents was low (0.002% did not agree to the new licence), their harsh opposition, mostly expressed in OSM forums, seemed to have had an effect on both the enrollment and initial contribution rates. These forums have an important role in online communities, and in the context of a peer production projects such as OSM. They build the community by sharing its values (Aknouche and Shoan 2013, Von Krogh *et al.* 2012), by developing rules and norms (Venkatesh *et al.* 2003, Fishbein and Ajzen 1975, Taylor and Todd 1995) and by discussing community issues. During the conflict, individuals' values were questioned regarding the relevance of the licence change or the validity of the process. It seems to have undermined people's perceptions about actual values and beliefs of the community, which may have refrained them from engaging until the situation was resolved. The fact that during this period many registered members refrained from making an initial contribution also confirmed that many lurkers were probing the project to see if the community were healthy and could suit their expectations (Preece *et al.* 2004, Amichai-Hamburger *et al.* 2016). The effect of the conflict decreased as the proportion of people accepting the licence increased until the licence change was finally approved.

2.4.4 Enrollment and the diffusion of innovations

Interestingly, participants' behaviours in each of these phases were also found to be very similar to those described by Rogers (1983) in characterizing people's behaviours in the early phases of the diffusion of an innovation. In early phases of the diffusion of an innovation, Rogers categorizes participants as "Innovators," "Early adopters" and "Early majority." The "Innovators" are described as participants that seek to be involved in the

implementation of a new idea, are venturesome people that are highly skilled and that can apply complex technical knowledge. This description matches the behaviour of the participants who enroll in the first phase of the project's development. "Innovators" are also known to develop diversified social relationships and friendship among other "Innovators," which could explain the origin of participants' awareness about the project in its "close encounter" phase. The "Early adopters" are described as having a great degree of opinion leadership, provide advice about an innovation and often serves as a role model for other participants and to keep this status they must make judicious decision about an innovation. This characterizes the origin of the events that triggered most of outstanding enrollments that happened during the second phase of the project when people seem to seek for authoritative approval. Finally, the "Early majority" is said to deliberate for some time before being involved but, once done, they follow with a "deliberate willingness." This may characterize participants that enroll in the third phase of the project, both from the development perspective (they waited until everything was fully operational) and from the origin of their enrollment as they decided to move in after large organizations had provided some credibility to the project.

2.5 Conclusions

The research aimed to uncover some of the events that affect the enrollment and the contribution to VGI projects. On the one hand, the scientific literature mostly studied online participants' motivation and interests through online surveys. However, surveys provide only time-specific information and usually offer no guarantee that those who

chose to reply to a survey are representative of the studied community. On the other hand, researchers have assessed participants' behaviour regarding the volume of their contributions or the nature of content they provided. However, none addressed the way people enrolled and made a first contribution or how it evolved over time. A detailed analysis of the evolution of enrollments and initial contributions to the OSM project identifies trends in the nature of events that correlated with significant changes in new participants' behaviour.

Our study showed important correlations between different types of events and the effects they may have had on the recruitment and participation of individuals in a large VGI project. Specifically, elements such as technological improvements to the infrastructure, media coverage, and recognition of the project by other communities were shown to correlate with direct increase in recruitment and participation. We also found that internal conflicts within a community can harm a project, even if it results only from a very small group of people. Furthermore, unpredictable consequences of such conflicts may affect the project on the long-term as shown by the spamming of the OSM registration process.

We finally established comparisons with the "Diffusion of innovation" theory which indicate that the profile of participants who enroll in the project change over time. As their profile changes, their behaviour is expected to change as well, which should be considered when analyzing their contributions over time. According to the distribution of the different profiles proposed by Rogers (1983), our results suggest the OSM participants

may still be issued from an “Early majority” which would predict a long life to the project before exhausting participants from following phases (Rogers 1983). Those findings can help online communities to create strategies for growing and reinforcing their membership according to the profile of the participants as the project evolve or by mitigating conflicts, all actions being oriented toward contributors being enthusiastic about their participation.

2.6. References

Ahmouda, A. and Hochmair, H.H., 2017, Using Volunteered Geographic Information to measure name changes of artificial geographical features as a result of political changes: a Libya case study. *GeoJournal,* 82 (1), 1-19.

Aknouche, L. and Shoan, G., 2013, *Motivations for Open Source Project Entrance and Continued Participation.* Thesis (Master). Lund University.

Amichai-Hamburger, Y., *et al.*, 2016, Psychological factors behind the lack of participation in online discussions. *Computers in Human Behavior,* 55, 268-277.

Arsanjani, J.J., *et al.*, 2015, An exploration of future patterns of the contributions to OpenStreetMap and development of a Contribution Index. *Transactions in GIS,* 19 (6), 896-914.

Borst, W.A.M., 2010. *Understanding Crowdsourcing—Effects of motivation and rewards on participation and performance in voluntary online activities.* 1st ed. Rotterdam (NLD): Erasmus University of Rotterdam.

Brown, J.J. and Reingen, P.H., 1987, Social ties and word-of-mouth referral behaviour. *Journal of Consumer research,* 14 (3), 350-362.

Bryant, S.L., Forte, A. and Bruckman, A., 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work,* November 6-9 Sanibel Island (USA). New York (USA): ACM, 1-10.

Budhathoki, N.R., 2010, *Participants’ motivations to contribute geographic information in an online community.* Thesis (PhD). Graduate College of the University of Illinois.

Budhathoki, N.R., Nedovic-Budic, Z. and Bruce, B., 2010, An interdisciplinary frame for understanding volunteered geographic information. *Geomatica,* 64 (1), 11-26.

Chakraborty, M., *et al.*, 2016, Recent developments in social spam detection and combating techniques: A survey. *Information Processing & Management,* 52 (6), 1053-1073.

Clary, E.G., 1998, Understanding and assessing the motivations of volunteers: A functional approach. *Journal of personality and social psychology,* 74 (6), 1516-1530.

Corcoran, P., Mooney, P. and Bertolotto, M., 2013, Analysing the growth of OpenStreetMap networks. *Spatial Statistics,* 3, 21-32.

DiBiase, D.*, et al.*, 2006. *Geographic Information Science &Technology—Body Of Knowledge.* 1st ed. Washington (USA): Association of American Geographers.

Downs, R.M. and DeSouza, A., 2006. *Learning to think spatially: GIS as a support system in the K-12 curriculum.* 1st ed. Washington (USA): The National Academies Press.

Fishbein, M. and Ajzen, I., 1975. *Belief, attitude, intention, and behaviour: An introduction to theory and research.* Reading (USA): Addison-Wesley.

Garling, C. 2012, *Google workers caught 'vandalizing' open source maps*, Wire.com (Business), January 17 [online]. Available from: https://www.wired.com/2012/01/osm-google-accusation/.

Girres, J. and Touya, G.G., 2010, Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS,* 14 (4), 435-459.

Goodchild, M.F., 2007, Citizens as sensors: the world of volunteered geography. *GeoJournal,* 69 (4), 211-221.

Google, 2011. *Google search and search engine spam* [online]. Available from: https://googleblog.blogspot.ca/2011/01/google-search-and-search-engine-spam.html.

Google, 2012. *Fighting Spam* [online]. Available from: https://www.google.ca/insidesearch/howsearchworks/fighting-spam.html.

Gyongyi, Z. and Garcia-Molina, H., 2005. Web spam taxonomy. *First international workshop on adversarial information retrieval on the web (AIRWeb 2005),* May 10-14 Chiba, Japan.

Haklay, M. and Weber, P., 2008 OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing,* 7 (4), 12-18.

Hemetsberger, A. and Pieters, R., 2003. *When consumers produce on the internet: the relationship between cognitive-affective, socially-based, and behavioral involvement of prosumers* [online]. CiteSeerX. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.9299&rep=rep1&type=pdf.

Horita, F.E.A., *et al.*, 2013. The use of volunteered geographic information (VGI) and crowdsourcing in disaster management: a systematic literature review. *Nineteenth Americas Conference on Information Systems,* August 15 − 17 Chicago, Illinois, USA.

Houle, B.B.J., 2005, A Functional Approach to Volunteerism: Do Volunteer Motives Predict Task Preference? *Basic and applied social psychology,* 27 (4), 337-344.

Hristova, D., *et al.*, 2013. The Life of the Party: Impact of Social Mapping in OpenStreetMap. *International Conference On Web And Social Media Papers,* July 8–11 Cambridge, Massachusetts, USA. Palo Alto, California, USA: The AAAI Press, 234-243.

Hyndman, R.J. and Athanasopoulos, G., 2014. *Forecasting: Principles and Practice.* 1st ed. Melbourne (AUS): OTexts.

Keßler, C. and de Groot, René Theodore Anton, 2013. Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap. In: D. Vandenbroucke, B. Bucher and J. Crompvoets, eds. Geographic Information Science at the Heart of Europe. Springer International Publishing., 21-37.

Kimura, A.H. and Kinchy, A., 2016, Citizen Science: Probing the Virtues and Contexts of Participatory Research. *Engaging Science, Technology, and Society,* 2, 331-361.

Mashhadi, A., Quattrone, G. and Capra, L., 2015. The impact of society on volunteered geographic information: The case of OpenStreetMap. In: J. Jokar Arsanjani*, et al.*, ed. OpenStreetMap in GIScience. Berlin: Heidelberg: Springer, 125-141.

McLeod, A.I., Yu, H. and Mahdi, E., 2011. Time series analysis with R. In: C.R. Rao, ed. Time Series Analysis: Methods and Applications. Oxford (GBR): Elsevier, 661-707.

Metzger, M.J., 2010, Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of communication,* 60 (3), 413-439.

Michelucci, P. and Dickinson, J.L., 2016, The power of crowds. *Science,* 351 (6268), 32-33.

Mooney, P. and Corcoran, P., 2013, Analysis of Interaction and Co-editing Patterns amongst OpenStreetMap Contributors. *Transactions in GIS,* 18 (5), 633-659.

Neis, P. and Zipf, A., 2012, Analyzing the Contributor Activity of a Volunteered Geographic Information Project—The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information,* 1 (2), 146-165.

Nov, O., 2007, What motivates wikipedians? *Communications of the ACM,* 50 (11), 60-64.

Nov, O., Arazy, O. and Anderson, D., 2011. Technology-Mediated Citizen Science Participation: A Motivational Model. *Proceeding of the Fifth International AAAI Conference on Weblogs*

and Social Media, July 17-21 Barcelona (ESP). Menlo Park (USA): The AAAI Press, 249-256.

OpenStreetMap contributors, 2014. *Main Page* [online]. OpenStreetMap Wiki. Available from: http://wiki.openstreetmap.org/wiki/Main_Page [Accessed 2017-06-19].

OpenStreetMap Foundation, 2017. *OpenStreetMap blog* [online]. Available from: https://blog.openstreetmap.org/ [Accessed 2017-04-07].

Penner, L.A., 2002, Dispositional and organizational influences on sustained volunteerism: An interactionist perspective. *Journal of Social Issues,* 58 (3), 447-467.

Poiani, T.H., *et al.*, 2016. Potential of collaborative mapping for disaster relief: A case study of OpenStreetMap in the Nepal earthquake 2015. *49th Hawaii International Conference on System Sciences (HICSS),* January 5-8 Koloa, HI, USA. IEEE, 188-197.

Preece, J., Nonnecke, B. and Andrews, D., 2004, The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior,* 20 (2), 201-223.

R Core Team, 2016. *R: A Language and Environment for Statistical Computing.* Vienna (AUT): R Core Team.

Rehrl, K.*, et al.*, 2013. A conceptual model for analyzing contribution patterns in the context of VGI. In: J.M. Krisp, ed. Progress in Location-Based Services, Lecture Notes in Geoinformation and Cartography. Berlin (DEU): Springer-Verlag, 373-388.

Riesch, H. and Potter, C., 2014, Citizen science as seen by scientists: Methodological, epistemological and ethical dimensions. *Public Understanding of Science,* 23 (1), 107-120.

Rogers, E.M., 1983. *Diffusion of Innovations.* 3rd ed. New-York (USA): The Free Press.

Ryan, R.M. and Deci, E.L., 2000, Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology,* 25 (1), 54-67.

Schneider, A., Von Krogh, G. and JäGer, P., 2013, "What's coming next?" Epistemic curiosity and lurking behaviour in online communities. *Computers in Human Behavior,* 29 (1), 293-303.

Soden, R. and Palen, L., 2014. From crowdsourced mapping to community mapping: The post-earthquake work of OpenStreetMap Haiti. *COOP 2014-Proceedings of the 11th International Conference on the Design of Cooperative Systems,* May 27-30 Nice (FRA). Springer, 311-326.

Stebbins, R.A., 2015. *Serious leisure: A perspective for our time.* 2nd ed. (USA): New Brunswick: Transaction Publishers.

Steinmann, R., *et al.*, 2013. Contribution Profiles of Voluntary Mappers in OpenStreetMap. *Online proceedings of the International Workshop on Action and Interaction in Volunteered Geographic Information (ACTIVITY) at the 16th AGILE Conference on Geographic Information Science,* May 14 Leuven (BEL).

Sun, N., Rau, P.P. and Ma, L., 2014, Understanding lurkers in online communities: A literature review. *Computers in Human Behavior,* 38, 110-117.

Taylor, S. and Todd, P.A., 1995, Understanding information technology usage: A test of competing models. *Information systems research,* 6 (2), 144-176.

Tichenor, P.J., Donohue, G.A. and Olien, C.N., 1970, Mass media flow and differential growth in knowledge. *Public opinion quarterly,* 34 (2), 159-170.

Venkatesh, V., *et al.*, 2003, User acceptance of information technology: Toward a unified view. *MIS quarterly,* 27 (3), 425-478.

Von Krogh, G., *et al.*, 2012, Carrots and rainbows: Motivation and social practice in open source software development. *MIS Quarterly,* 36 (2), 649-676.

Weait, R., 2011. *OSM Licence Upgrade—Phase 4 coming soon* [online]. Blogs.OpenStreetMap.org. Available from: https://blog.openstreetmap.org/2011/06/14/osm-license-upgrade-phase-4-coming-soon/ [Accessed 2016-05-08].

Yamak, Z., Saunier, J. and Vercouter, L., 2016. Detection of Multiple Identity Manipulation in Collaborative Projects. *Proceedings of the 25th International Conference Companion on World Wide Web,* International World Wide Web Conferences Steering Committee, 955-960.

Zook, M.A., *et al.*, 2010, Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. *World Medical & Health Policy,* 2 (2), 7-33.

Chapter 3: Contributors' Withdrawal from Online Collaborative Communities - the Case of OpenStreetMap

Abstract: Online collaborative communities are now ubiquitous. Identifying the nature of the events that drive contributors to withdraw from a project is of prime importance to ensure the sustainability of those communities. Previous studies used *ad hoc* criteria to identify withdrawn contributors, preventing comparisons between results and introducing interpretation biases. This paper compares different methods to identify withdrawn contributors, proposing a probabilistic approach. Withdrawals from the OpenStreetMap (OSM) community are investigated using time series and survival analyses. Survival analysis revealed that participants' withdrawal pattern compares with the life cycles studied in reliability engineering. For OSM contributors, this life cycle would translate into three phases: "Assessment," "Engagement" and "Detachment." Time series analysis, when compared with the different events that may have affected the motivation of OSM participants over time, showed that an internal conflict about a licence change was related to largest bursts of withdrawals in the history of the OSM project. This paper not only illustrates a formal approach to assess withdrawals from online communities, but also sheds new light on contributors' behaviour, their life cycle, and events that may affect the length of their participation in such project.

Keywords: Chebyshev's inequality; circadian cycle; time series analysis; survival analysis; life cycle; OSM history; contributors' behaviour

3.1. Introduction

With the advent of the Web 2.0, large communities have developed around online collaborative projects that allow people to contribute data. Examples include platforms that allow sharing of in situ observations (e.g., the Audubon Society for birdwatching), identification of features from images (e.g., Zooniverse), the sharing of general (e.g.,

Wikipedia) and technical knowledge (e.g., PostgreSQL), and the mapping of people's neighborhoods (e.g., OpenStreetMap). Every day, millions of people visit web sites from online communities like Wikipedia.org or OpenStreetMap.org (SimilarWeb Ltd. 2017). Researchers are increasingly referring to these communities as a valuable work force and important source of data (Kimura and Kinchy 2016, Michelucci and Dickinson 2016)

These successful communities may have hundreds of thousands of active contributors, but all do not contribute in the same way. Among those who contribute, a majority of them will only participate once (Panciera *et al.* 2009, Neis and Zipf 2012), leaving most transactions to a small group of dedicated contributors (Nielsen 2006, Ochoa and Duval 2008). Even if the proportions may slightly change between communities (Neis and Zipf 2012), this typical participation model is referred to as the 90–9–1 rule (Nielsen 2006), stating that 90% of the members of a given online community will not contribute anything, 9% will contribute sporadically, and the remaining 1% will be dedicated contributors. In this context, the withdrawal of participants who maintained their participation beyond an initial period of engagement is a significant loss for a community (Balestra *et al.* 2017).

Studies have looked at the life cycle of online contributors (Neis and Zipf 2012, Ciampaglia and Vancheri 2010, Ortega and Izquierdo-Cortazar 2009, Panciera *et al.* 2010, Zhang *et al.* 2012), but the results can be hard to compare. The use of *ad hoc* criteria to identify withdrawn contributors prevents comparisons between studies, in addition to introducing biases and interpretation errors. Most collaborative online projects have no

formal mechanism to determine who withdrew from the project. Since participants freely decide when they contribute, based on their spare time, it is then difficult to distinguish between participants who left a project from those who are waiting for some free time to contribute again.

Assessing withdrawals from online projects and identifying the nature of the events that drive contributors to leave a community is thus of prime importance. Such knowledge is required to monitor the health of an online community and to minimize contributor withdrawal, particularly when changes are to be made to the participatory environment.

In order to analyze this phenomenon, about 10 years of withdrawals from the OpenStreetMap (OSM) community were investigated. Different statistical approaches were explored to model participants' behaviour based on the history of their daily contributions. Using the history of daily contributions required first eliminating potential biases caused by the location of contributors. A probabilistic procedure was then developed to identify the contributors who left the project according to their historical behaviour. The resulting daily count of withdrawals was analyzed using both survival and time series analyses.

Survival analysis was used to model the proportion of OSM participants who were still considered active in the project after a given period of time (i.e., survival curve). The resulting model was also used to generate the "hazard curve" of OSM participants. Hazard curves are often used to characterize life cycles of different domains, such as demography

or reliability engineering, and may provide similar insight about OSM contributors.

Time series analysis was used to decompose daily withdrawals in their different components (i.e., trend, seasonal and random). Once decomposed, significant variations of resulting components were compared with the different events that dotted the OSM history to identify which ones may have affected the motivation of OSM participants over time (Bégin *et al.* 2017).

This paper describes the distribution functions used to characterize the frequency of contributions from participants and discusses the results. The origin of the bias induced when using UTC timestamps to determine the dates of the contributions is explained, and the method used to correct the dates is described. The life expectancy and the survival rates of OSM contributors are presented with the results of a time series analysis. Finally, the paper reports on the events in the OSM project that correlated with large numbers of withdrawals from the community over years.

3.2. Materials and Methods

The OpenStreetMap project was chosen because the project's history is well documented and the data are freely available. The OSM project aims to create a comprehensive map of the world built on the interests and the local knowledge of its community (Mooney and Corcoran 2012, Napolitano and Mooney 2012, Bright *et al.* 2017). The project uses a Wiki approach to enable its community to create and improve the map. With currently more than 3 million registered users (OpenStreetMap contributors 2013), it has become one of

the most successful peer-production projects of the Web and is the largest mapping project in the world. The chronicle of the project's history (e.g., technical improvements, normative changes, social activities) is maintained in the project's wiki documentation (OpenStreetMap contributors 2014b) and a record of all the contributions is made available on a regular basis through OSM history dump files (OpenStreetMap contributors 2014a). These files contain all transactions made since the first contribution and include the virtual containers (i.e., changesets) in which the edits were provided. These changesets identify the contributors who submitted changes, the temporal extent of each editing session, and a minimum bounding rectangle covering all the features edited during the session.

3.2.1. Data Retrieval

As part of a larger project, a history dump file released on 1 September 2014, was downloaded from the OSM web site to access the records of contributions made to the project since 9 April 2005 (i.e., the first edits). FME workbenches (Safe Software 2015.0) were developed to extract and load the data contained in the history dump file to a PostgreSQL (9.3) database. The resulting 2 TB database included 25 M changesets that were used in this study. Statistical analyses and visualizations presented in this paper were carried out using R software (v.3.2.1).

The frequency of contributions (i.e., the number of continuous time intervals an individual has invested in the project) cannot be determined from the number of changesets a contributor provided. The number of changesets and the time span of each of

these changesets largely depend on the OSM application interface (API) and the mapping application used by the contributor. First, the OSM API applies constraints regarding the time over which a changeset has been opened by automatically closing them either after being inactive for one hour, or after being active for 24 h. Second, OSM mapping applications have different schema for creating changesets. The same editing session may then produce various numbers of changesets, according to the application used and its configuration. However, the changesets' creation timestamps were exploited to identify on which days a contributor was active.

In order to link potential bursts of withdrawals from the community with events from the project's history, a comprehensive event repository was built by retrieving the entire history of the project from OSM Wiki pages (OpenStreetMap contributors 2014b) and some OSM mailing lists(OpenStreetMap contributors 2017b) (i.e., "talk," "dev" and "legal" mailing lists). The period covered by the repository matched the time span of the history dump file. The events were classified according to an adapted version of the Wiki page's nomenclature and OSM event classification (OpenStreetMap contributors 2017a) to include development milestones, media news and internal announcements (i.e., blogs and mailing lists).

3.2.2. Assessing the frequency of contributions

The frequency of contributions of each participant has been derived from the UTC timestamps of their changesets. UTC timestamps cannot be used directly to extract the dates of contributions as it could introduce a bias due to the contributor's geographic

location and the local time at which the contributions were usually made. The number of distinct dates extracted from the changesets can double when the local time at which the contributions are made falls around midnight GMT. In order to circumvent the problem, we needed to aggregate individuals' contributions in 24-h units that would not be affected by this temporal reference. Two approaches were compared to define a daily contribution timeframe for each individual, the first one based on the proximity of contributions, the other based on contributors' circadian behaviour.

The first approach aimed at aggregating contributions by using hierarchical clustering on the time interval (i.e., distance) between changesets. The approach was based on the fact that, when the participants have some free time to contribute, the changesets generated during their editing sessions will form clusters in time as demonstrated by Halfaker (Halfaker *et al.* 2015) for different online communities. The closer the changesets, the higher the odds the edits were made during the same editing session and consequently on the same day (from contributors' point of view). For each contributor, clusters of changesets were formed by iteratively grouping the nearest changesets using the nearest-neighbour chain algorithm (Day and Edelsbrunner 1984). The algorithm was chosen because of its relative simplicity to implement as a recursive function in PostgreSQL. When a cluster was about to extend over more than 24 h, it was removed from the process and considered as a one-day contribution. After all the contribution clusters were removed (i.e., any new cluster would span over 24 h), the inter-cluster times were rounded to one-day units to obtain the number of days spent by a contributor between each contribution.

The second approach aimed at identifying the circadian cycle of each contributor in order to apply an offset to the UTC timestamps and consequently to adjust the date of contributions. The circadian cycle partition of a contributor was defined as the time (UTC) at which a contributor was usually inactive (i.e., potentially asleep) according to the history of its contributions. The UTC offset was computed by averaging hours over the longest contiguous interval of time for which the number of contributions was at its minimum. The number of contributions was counted over 24 one-hour bins (0 h—23 h). Corresponding bins were duplicated over four hours on each side (−4 h, −3 h … 26 h, 27 h) to smooth contributions' count with a nine-hour moving average window. Once a UTC offset was obtained for each contributor, it was applied to their changesets' UTC timestamps prior to extract the distinct dates of their contributions (i.e., active days). Changesets' creation timestamps were used since only participants can trigger them while closing timestamps could result from an API operation.

Both approaches were compared and assessed using a subset of about fifty contributors at both ends of the activity spectrum. The subsets covered both new (active days <10) and accomplished (active days> 1000) contributors. The approach that provided the most reasonable estimate of contributors' active days for both subsets was used to identify the number of contributions (active days) and the number of days between these contributions. Since a reasonable estimate had to be compatible with human behaviour, the time spent by participants contributing on each active day was measured for each method. The higher the number of days an outstanding time was spent contributing (i.e., 12–24 h), the less the method was considered compatible.

3.2.3. Identifying Withdrawn Contributors

Due to the irregular nature of contributions made by volunteers on online communities, it can be hard to discriminate participants who are waiting for time to contribute again from others who simply withdrew from a project. Results from the analysis described above were used to model the frequency of contributions and identify a time threshold after which an inactive contributor should be considered as being withdrawn (i.e., has definitely left the project) with, say, a 95% probability. Three models were used to identify such threshold. The first two used a global approach based on the contributions from all the participants while the last one considered the history of contributions of individual participants.

First, the potential theoretical distribution of delays was identified based on kurtosis and skewness methods. The "descdist" procedure (from R's "fitdistrplus" package) was used to identify the distribution using a "Cullen and Frey" graph for discrete values (Cullen and Frey 1999, Delignette-Muller and Dutang 2015) with 100 bootstrap samples. The proposed distribution was examined to model the delays and identify withdrawn contributors.

Second, the 95^{th} percentile of delays between each sequential contribution was computed and plotted on a log-log graph, providing threshold values that can be used to identify withdrawn contributors. The graph was assessed on both new and accomplished contributors.

Third, since the history of contributions of each individual is available, we used the

Chebyshev inequality described in Equation (3-1) to assess the contributions of each participant and set individuals' threshold:

$$\boldsymbol{P}(|\boldsymbol{X}-\boldsymbol{u}|\geq\boldsymbol{\epsilon})\leq\frac{\boldsymbol{\sigma^2}}{\boldsymbol{\epsilon^2}} \tag{3-1}$$

On the left side of the inequality, P is the probability that the interval of time since the participant's last contribution (X) is larger or equal to a given value (ε) when compared to the average interval (u) between its contributions. The right side of the inequality shows that this probability is less or equal to the ratio of the variance of the intervals between contributions (σ^2) over the square of the value provided on the left side of the equation (ε^2).

Chebyshev's inequality was chosen because it can be applied to any arbitrary distribution, something expected in our context. However, Equation (3-1) determines the probability for both sides of the distribution while we are only interested in the upper bound (i.e., the maximum delay expected from a given contributor). Furthermore, the equation requires the population's mean and variance while we consider having only a sample of the delays a contributor will experience during its lifespan in the project, unless the contributor has already left the community. Consequently, we used a version of the one-sided Chebyshev inequality adapted to samples (user:Cardinal 2014), as described by Equation (3-2):

$$\boldsymbol{P}(\boldsymbol{X_n}-\overline{\boldsymbol{X}}\geq\boldsymbol{\epsilon s})\leq\frac{\mathbf{1}}{\mathbf{1}+\frac{\boldsymbol{n}}{\boldsymbol{n-1}}\,\boldsymbol{\epsilon^2}} \tag{3-2}$$

In order to determine that a participant has withdrawn from a project with a given probability (P), the time since its last contribution (X_n) must differ by at least a given threshold ($\boldsymbol{\epsilon s}$) from average delays ($\overline{X}$) experienced by the participant. This probability is smaller or equal to the right side of the inequality that takes into account the size of the sample, where (n) is the number of delays, ($\mathbf{s}$) is the standard deviation of the delays and ($\boldsymbol{\epsilon}$) is a constant specific to each participant. The constant is obtained from equation 3-3.

$$\boldsymbol{\epsilon} \leq \sqrt{\frac{\mathbf{1}-\boldsymbol{P}}{\boldsymbol{P}(\frac{\boldsymbol{n}}{\boldsymbol{n}-\mathbf{1}})}} \tag{3-3}$$

Equations (3-2) and (3-3) were used to determine individuals' thresholds for the time interval since their last contribution. The contributors were considered withdrawn with a 95% probability (P) when the interval between the creation of the history dump and their last contribution reached this threshold. In cases where the participants did not have enough contributions to compute delays' standard deviation (i.e., fewer than three contributions), we used the average threshold of people having made three contributions.

Finally, the subsets of participants from both ends of the activity spectrum were used again to assess the most appropriate method to identify withdrawn contributor from the distribution identified by the Cullen and Frey graph, the 95th percentile of delays, and the sample version of the one-sided Chebyshev inequality. The method was selected by comparing the proportions of contributions that happened outside the threshold established by each method using the history of contributions from our subset of participants. The nearer the proportion is to 5%, the more adequate is the method.

3.2.4. Survival Analysis

Survival analysis provides a set of methods that allow for modelling the probability that an event occurred (e.g., death, withdrawal) over a given period of time. The methods deal with two types of observations, those for which the observed event occurred, and those for which the event did not occur during the period under consideration. In cases the event did not occur within this period, the observations must be censored. Censored data (i.e., a type of missing data) are observations for which the information was measured accurately within the studied period but for which we only know that the survival span was longer than the observed period. The survival analysis is preferred to standard regression models because it adequately handles censored observations, avoiding potential bias in such analysis.

A survival analysis (Kleinbaum and Klein 2006, Therneau and Lumley 2017) was run using the R "survival" package to calculate the probability that an OSM contributor would still be active after a given time in the project. We estimated and plotted survival curves using a non-parametric estimator of the survival function (i.e., the Kaplan-Meier method). The contributors not considered as withdrawn at the end of the period covered by our study (1 September 2014) were identified as censored observations.

Kaplan-Meier estimators were computed for the entire OSM population, and then for years at which participants first contributed (i.e., strata computation). Using the resulting survival curves, we computed and plotted the instantaneous rate of withdrawal over time, also known as the hazard function. This function provides the proportion of

active contributors that are expected to withdraw from the project at a given point in time. It illustrates at which points in the life cycle of contributors the odds they withdraw from the project are higher, stable, or lower. Since the results vary on a daily basis, they were filtered using a moving average on a 30-day window.

3.2.5. Time series analysis

A time series analysis assumes the data result from a stochastic process, dividing the process into a deterministic trend, seasonal and centred random components (McLeod *et al.* 2011, Hyndman and Athanasopoulos 2014). The daily counts of withdrawn contributors were considered as resulting from such a stochastic process. Variations in the different components can show changes in the interest of the participants to contribute to the project. However, one must consider the volume of new contributors in interpreting any variations because withdrawals depend on them, particularly since most participants contribute for only a very short period of time (Panciera *et al.* 2009, Neis and Zipf 2012). Consequently, a time series of both withdrawn and new contributors were computed.

The time series were divided into their components using the R package "decompose" procedure (R Core Team 2016). The procedure first determines the trend component by using a moving average on observed data and removes it from the time series. The window used in this process is determined by the cyclical variations expected in the data (i.e., seasonal). The length of the seasonal variations was set to a year, resulting in 182 days without value on each side of the trends components. The seasonal variations were then computed by averaging resulting observations for each of the 365 time units and

the results duplicated over the whole range of observations. Finally, the centred random component is what remains after having removed both the trend and the seasonal values from observed data. An additive decomposition was chosen over a multiplicative one to limit the influence of early years of the project in the analysis. Given the small number of participants at that time, any change represented a large proportion of the population using a multiplicative decomposition, which in turn would have had a large impact on the resulting seasonal and random components later in time (Bégin *et al.* 2017).

Variations in withdrawals and the number of new contributors were compared for each component. Outstanding variations in withdrawal components that were not correlated with variations from new contributors were identified and linked to potential explanatory events found in our inventory. The number of participants who withdrew from the project was estimated by adding positive random component values over 21 days surrounding each event.

3.3. Results

We identified 464,858 distinct contributors from the 25.1 M changesets found in an OSM history dump retrieved on 1 September 2014. The dump spanned a period of 3433 days (almost 10 years), from first to last registered contributions. The 8381 changesets created by anonymous users were not used in the analyses. This option to remain anonymous was removed for new contributors in Fall 2007 and for all participants with the advent of API 0.6 in Spring 2009. Furthermore, 400-450 contributors who declined the CT/ODbL licence implemented in 2012 (Weait 2011, OpenStreetMap administrator 2016) were not

considered either since their data were removed from the database and their contributions did not appear in the dump.

Over 3570 events related to the history of the OSM project were retrieved from the OSM Wiki and from forums' threads, covering the project's history from 2005 to 2014. Events were classified into seven categories (Table 3-1).

Table 3-1 Classification of events related to the OSM project (2005-2014).

Category	**Category description**	Number
Meeting	Administrative, development and social activities.	1350
Upgrade	Infrastructure and software upgrade implementation.	135
Forum	Mailing lists announcements and OSM Foundation blog.	52
Licence	Contributor terms and ODbL[1] licence change milestones.	8
Mapping	Mapping parties/efforts, including humanitarian activities.	725
Conference	Conferences mentioning/discussing the OSM project.	369
Media	Media coverage about OSM or related topics.	939

[1] OSM switched to an Open Database Licence (ODbL) after a lengthy process that lasted almost four years.

3.3.1. Assessing the Frequency of Contributions in Days

Results from the nearest-neighbour chain algorithm estimated to 4.52 M the number of days OSM participants contributed, with an average of 9.72 days per contributor, and up to 2373 days for the most active ones. Results from the circadian cycle algorithm estimated to 5.03 M the number of days OSM contributors were active, with an average of 10.83 days per contributor, and a maximum of 2465 days for one of the contributors.

The comparison of both approaches shows that the nearest-neighbour chain algorithm generated five times more occurrences of contribution spans longer than 12 h

for a day (50,579 days) than the circadian cycle (10,875 days). This was further analyzed by comparing activities over long contribution span clusters with the UTC offsets of their contributors. The result shown that the changesets grouped under long span clusters were usually split by a period of inactivity around contributors' UTC offsets (i.e., contributors' middle of the night). Using our subset of new and accomplished contributors, we found the average daily contribution span was 58% longer for the nearest-neighbour chain algorithm in the first group and 44% longer for the second group. Similarly, the longest daily contribution span was of 24 h for the nearest-neighbour chain algorithm and of 20 h for the circadian cycle algorithm. The circadian cycle algorithm then provided results that were more compatible with expected human behaviours for both new and accomplished participants. Consequently, the circadian cycle algorithm was used to identify contributors' active days and then compute the time they waited between two consecutive active days (i.e., contributors' delays).

3.3.2. Identifying Withdrawn Contributors

The first approach used the skewness and kurtosis of contributors' delays (i.e., the Cullen and Frey graph) to suggest potential models of distributions for the delays and identify withdrawal thresholds (Figure 3-1). Results suggested a negative binomial distribution. A negative binomial distribution is the distribution of a random variable that gives the expected number of trials required prior a given number of successes (r) to happen (for instance, obtaining a given result twice when throwing dice). Since in our case the number of trials, failures, and successes are integers (days), and we are waiting for a

next contribution to happen (r = 1), the data would have a geometric distribution (i.e., a special case of the negative binomial distribution), as long as the probability remains the same over all trials. In other words, contributing on a given day could be seen as the successful result of a dice game, in which all OSM participants would use the same dice.

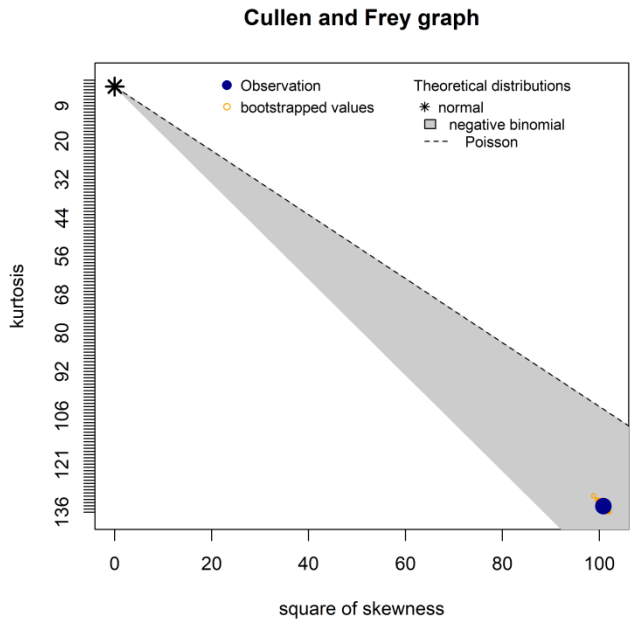


Figure 3-1 Cullen and Frey graph of delays between contributions of OSM participants with 100 bootstrap samples.

In the case of a geometric distribution, the probability of being successful (i.e., to contribute on a given day) is inversely related to the average number of trials required, which in our case is the average delay between contributions (in days). Using the 4.57 M delays experienced by those who contributed at least twice to the OSM project, we found that on average, an OSM contributor waited 19.51 days between two consecutive contributions, with the longest delay being of 3118 days (i.e., over 8.5 years).

Using the dice game analogy, OSM participants did not use the same dice since

they show a broad spectrum of frequency of contributions. Furthermore, assuming that each participant would keep playing the same game with the same number of dice all over their life span in a project is not realistic. Consequently, identifying withdrawn contributors from the above statistical model was not considered realistic either.

The second approach used the 95th percentile of the delays between each sequential contribution illustrated here in a log-log plot (Figure 3-2).

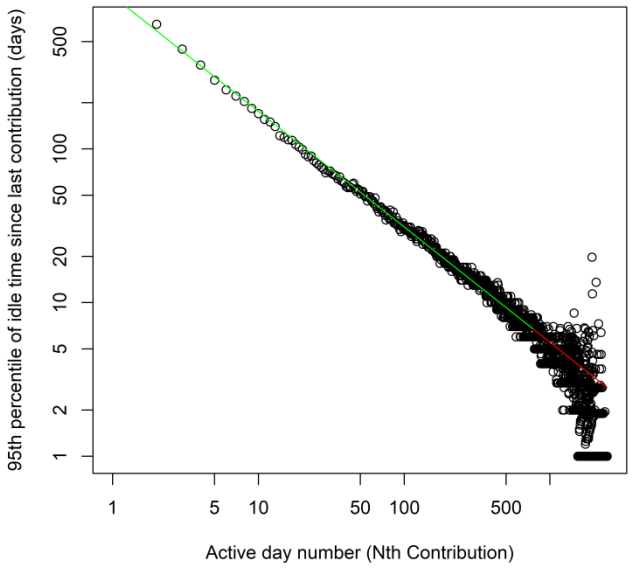


Figure 3-2 The 95th percentile of delays (days) between a Nth contribution and the previous one. An exponential model of the distribution covering 99.9% of contributors (i.e., a subset) is drawn on the log-log graph (green line). The model was extrapolated for the remaining 0.1% of contributors (red line) where delays were diverging.

The curve shows that new participants may take years before contributing again since at least 5% of them waited more than a year between one of their first four active days. It also shows that, as the number of active days gets higher, the delays between

contributions become smaller. An exponential decay model was built by fitting a linear equation on the log transform of both the percentiles and active day numbers to characterize the behaviour of 99.9% of contributors (green line). We chose to exclude from the model the percentiles derived from the remaining 0.1% of contributors since their values started to disperse unevenly after about 765 active days. These values were affecting the adjustment of the model with 69% of available measurements representing only 0.1% of contributors. The resulting equation is shown below:

$$P_{95} = e^{-0.75\log(N)+6.898}, \quad (3\text{-}4)$$

where P_{95} is the number of days after which 95% of participants will have contributed again after a previous active day, and N is the current contribution (active day). The resulting model coefficients (p <0.001) produced an adjusted R-squared of 0.986 (green line). The model was extrapolated to cover the remaining contributions (red line). However, we found that the graph tends to underestimate actual delays experienced by individual participants. For new participants, 26% experienced a delay longer than the 95th percentiles defined in above equation (3-4), while we were expecting around 5%. For accomplished contributors, this proportion rises to 74%. Since the 95th percentiles were determined from the delays of all participants (which count a few bots), those who kept contributing for a larger number of days pulled the model to shorter delays as the frequencies of their contributions were higher (as defined by the model). Interestingly, the fact that the more the participants have contributed, the less time they wait until their next contribution may suggest behaviour that is typical of an addictive process (Rozaire *et al.*

2009, Vaghefi and Lapointe 2014, OpenStreetMap contributors 2017c).

The Chebyshev inequality determined the time threshold after which a contributor should be considered as being withdrawn with a 95% probability. Since Chebyshev's inequality requires at least two observations to compute a threshold, participants having fewer than three contributions had their thresholds set to 598 days, the average threshold value of participants having three contributions. The resulting thresholds were compared to the time actually spent by the participants between each contribution. We found that 7% of new contributors experienced at least one delay longer than the estimated threshold, and 3.8% of accomplished contributors could have been identified as being withdrawn from the project more often than expected (i.e., 5% of the delays). These results are consistent with the proportion expected from the analysis and were considered appropriate to run the remaining analyses.

The Chebyshev inequality built on individuals' history has provided a better estimate of the thresholds than those obtained from statistics using the whole OSM population. Individuals' thresholds obtained from Chebyshev's inequality were then compared to the time lapse between contributors' last participation and 1 September 2014. Participants for which the time lapse was longer than their individual thresholds were considered withdrawn from the project.

3.3.3. Survival Analysis

The Kaplan-Meier estimator used to model survival rates of participants in the OSM project reveals variations in withdrawals of participants over years (Table 3-2).

Table 3-2 Withdrawals per year of first contribution. For each year, "Joined" is the number of people who made a first edit in that year, "Quit" is the number of concerned people who have withdrawn from the project so far, "Rate" is the corresponding proportion of withdrawals, and "Median" is the number of days over which at least 50% of participants contributed to the project.

Year	Joined	Quit	Rate	Median
2005	83	41	49%	3143
2006	432	218	50%	2733
2007	4820	3240	67%	1036
2008	26545	20409	77%	111
2009	61566	52044	85%	1
2010	58547	49698	85%	1
2011	65516	55917	85%	1
2012	87582	73833	84%	1
2013	86319	9278	11%	NA*
2014	73447	4220	6%	NA*
All	464857	268898	58%	28

- Participants who made a first contribution after January 2013 should not be considered since the majority of them were assigned a threshold of 598 days as they contributed fewer than three times. Consequently, their thresholds were not reached yet at the time the history dump was created.

Table 3-2 shows that half of participants who enrolled during the 2005–2007 period were still active in September 2014, while 85% of those who enrolled after 2009 withdrew from the project prior to that date. Similar turning points in participants' behaviour were found in OSM's enrollment history (Bégin *et al.* 2017) and were linked to early stages of the Diffusion of Innovation theory (Rogers 1983). After 2009; half of withdrawn participants contributed only once, as shown by the median values. Combining all the above participants, the analysis produced a survival curve that is shown in Figure 3-3. The model estimated that 64% of OSM participants "survived" their first active day, while 11% would have been active after almost 10 years (3335 days). After a steep drop of the survival rate, the slope rapidly decreases to eventually become constant.

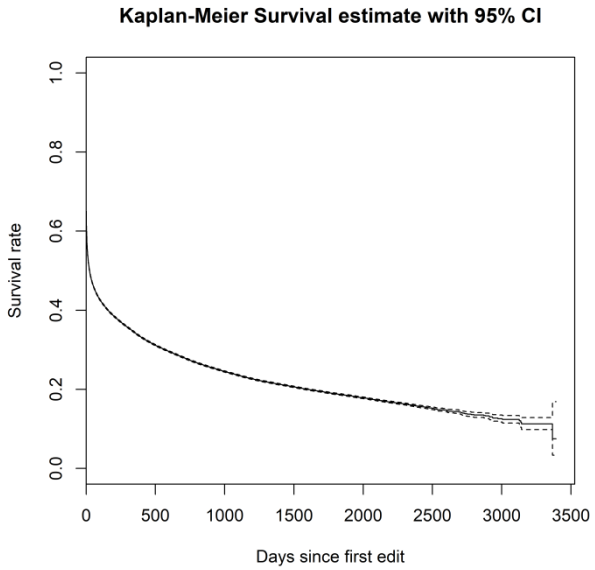


Figure 3-3 Survival curve of OSM contributors with 95% confidence intervals.

This characteristic is more easily understood from the hazard function that assesses the rate of withdrawal of participants who keep contributing to the project. The plot of the hazard function is presented in Figure 3-4. The curve shows a bathtub profile familiar to reliability engineering and system safety domains (Wang *et al.* 2002). These curves are used to characterize the rate of failure of different systems or manufactured objects and are used to split life cycles into three stages. The first stage is called “Early failures” and shows an initial steep drop in the failure rates, where weaker components rapidly fail after an item is put into service. The next stage is referred to as the “useful life” of equipment, where failure rates are low and relatively constant and result from random events. The last one is called the “wear out” stage, in which cumulative damages eventually trigger cascade failures of the components.

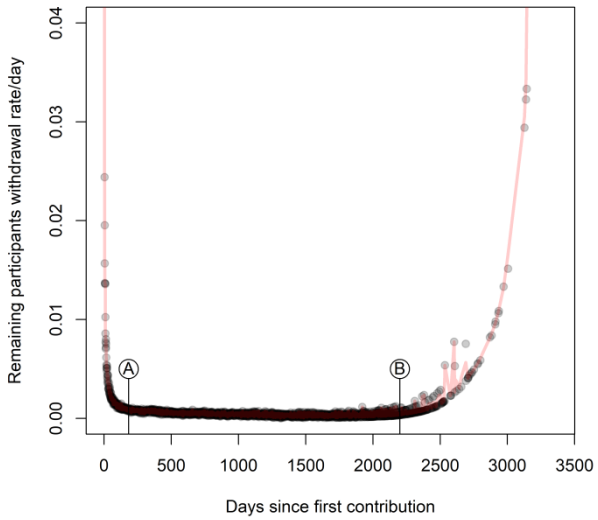


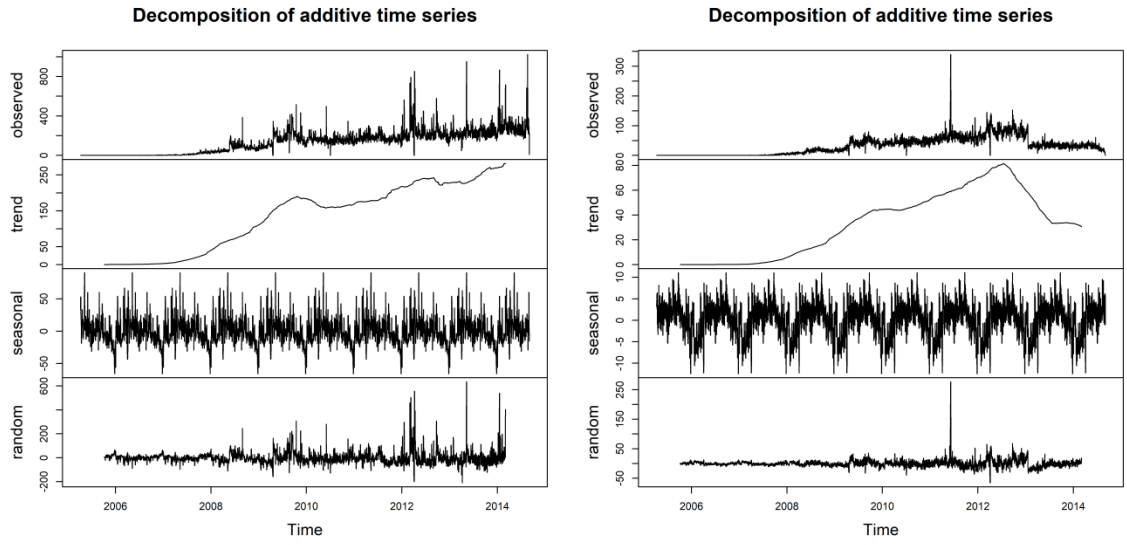
Figure 3-4 Hazard function of OSM participants, where dark dots are the proportion of remaining participants who withdrew at a given time and the red line is a moving average of the data. The first and last points of the distribution are not shown. Tags A and B delimit a segment of the curve where withdrawal rates are low and almost constant.

When using similar definitions with OSM (Figure 3-4), one can observe that the early defect rates are high with 36% of withdrawals happening on the first day (not shown on the graph). The daily rates then drop rapidly to stabilize around 0.1% after six months. By this time, about 60% of contributors will have left the project. The second stage, delimited by tags A and B (Figure 3-4), shows stabilized daily rates. These rates slightly decrease over time to reach a minimum of 0.023% (i.e., 8% on an annual basis) after 1670 active days. The rates then increase to reach 0.04% after six years (2192 days). By this time, about 80% of contributors will have left the project. The last stage sees the rates of withdrawal increasing exponentially to reach 33% (not shown on the graph). This rate results from the withdrawal of one of the three oldest participants who quit the project after having contributed over 3367 days. This last stage concerns early OSM contributors

since the span of the history dump used in this research was 3432 days and the longest individual span was 3381 days.

3.3.4. Time Series Analysis

The data used in the analysis were a continuous sequence of discrete time-ordered number of withdrawals from the OSM project, as identified previously. A first analysis was run on all OSM participants who withdrew from OSM. The variations in the number of both withdrawals and new contributors proved to be highly correlated (Spearman's rank correlation rho = 0.721 p <0.001), which means that the events that triggered a large volume of new contributors did the same for withdrawals since 36% of these new contributors withdrew on the same day. In order to reduce this correlation, the same analysis was run with participants who contributed more than once to the project. The resulting analysis presented an outstanding peak of withdrawals in mid-2011, which was not visible on results from all participants. The height of the peak affected the computation of seasonal and the random components. To remove the effect from the seasonal component, observed values were replaced by trend values over the event interval. A second analysis was run and the peak was added back on observed and random components. Figure 3-5 presents the time series of new contributors and the adjusted time series of the withdrawn contributors.



(a) **(b)**

Figure 3-5 Compared time series analysis plots for participants who contributed more than once where (a) shows the time series for new contributors and (b) shows the time series for withdrawn contributors with seasonal and random components adjusted for the peak event. Both graphs show the observed values, trend, seasonal, and random components that indicate the estimated number of contributors.

As expected, seasonal and trend variations look similar on both graphs, although the trend of withdrawals (Figure 3-5b) should not be considered after it started declining in mid-2012. This decline resulted from participants who began contributing after this date and for whom the probability of withdrawal had not yet reached 95% when the history dump file was created. Random variations show numerous peaks on both distributions. These peaks identify days when unusual volumes of participants (i.e., small or large) first contributed or withdrew from the project. These unusual volumes of withdrawals were manually identified on the graph, and potential explanations were searched from the event inventory. Outstanding variations of withdrawals that were synchronized with variations of the number of new contributors were excluded from our selection. These included all

negative peaks of withdrawals since they were all related to OSM database downtime and the events that potentially brought burst of new participants as identified by the literature (Bégin *et al.* 2017). The remaining outstanding withdrawal events are identified in Figure 3-6.

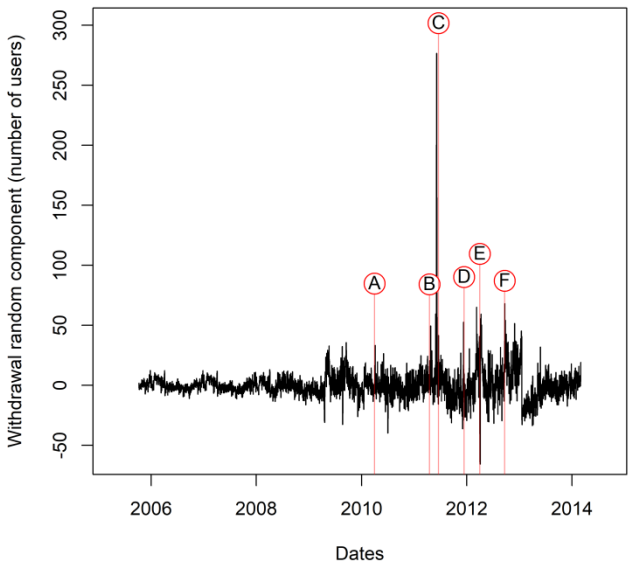


Figure 3-6 Random components of withdrawals from the OSM project and largest outstanding events (A–F). The sharp drop seen after the last event (F) is an artefact of the 598-day threshold assigned to new contributors, and the time at which the history dump file was created.

In addition to the main peak (C), five other peaks were identified in the graph. The potential explanatory events of these peaks are identified in Table 3-3. The cyclic variations visible at the left of the first event (A) are residual from the seasonal variations (Figure 3-5a seasonal) and the large withdrawals correlate with bursts of new contributors following large mapping parties after the implementation of API 0.6.

Table 3-3 Outstanding random variations of withdrawals from OSM with associated explanatory events. "Id" refers to the labels of Figure 3-6. "Quit" is the estimated number of withdrawn contributors.

Id	Date	Quit	Associated explanatory event description
A	2010-04-01	136	Ordnance Survey began releasing data for free reuse.
B	2011-04-17	255	ODbL: Unsettled users must make their choice in order to contribute.
C	2011-06-19	1117	ODbL: Users who did not agree with the new licence were blocked.
D	2011-12-13	111	ODbL: Treads about what data should be removed from the database.
E	2012-04-01	501	ODbL: Planned non-ODbL data removal and Blog. announcements
F	2012-09-20	419	Import guidelines now require dedicated accounts.

Interestingly, the first peak of withdrawals (A) seems related to the origin of the OSM project itself (Al-Bakri and Fairbairn 2011, Koukoletsos 2012). The last peak (F) could be related to participants who have imported or were to import data to the OSM database. In such a case, the volume of withdrawn contributors should correspond to those who have changed the nature of their activities at this time or before since at the same time the number of new contributors increased without any other explanation according to the event inventory.

The remaining peaks of withdrawals correlate with specific milestones or discussions about the licence change. The largest peak (C) happened in the days before the accounts of users who did not agree to the CT/ODbL licence were to be deactivated. It is important to recall that the data from these contributors were later removed from the databases and consequently do not appear in our results. These peaks could represent contributors who accepted the new licence in order not to see their work removed from the database (OpenStreetMap contributors 2017d), or subsequently lost their motivation to contribute when the process resulted in a data loss.

3.4. Discussion

The results obtained from the different analyses and procedures have not only allowed for identifying withdrawn contributors from an online community, but also suggest potential explanations about the origin of collective withdrawals from OSM. Those results have also shed some light on OSM contributors' behaviour and life cycle.

3.4.1. Assessing Withdrawals from an Online Community

According to communities' conventions about withdrawals, if any, contributors may announce their decisions to quit using templates or messages in their personal profiles[25]. However, in order for the decision to be made public, contributors must care about respecting community conventions and their decision must be taken consciously. We suspect this happens mostly on specific circumstances such as health problems, personal obligations or a conflict with the community (e.g., OSM licence change), as illustrated in some OSM users' profiles (OpenStreetMap contributors 2017d). The vast majority of contributors rather withdraw from a project by simply postponing their next contributions indefinitely because the priority they give to the activity slowly dropped (Vázquez et al., 2006), along with their motivation to contribute. This supports the need to use a statistical approach that depends only on actual contributions made by participants.

The challenge in identifying withdrawn contributors was twofold. First, using statistical models derived from the contributions of a whole population would not have permitted an analysis of individuals' behaviour. The use of Chebyshev's inequality to

[25] https://en.wikipedia.org/wiki/Template:Retired

assess the contributions of each participant has proven to provide accurate decisions about individuals' withdrawal. The main drawback of the method is that it took 798 days before confirming a one-time OSM contributor had left the project with 95% certainty, which is much shorter in most of the cases. According to Figure 3-3, about 75% of contributors have left the project at this time but the status of these one-time contributors cannot be confirmed with a 95% certainty until the threshold is reached. However, the length of this threshold for one-time contributors will vary according to the studied community and the required level of certainty. Second, in order to identify withdrawn participants based on the history of their contribution, one must identify the frequency at which they contributed to the project. We demonstrated that the UTC timestamps used to make such an assessment can lead to very different results depending on contributors' location and the time at which they usually contribute. The resulting frequency of contributions may even double in certain circumstances, something that has to our knowledge not been mentioned in the literature. Such bias could induce interpretation error when assessing contributions based on participants' locations (i.e., country, continent). Determining individuals' circadian cycle based on the UTC timestamps of their contribution proved to be a simple and efficient approach. Identifying the time at which the volume of contributions is at its minimum for each contributor better reflects individuals' natural cycles, even with fewer than 10 contributions, as we found when assessing changesets' clustering using a nearest-neighbour algorithm.

3.4.2. Withdrawals from the OSM Project

Examining the withdrawals from the OSM project over time proved to be more complex than expected, considering the relationship between withdrawal and enrollment rates. However, although the origin of long-term variations of withdrawals could not be differentiated from those of enrollment, we were able to identify specific events that correlated with collective withdrawals of participants.

The first outstanding event originated from outside the project when the original *raison d'être* of the project disappeared for many contributors after the British national mapping agency (i.e., the Ordnance Survey) began releasing data for free use. This is a risk any crowdsourcing projects can face when participants' needs can suddenly be better met through another source. In this case, a new authoritative source of free geographic data has potentially caused some local contributors to leave the project. However, considering the number of withdrawals directly related to this event, the individual needs the OSM project was meeting must have been larger for most participants, as suggested in the literature about the motivations of online participants (OpenStreetMap contributors 2017d, Vázquez *et al.* 2006, Barabási 2005, Chacon *et al.* 2007, Nov *et al.* 2011, Aknouche and Shoan 2013).

The main source of withdrawals from the OSM project was related to events that were internal to the project. The licence change process and related discussions in OSM forums may have resulted in the withdrawal of about 2000 contributors (Table 3-3) to which we must add the 400–450 contributors who declined the CT/ODbL licence

(Weait 2011, OpenStreetMap administrator 2016). Overall, 1% of OSM contributors left the project during burst of withdrawals that seemed related to this process.

If shared interests, values, and beliefs bring contributors together in a collaborative project like OSM (Aknouche and Shoan 2013, von Hippel and von Krogh 2003), it necessarily translates into a collective identity (Houle 2005) that in turn should result in collective behaviour regarding the events that pave the way to the project. The licence change may have highlighted differences in the values and beliefs of participants, resulting in the collective withdrawal of people whose values were jostled in the process (Table 3-3 and Figure 3-6). The fact that these withdrawals happened over different events simply reflects differences in the collective identity of those people (Houle 2005).

The last event identified in Table 3-3 may have shed light on the volume of participants who are concerned by data imports. When a change to the import guidelines required contributors to use dedicated accounts for import and for casual mapping, a large number of users seem to have withdrawn from the project (Table 3-3). Since this event simultaneously generated an increase in both the number of new and withdrawn contributors, the latest is probably not related to people who left the project, but rather people who considered not having the same type of contribution anymore (i.e., imports or casual mapping) and decided to leave their previous account to adjust to the new guidelines.

The withdrawals from the OSM project may reveal situations where a community is confronted to new challenges that cannot be overcome by all its participants (Balestra *et*

al. 2017). The challenges online communities face in preventing contributors from withdrawing are twofold. First, changes related to the technical aspects of the participation (e.g., new rules, technical requirements) may trigger withdrawals even when changes can be considered as being positive for the community. This is not necessarily because the learning curve could be too steep, but it might also be that some contributors are not enough motivated anymore (the wear-out stage). Second, interventions and changes that may hurt personal values or beliefs of the participants (e.g., changes in project's objectives, better alternatives, and internal conflicts) seem to have triggered large numbers of withdrawals in an otherwise strong and healthy community. In this case alternatives are limited since our results have shown that multiple collective identities can coexist in the same project, where going towards one group means moving away from another one.

3.4.3. Contributors' Behavior

As shown by Barabási (2005) and Vázquez *et al.* (2006), people contribute through bursts of rapidly occurring events separated by long periods of inactivity. The main difference between new and accomplished contributors should then be the length of their activity bursts, this length being much longer for the latter. Figure 3-2 reveals such long periods of inactivity for new contributors and the long periods of rapidly occurring contributions from accomplished ones.

When participants engage in the project, they seem to assess the project to determine whether they find it relevant, enjoyable, or both (Bégin *et al.* 2017). The contributors will consider a project as relevant if it meets their needs, desires, or

aspirations, whether because of the project's objectives (Chacon *et al.* 2007, Nov *et al.* 2011, Aknouche and Shoan 2013, von Hippel and von Krogh 2003) or because of the nature of the tasks (Houle 2005, Borst 2010, Hemetsberger and Pieters 2003). They will find a project enjoyable if their participation provides them distraction or even fun (Budhathoki *et al.* 2010, Aknouche and Shoan 2013, Nov *et al.* 2011). According to the Self-Determination Theory (Ryan and Deci 2000), an important motivation to keep contributing is self-efficacy (Davis 1989, Hemetsberger and Pieters 2003). This is the perception the individuals gain about their capacity to fulfill the required tasks as they contribute. When they are successful, individuals gain a feeling of control, competency, and autonomy that motivates them to keep contributing, while unsuccessful attempts may lead them to lose their motivation and stop contributing.

Figure 3-4 shows that this phase seems to last up to six months, when the daily rates of withdrawals fall from 35% to 0.1% when they stabilize. During this phase, about 60% of the participants will have withdrawn from the project. We would call this period the "Assessment" phase, a period over which participants are estimating the costs and benefits of contributing to the project (Nov *et al.* 2011, Aknouche and Shoan 2013). During this phase, the knowledge and skills required to contribute geographical information (DiBiase *et al.* 2006, Downs and DeSouza 2006, Jones and Weber 2012) can certainly be an obstacle for OSM contributors, which makes the project's learning curve steeper than the average collaborative project. Consequently, one would expect the rate of withdrawal to be higher with such a project than with other projects such as Wikipedia. However, the literature suggests the contrary, since about 60% of Wikipedia contributors

withdraw within the first day (Zhang *et al.* 2012, Panciera *et al.* 2009), while a similar rate was found only after six months for OSM. An explanation might be that while learning to contribute, participants are less inclined to withdraw from a project. Such behaviour may be seen in communities of practice where legitimate peripheral participation (Lave and Wenger 1991, Wenger 1998) is an important learning mechanism in which new participants slowly move from the periphery to the core of an activity. The longer it takes to grasp the nature of an activity, the longer it may take to assess the costs and benefits of engaging in such an activity. Interestingly, a similar "Assessment" phase (Figure 3-5) has been illustrated in another volunteered geographical information (VGI) project where the rates of withdrawals seemed to stabilize after about six months (Panciera *et al.* 2010).

If the project meets the needs of the participants, they seem to engage with the project for the long-term since daily rates of withdrawal stay low for a period of about six years. Given that such long-term engagement is frequent in collaborative projects (Danescu-Niculescu-Mizil *et al.* 2013, Arazy *et al.* 2017, Panciera *et al.* 2009, Zhang *et al.* 2012), we have called this period the "Engagement" phase. Over the first half of the period, the daily rates dropped from 0.1% to almost nothing (0.004%) before rising again over the second half to reach 0.04%. Referring to concepts used in reliability engineering, we consider the time at which the rates reached their minimum (i.e., 3.5 years after the first contribution) as a pivotal point where contributors seem to switch from an adaptation-dominated process to a cumulative-damage-dominated process (Wang *et al.* 2002). During the adaptation-dominated process, contributors adapt to the community's norms and rules, learn how to contribute and master available tools, and develop a feeling of self-efficacy.

During the cumulative-damage-dominated process, the many events that over years brought irritation or annoyance to the participants start affecting their motivation to keep contributing. It is a period in which contributors may become less inclined to adapt to an evolving project and a never-ending flow of unexperienced contributors. This type of behaviour (adaptation—conservatism) has already been mentioned in the literature regarding the vocabulary used by participants in online communities (Danescu-Niculescu-Mizil *et al.* 2013).

We called the last period experienced by participants, after having contributed to the project for over six years, the "Detachment" phase. Results have shown that the daily rates of withdrawal increase exponentially over this period (Figure 3-4). However, the analyses also revealed that only half of early contributors (2005–2006) withdrew from the project (Table 3-2). This special commitment to the project contrasts with withdrawals from later participants, which reached 85% after 2009. According to Budhathoki (Budhathoki 2010), a large proportion of these early contributors were also project developers or people who had an impact on its development, which could explain the discrepancy.

Another interesting finding made about contributors' behaviour is the time they spent between contributions, as the number of their contributions increases (Figure 3-2). The fact that this pattern of participation is similar to what would be expected from an addictive process might be linked to contributors' motivation. Providing geographic data to a project like OSM is a complex task (DiBiase *et al.* 2006, Downs and DeSouza 2006),

which may increase the pleasure gained by participants from fulfilling the task (learning, self-efficacy, self-actualization, self-expression), contemplating the outcome (fun, instrumentality), or using the result (meeting own need), as described by Budhathoki (Budhathoki *et al.* 2010). The more they contribute and master the process, the more pleasure they derive from it, and the higher priority they will give to the activity during their free time. The latest mechanism has even been used to explain the "bursty" nature of human behaviour when engaging in online activities (Barabási 2005). However, since the number of active days (Figure 3-2) and the time span of the project are related, some have suggested that new participants may have had fewer opportunities to contribute (lower frequency) than older participants (higher frequency) because of the OSM map saturation (Rehrl and Gröchenig 2016) in many Western countries (Neis *et al.* 2013). An analysis of the number of participants who contributed frequently (more than once a week) against their years of enrollment revealed that there was no such relationship, with the number of recurring contributors being even higher in recent years.

Finally, the rates of withdrawal have shown variations over the years, a phenomenon similar to that identified within OSM enrollment and linked to the early phases of the Diffusion of Innovation theory (Bégin *et al.* 2017, Jones and Weber 2012). This might result from a stronger engagement of early participants who developed the project, while the latest participants got involved once the project's infrastructure was mostly set up (Rogers 1983, Shepherd and Kuratko 2009).

3.5. Conclusions

Online collaborative communities have grown in importance, with millions of people visiting or consulting their web sites every day. For this reason, assessing withdrawals from online projects and identifying events that drive the contributors to leave a community is of prime importance since the proposed contents rely on those contributors.

This study compared different methods to identify the contributors who have left a community. All these methods required assessing the frequency of contributions over time but the literature had not yet assessed the biases that could result from assessing this frequency according to participants' location and schedules. We developed a method based on contributors' circadian cycles that proved to be a simple and efficient approach to avoid such biases when using UTC timestamps. Our results show that assessing the withdrawal of individual participants required estimating individual behaviour from the history of their own contributions. Accurately identifying withdrawn contributors should have provided reliable results when assessing withdrawals from the OSM community over time. Contrarily to previous studies that relied on ad hoc criteria to identify withdrawn contributors, the use of both the participants' circadian cycles and Chebyshev's inequality provides a transparent and reproducible approach when analyzing and comparing the behaviour of contributors within and between online communities.

The different procedures and analyses achieved in this research have not only illustrated an effective approach to assess withdrawals from online communities, but also

shed light on contributors' behaviour, their life cycle, and the events that may affect the length of their participation in such a project. Our results suggest the origin of withdrawals from an online community is twofold.

First, collective withdrawal can result from changes in the environment that cause participants to question their primary motivation for enrolling in a given community. These changes may lessen the need for the participants to contribute to a project, either because the need does not exist anymore or the need is better fulfilled elsewhere. Internal conflicts seem to be a major threat to the well-being of a community. Such conflicts often result from differences in values and beliefs between the members of a community, and these disagreements may be difficult to resolve collectively. Other changes that are internal to a project may also trigger withdrawals on a smaller scale in the event of a change in the community's norms and rules, contribution tools, or communication interfaces.

Second, contributors' withdrawal seems to be determined by three different phases. There is first a short "Assessment" phase, when contributors probe the project and determine if they will engage in the long term. A large of the participants will withdraw from a project during this phase. A longer "Engagement" phase follows, during which withdrawal rates are low and relatively constant. Finally, a "Detachment" phase will come when years of wear and tear have exhausted the determination of many remaining participants. However, we were not able to establish a maximum lifespan for OSM contributors since half of those who engaged in the early years of the project were still

active.

This research has highlighted very simple mechanisms that can explain most withdrawals from an online collaborative project, from both individual and collective perspectives. Understanding the processes that determine withdrawals from an online community can help with intervening and minimizing their effects. It may then be possible to minimize withdrawals by directing efforts to appropriate phases ("Assessment" or "Detachment"), or to transform the life of a project without generating conflicts, taking into account that all contributors do not have the same sensibilities, values, and beliefs.

3.6. References

Aknouche, L. and Shoan, G., 2013, *Motivations for Open Source Project Entrance and Continued Participation.* Thesis (Master). Lund University.

Al-Bakri, M. and Fairbairn, D., 2011. User generated content and formal data sources for integrating geospatial data. In: A. Ruas, ed. *Proceedings of the 25th International Cartographic Conference,* July 3-8 Paris (FRA). Paris (FRA): International Cartographic Association, 1-8.

Arazy, O., *et al.*, 2017. On the "how" and "why" of emergent role behaviours in Wikipedia. *Conference on Computer-Supported Cooperative Work and Social Computing, .*

Balestra, M., *et al.*, 2017. Investigating the Motivational Paths of Peer Production Newcomers. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* May 06-11 Denver (USA). New York (USA): ACM, 1-5.

Barabási, A., 2005, The origin of bursts and heavy tails in human dynamics. *Nature,* 435 (7039), 207-211.

Bégin, D., Devillers, R. and Roche, S., 2017, Contributors' Enrollment in Collaborative Online Communities: The Case of OpenStreetMap. *Geo-spatial Information Science,* 19 (3), 282-295.

Borst, W.A.M., 2010. *Understanding Crowdsourcing—Effects of motivation and rewards on participation and performance in voluntary online activities.* 1st ed. Rotterdam (NLD): Erasmus University of Rotterdam.

Bright, J., De Sabbata, S. and Lee, S., 2017, Geodemographic biases in crowdsourced knowledge websites: Do neighbours fill in the blanks? *GeoJournal,* [Online], 1-14.

Budhathoki, N.R., 2010, *Participants' motivations to contribute geographic information in an online community.* Thesis (PhD). Graduate College of the University of Illinois.

Budhathoki, N.R., Nedovic-Budic, Z. and Bruce, B., 2010, An interdisciplinary frame for understanding volunteered geographic information. *Geomatica,* 64 (1), 11-26.

Chacon, F., Vecina, M.L. and Davila, M.C., 2007, The Three-Stage Model of Volunteers' Duration of Service. *Social behaviour and personality,* 35 (5), 627-642.

Ciampaglia, G.L. and Vancheri, A., 2010. Empirical Analysis of User Participation in Online Communities: the Case of Wikipedia. *Proceeding or the 4th International AAAI Conference on Weblogs and Social Media,* May 23-26 Washington (USA). Menlo Park (USA): The AAAI Press, 219-222.

Cullen, A.C. and Frey, H.C., 1999. *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs.* 1st ed. New York (USA): Plenium Press.

Danescu-Niculescu-Mizil, C., *et al.*, 2013. No country for old members: User lifecycle and linguistic change in online communities. *Proceedings of the 22nd international conference on World Wide Web,* ACM, 307-318.

Davis, F.D., 1989, Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly,* 13 (3), 319-340.

Day, W.H. and Edelsbrunner, H., 1984, Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification,* 1 (1), 7-24.

Delignette-Muller,M.L. and Dutang,C., 2015. *R Fitdistrplus Package—An R package for fitting distributions.*

DiBiase, D*., et al.*, 2006. *Geographic Information Science &Technology—Body Of Knowledge.* 1st ed. Washington (USA): Association of American Geographers.

Downs, R.M. and DeSouza, A., 2006. *Learning to think spatially: GIS as a support system in the K-12 curriculum.* 1st ed. Washington (USA): The National Academies Press.

Halfaker, A., *et al.*, 2015. User session identification based on strong regularities in inter-activity time. *Proceedings of the 24th International Conference on World Wide Web,* May 18-22 Florence (ITA). New York: ACM, 410-418.

Hemetsberger, A. and Pieters, R., 2003. *When consumers produce on the internet: the relationship between cognitive-affective, socially-based, and behavioral involvement of*

prosumers [online]. CiteSeerX. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.9299&rep=rep1&type=pdf.

Houle, B.B.J., 2005, A Functional Approach to Volunteerism: Do Volunteer Motives Predict Task Preference? *Basic and applied social psychology,* 27 (4), 337-344.

Hyndman, R.J. and Athanasopoulos, G., 2014. *Forecasting: Principles and Practice.* 1st ed. Melbourne (AUS): OTexts.

Jones, C.E. and Weber, P., 2012, Towards Usability Engineering for Online Editors of Volunteered Geographic Information: A Perspective on Learnability. *Transactions in GIS,* .

Kimura, A.H. and Kinchy, A., 2016, Citizen Science: Probing the Virtues and Contexts of Participatory Research. *Engaging Science, Technology, and Society,* 2, 331-361.

Kleinbaum, D.G. and Klein, M., 2006. *Survival analysis: a self-learning text.* 2nd ed. New York (USA): Springer Science & Business Media.

Koukoletsos, T., 2012, *A Framework for Quality Evaluation of VGI linear datasets.* Thesis (PhD.). University College London.

Lave, J. and Wenger, E., 1991. *Situated learning: Legitimate peripheral participation.* 1st ed. Cambridge (GBR): Cambridge university press.

McLeod, A.I., Yu, H. and Mahdi, E., 2011. Time series analysis with R. In: C.R. Rao, ed. Time Series Analysis: Methods and Applications. Oxford (GBR): Elsevier, 661-707.

Michelucci, P. and Dickinson, J.L., 2016, The power of crowds. *Science,* 351 (6268), 32-33.

Mooney, P. and Corcoran, P., 2012. Who are the contributors to OpenStreetMap and what do they do? *Proceedings of the GIS Research UK 20th Annual Conference,* April 11-13 Lancaster (GBR). Lancaster (GBR): Lancaster University, 355-360.

Napolitano, M. and Mooney, P., 2012, MVP OSM: A Tool to identify Areas of High Quality Contributor Activity in OpenStreetMap. *The Bulletin of the Society of Cartographers,* 45 (1), 10-18.

Neis, P., Zielstra, D. and Zipf, A., 2013, Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions. *Future Internet,* 5 (2), 282-300.

Neis, P. and Zipf, A., 2012, Analyzing the Contributor Activity of a Volunteered Geographic Information Project—The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information,* 1 (2), 146-165.

Nielsen, J., 2006. *The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities* [online]. Nielsen Norman Group. Available from: http://www.useit.com/alertbox/participation_inequality.html [Accessed 2012-10-26].

Nov, O., Arazy, O. and Anderson, D., 2011. Technology-Mediated Citizen Science Participation: A Motivational Model. *Proceeding of the Fifth International AAAI Conference on Weblogs and Social Media,* July 17-21 Barcelona (ESP). Menlo Park (USA): The AAAI Press, 249-256.

Ochoa, X. and Duval, E., 2008. Quantitative analysis of user-generated content on the web. In: D. De Roure and W. Hall, eds. *Proceedings of the First International Workshop on Understanding Web Evolution (WebEvolve2008): A prerequisite for Web Science,* April 22 Beijing (CHN). 19-26.

OpenStreetMap administrator, 2016. *ODbL disagreed users Ids* [online]. planet.openstreetmap.org. Available from: http://planet.openstreetmap.org/users_agreed/users_disagreed.txt [Accessed 2016-15-23].

OpenStreetMap contributors, 2013. *Stats* [online]. OpenStreetMap Wiki, . Available from: http://wiki.openstreetmap.org/wiki/Stats [Accessed 2012-01-21].

OpenStreetMap contributors, 2014a. *Complete OSM Data History* [online]. OpenStreetMap Wiki. Available from: http://planet.openstreetmap.org/planet/full-history/ [Accessed 2014-07-03].

OpenStreetMap contributors, 2014b. *Main Page* [online]. OpenStreetMap Wiki. Available from: http://wiki.openstreetmap.org/wiki/Main_Page [Accessed 2017-06-19].

OpenStreetMap contributors, 2017a. *Events category template* [online]. OpenStreetMap Wiki. Available from: http://wiki.openstreetmap.org/wiki/Template:Cal/doc [Accessed 2017-04-07].

OpenStreetMap contributors, 2017b. *OSM mailing lists* [online]. OpenStreetMap Wiki. Available from: http://wiki.openstreetmap.org/wiki/Mailing_lists [Accessed 2017-04-07].

OpenStreetMap contributors, 2017c. *OSM purity self-test* [online]. OpenStreetMap Wiki. Available from: http://wiki.openstreetmap.org/wiki/OSM_purity_self-test [Accessed 2017-04-07].

OpenStreetMap contributors, 2017d. *User:TimSC/Quit* [online]. OpenStreetMap Wiki. Available from: http://wiki.openstreetmap.org/wiki/User:TimSC/Quit [Accessed 2017-04-07].

Ortega, F. and Izquierdo-Cortazar, D., 2009. Survival analysis in open development projects. *Proceedings of the 2009 ICSE Workshop on Emerging Trends in Free/Libre/Open Source Software Research and Development,* May 18 Vancouver (CAN). Washington (USA): IEEE Computer Society, 7-12.

Panciera, K., Halfaker, A. and Terveen, L.G., 2009. Wikipedians are born, not made: a study of power editors on Wikipedia. *Proceedings of the ACM 2009 international conference on Supporting group work,* May 10-13 Sanibel Island (USA). New York (USA): ACM, 51-60.

Panciera, K., *et al.*, 2010. Lurking? cyclopaths? : a quantitative lifecycle analysis of user behaviour in a geowiki. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* April 10-15 Atlanta (USA). New York (USA): ACM, 1917-1926.

R Core Team, 2016. *R: A Language and Environment for Statistical Computing.* Vienna (AUT): R Core Team.

Rehrl, K. and Gröchenig, S., 2016, A Framework for Data-Centric Analysis of Mapping Activity in the Context of Volunteered Geographic Information. *ISPRS International Journal of Geo-Information,* 5 (3), 37.

Rogers, E.M., 1983. *Diffusion of Innovations.* 3rd ed. New-York (USA): The Free Press.

Rozaire, C., *et al.*, 2009, Qu'est-ce que l'addiction ? *Archives de politique criminelle,* 31 (1), 9-23.

Ryan, R.M. and Deci, E.L., 2000, Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology,* 25 (1), 54-67.

Shepherd, D.A. and Kuratko, D.F., 2009, The death of an innovative project: How grief recovery enhances learning. *Business horizons,* 52 (5), 451-458.

SimilarWeb Ltd., 2017. *Analyze any Web site or App—Home page* [online]. Available from: https://www.similarweb.com/ [Accessed 2017-01-06].

Therneau,T.M. and Lumley,T., 2017. *R Survival Package—Survival Analysis.* Fermanagh (IRL): CRAN.

user:Cardinal, 2014. *Does a sample version of the one-sided Chebyshev inequality exist?*[online]. Stack Exchange network. Available from: https://stats.stackexchange.com/a/82694/82725; [Accessed 2016-09-05].

Vaghefi, I. and Lapointe, L., 2014. When too much usage is too much: Exploring the process of it addiction. *Proceedings of the 2014 47th Hawaii International Conference on System Sciences,* January 06-09 Hawaii (USA). Washington (USA): IEEE Computer Society, 4494-4503.

Vázquez, A., *et al.*, 2006, Modeling bursts and heavy tails in human dynamics. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics,* 73 (3), 1-19.

von Hippel, E. and von Krogh, G., 2003, Open source software and the "private-collective" innovation model: Issues for organization science. *Organization science,* 14 (2), 209-223.

Wang, K.S., Hsu, F. and Liu, P., 2002, Modeling the bathtub shape hazard rate function in terms of reliability. *Reliability Engineering & System Safety,* 75 (3), 397-406.

Weait, R., 2011. *OSM Licence Upgrade—Phase 4 coming soon* [online]. Blogs.OpenStreetMap.org. Available from: https://blog.openstreetmap.org/2011/06/14/osm-license-upgrade-phase-4-coming-soon/ [Accessed 2016-05-08].

Wenger, E., 1998. *Communities of practice: Learning, meaning, and identity.* 1st ed. Cambridge (GBR): Cambridge university press.

Zhang, D., Prior, K. and Levene, M., 2012. How long do Wikipedia editors keep active? In: C. Lampe and D. Cosley, eds. *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration,* August 27-29 Linz (AUT). New York (USA): ACM, 1-4.

Chapter 4: The Life Cycle of Contributors in Collaborative Online Communities - the Case of OpenStreetMap

Abstract: Over the last two decades, online communities have become ubiquitous, with millions of people accessing collaborative project websites every day. Among them, the OpenStreetMap project (OSM) has been very successful in collecting/offering volunteered geographic information (VGI). Very different behaviours are observed among OSM participants, which translate into large differences of lifespan, contribution levels (e.g. Nielsen's 90-9-1 rule) and attitudes toward innovations (e.g. Diffusion of innovation theory—DoIT). So far, the literature has defined phases in the life cycle of contributors only based on the nature of their contributions (e.g. role of participants, edits characteristics). Our study identifies the different phases of their life cycle from a temporal perspective and assesses how these phases relate to the volume and the frequency of the contributions from participants. Survival analyses were performed using both a complementary cumulative distribution function and a Kaplan-Meier estimator to plot survival and hazard curves. The analyses were broken down according to Nielsen and DoIT contributors' categories to highlight potential explanatory variables. This paper shows that two contribution processes combine with three major participation stages to form six phases in contributors' life cycle. The volume of edits provided on each active day is driven by the two contribution processes, illustrating the evolution of contributors' motivation over time. Since contributors' lifespan is a universal metric, our results may also apply to other collaborative online communities.

Keywords: OSM history; behaviour; lifespan; survival analysis

4.1. Introduction

Online communities have become ubiquitous features of today's life. Well-known communities have developed around social networking (e.g. Facebook, Instagram, LinkedIn), while others have focused on knowledge sharing projects such as Wikipedia,

birdwatching (e.g. Audubon), or mapping (e.g. OpenStreetMap). Millions of people visit those websites every day and the scientific community is increasingly referring to these communities as being both an important source of data and a valuable work force (Kimura and Kinchy 2016, Michelucci and Dickinson 2016). These collaborative projects require a continuous flow of new participants to compensate for those who withdraw after being active for a period of time.

In a previous study on volunteered geographic information (VGI) contributors (Bégin *et al.* 2017a), we observed that the rate at which people enroll in an online community depends on two main factors: interest in, and awareness of, a project. A project triggers the interest of people because of its appealing objectives (Chacon *et al.* 2007, Budhathoki *et al.* 2010, Nov *et al.* 2011), the nature of the tasks (Houle 2005, Borst 2010, Hemetsberger and Pieters 2003), or because people foresee their participation as being potentially enjoyable (Budhathoki *et al.* 2010, Nov *et al.* 2011). Awareness about a project, whether online or not, usually comes from credible acquaintances, colleagues or friends (Rogers 1983, Brown and Reingen 1987, Hemetsberger and Pieters 2003). When a project succeeds, awareness may also come from mass media, blogs or conferences. We also demonstrated that the many events that dot the history of a project have an influence on the number of people that register and contribute to it (Bégin *et al.* 2017a).

In another recent study (Bégin *et al.* 2017b), we found that the rate at which VGI contributors withdraw from an online community depends on how satisfied contributors are when contributing, and the time it takes them to get disinterested in the project.

Satisfaction arises from a project's ability to meet individuals' needs, desires and aspirations (Penner 2002, Clary 1998, Nov 2007, Budhathoki 2010). Factors such as required knowledge and skills, community norms and rules and other participants' behaviours may discourage most new contributors from pursuing their participation beyond the first few days. The same study (Bégin *et al.* 2017b) proposed three overarching stages in the life cycle of contributors adapted from the life cycle of complex systems in reliability engineering (Wang *et al.* 2002). First, an "Assessment" stage (i.e. early defects) over which a majority of participants withdraw after having estimated the costs and benefits of contributing to the project. Second, an "Engagement" stage (i.e. useful life), in which participants often contribute for years. Finally, a "Detachment" stage (i.e. wear out), being a period over which the rate of withdrawal increases exponentially. These stages are spread over two periods that are characterized by distinct contribution processes (Wang *et al.* 2002, Danescu-Niculescu-Mizil *et al.* 2013). These life-cycle stages reflect a learning-adaptation dominated process in which active contributors seek to adhere to evolving norms and tools, followed by a cumulative-damage dominated process in which contributors adopt a more conservative attitude and eventually withdraw (Bégin *et al.* 2017b).

The literature has so far characterized contributors' life cycle based solely on the nature of their contributions. This was done for instance based on contributors' motivations, knowledge and skills (Coleman *et al.* 2009), contributors' roles in the community (Cheung *et al.* 2005, Bryant *et al.* 2005, Preece and Shneiderman 2009) or the volume of their edits (Neis and Zipf 2012).

Very different levels of participation are observed between participants. This participation inequality is expressed by the 90-9-1 rule proposed by Nielsen (2006), in which 90% of the members do not contribute much data (Schneider *et al.* 2013, Sun *et al.* 2014), 9% contribute sporadically, and the remaining 1% produces most of the content (Ochoa and Duval 2008, Neis and Zielstra 2014, Ma *et al.* 2015). The history of a project has also an influence on the level of participation. Most collaborative online projects have been created over the last two decades, offering new ways of sharing information. In this context, the profile of participants might evolve over time as predicted by the Diffusion of Innovation Theory (DoIT). This theory (Rogers 1983) describes how an innovation diffuses through a population and characterizes participants according to the time at which they adopt the innovation during its diffusion process.

Unravelling the phases of contribution in an online project could help determine at what point in time the properties of contributions are likely to change (e.g. volume, content, quality). To the extent that both are related, this could shed a new light on the structure of valuable VGI contributions to GIScience. In order to better understand contributors' life cycle, its phases, and the potential relationships these phases may have with the data they provide, we analyzed both the lifespan and the contributions of the OpenStreetMap (OSM) participants. The distribution of contributors' lifespan was examined over years and survival analyses were performed to identify the different phases of their life cycle. Potential changes in the nature of their contributions were assessed for both the frequency and the volume of contributions at each phase. In addition, the analyses were broken down according to both the Nielsen (90-9-1) and DoIT categories to

understand the effects they may have on our results. Finally, we examined whether contributors' life cycle phases are defined by the time the participants spend in the project or the number of days they are actually active.

This paper analyzes the life cycle of online contributors from a temporal perspective. Section 2 describes the methods used to study contributors' lifespan and identifies both the phases of their life cycle and the nature of their contributions during each phase. Section 3 presents the evolution of contributors' lifespan over years and identifies specific events that seem to have had an impact on the project. The results from survival analyses are presented and broken down according to both DoIT and Nielsen categories. The different phases of contributors' life cycle are presented in detail and the impacts they have on the nature of the contributions are described. Finally, Section 4 discusses the evolution of contributors' lifespan and the nature of their contributions according to the different phases of their life cycle.

4.2. Materials and Methods

OpenStreetMap is a large collaborative project that aims to build a comprehensive map of the world. The OSM community uses a Wiki approach to create and improve the map by collecting the local knowledge from members (Mooney and Corcoran 2012, Napolitano and Mooney 2012, Bright *et al.* 2017). OSM has been widely studied by the GIScience community to understand key questions about VGI, both because of OSM's success and also because the documentation of the project is easily accessible (Sui *et al.* 2013, Capineri *et al.* 2016, Arsanjani *et al.* 2015).

A history dump file released by OSM on September 1, 2014, and was downloaded for the purpose of this study (OpenStreetMap contributors 2014a). The file contained all the contributions made to the OSM project. In addition to these contributions, the file included the virtual containers (i.e. changesets) in which the edits were supplied, identifying both the temporal and the geographical extents of each editing session as well as the contributors who made them. A detailed chronicle of project's history is maintained by OSM contributors (2014b) and was consulted when required. FME workbenches (Safe Software 2015.0) were developed to obtain contributors' registration timestamps from OSM website and to extract and load the 1 TB history dump file to a PostgreSQL v.9.3 database. Statistical analyses and visualizations were performed using the R software v.3.4.1.

First, the Nielsen 90-9-1 rule was used to categorize participants as being "Prolific" (i.e. having contributed 90% of the edits), "Casual" (i.e. the following 9%), and "Inactive" (i.e. remaining 1%). A cumulative sum of edits was then assigned to each contributor after having ordered them based on the volume of their respective edits, from largest to smallest. Second, DoIT categories were assigned to contributors based on the results from previous studies (Bégin *et al.* 2017a and 2017b). Contributors that registered prior to 2007 were identified as "Innovators," "Early adopters" were those who registered from 2007 to 2009 and "Early majority" was assigned to those who registered after 2009. The number of days between the history dump file creation and contributors' registration dates was computed to use the same reference system as their lifespan (i.e. days since the first contribution).

Some survival analyses require differentiation between withdrawn and active contributors. In order to make this distinction, we used a systematic method that identifies status from a statistical analysis of individuals' contributions (Bégin *et al.* 2017b). Withdrawn contributors were identified by comparing the time passed since their last contribution with the longest period of inactivity expected from their contribution history. The Chebyshev theorem was applied to the time spent between contributions estimating maximum duration with a 95% probability. Periods of inactivity were computed in days after having removed biases induced by contributors' location and time zone.

4.2.1. Contributions' span over time

A scatterplot of contributors' first and last edits was created to visualize general trends in the contribution spans over years. A "contribution span" is defined as the time interval between the first and last edits made by a contributor. The opening timestamp of the first changeset was used for the time of the first edit, and the closing timestamp of the last changeset was used as last edit. Previous studies linked variations in the number of new and withdrawn contributors (i.e. variations of lifespan) with specific events that dotted the history of the project (Bégin *et al.* 2017a, Bégin *et al.* 2017b). Consequently, a plot of contributions' span over time should help highlight the impact that specific events may have had on the life cycle of contributors. Due to the large number of contributors, we used R's "smoothScatter" procedure that plots kernel density estimates instead of actual data points (R Core Team 2016). Contrary to standard scatterplots, the density scatterplot shows the relative number of contributors represented by each point.

4.2.2. Survival analysis

Complementary cumulative distribution functions (CCDF) were used to measure the proportion of participants whose lifespan was greater than, or equal to, a given duration. This function, also called "Survival function," was used to globally assess contributors' lifespan, without discriminating between active and withdrawn participants. Inflection points on the CCDF graph were expected to show changes in contributors' engagement in the project. The analysis was run using an empirical cumulative distribution function (ecdf) from R software (2016). CCDF was obtained using equation 4-1.

$$CCDF = 1—(ecdf[x][x]), \quad (4\text{-}1)$$

where ecdf(x) generates a cumulative distribution function for x, and a call to this function for x (ecdf[x][x]), returns the percentiles of x. The complementary value is obtained from 1—(ecdf[x][x]). The distribution was plotted for the whole range of contribution spans, and different scales were used on each axis to support a manual identification of inflection points.

The Kaplan-Meier estimator was also used to model contributors' survival using the "survfit" procedure from R's "survival" package (Therneau and Lumley 2017). Participants that were considered as being active at the time of the analysis were "censored" for the procedure to consider their survival time (i.e. lifespan) as incomplete. Both survival and hazard function curves were derived from the analysis. Survival curves show the proportion of participants that are still active, while the hazard function curves show the daily rates of withdrawal. Hazard function curves are of particular interest since

they often illustrate the different phases in the life cycle of the studied phenomena (Weon 2016, Wang *et al.* 2002). Those curves were filtered using a moving average over a 30-day window and inflection points were identified manually on the resulting curves. The effects Nielsen and DoIT categories may have had on the life cycle were examined using strata analyses, breaking down the Kaplan-Meier analysis using these categories.

4.2.3. Identification of contributors' life cycle phases

The life cycle phases of OSM contributors were identified by comparing the inflection points found on the curves resulting from both survival analyses. Since our data are empirical and that no theoretical model could have located these inflection points, these points were identified manually.

Different metrics were defined to characterize each phase. First, the proportion of contributors that completed a given phase was established using Kaplan-Meier survival rates. The number of members belonging to each phase resulted from classifying all contributors based on the time they spent in the project. Estimating the number of active participants used the same process, but counted only those who were still considered as being active. The formal evaluation process used to differentiate withdrawn from active contributors may require long delays (possibly years) before confirming a contributor has left a project with 95% probability, particularly when they contributed for a short period of time, which is the case for OSM (Bégin *et al.* 2017b). For instance, while we know that 40% of contributors withdraw on the first day, it may take years before confirming they have withdrawn with a 95% probability. The proportion of active contributors was then

expected to be overestimated for early phases; to address this, an alternative estimation was used, as described in Equation 4-2.

$$ACp = Nc*Ep*Sp, \quad (4\text{-}2)$$

where p is the phase considered, ACp is the number of active contributors at the end of phase p, Nc the number of new contributors per day, Ep is the number of days since first edits at the end of phase p, and Sp is the proportion of active contributors at the end of phase p. The smallest estimates of active contributors derived from both methods (i.e. formal and alternative) were applied to each phase of the OSM project.

4.2.4. Volume and frequency of contributions

In order to assess the volume and the frequency of contributions at each project phase, the number of edits and the number of days since each user's previous contribution were registered for each active day. However, assessing contributions from the number of edits or the number of active days would introduce biases since both are correlated to contributors' lifespan and the duration of each project phase. Instead, the ratios of volume (edits per active day) over the frequency (days between edits) of edits were calculated. With this method, a specific behaviour will produce the same results, regardless of contributors' lifespan or duration of the project phases. Boxplot procedures were used to analyze the nature of contributions to reflect their long-tail distributions. Finally, considering that "Prolific" participants generate 90% of the data, their contributions were also evaluated separately from the overall population.

4.3. Results

The history dump file retrieved from the OSM website spanned over 3433 days. It contained 25.1M changesets related to 464,857 accounts, considered herein as distinct contributors. Within these changesets, 8381 had no associated contributors and were not used in the analysis. Overall, 58% of OSM contributors have withdrawn over years. The breakdown of both DoIT and Nielsen categories is presented in Table 4-1.

Table 4-1 OSM contributors according to DoIT and Nielsen categories, the number of participants (N) and their proportion (%).

DoIT Classification	N	%	Nielsen Classification [1]		N	%
Innovators	1453	0.3%	Prolific	(90% of data)	8189	1.8%
Early adopters	49866	10.7%	Casual	(9% of data)	41722	9.0%
Early majority	413538	89.0%	Inactive	(1% of data)	414946	89.2%

[1] The names provided here illustrate the nature of contributions according to Nielsen.

Regarding DoIT, the proportions are expected to evolve since other categories of participants are expected to join the project over years. For Nielsen's categories, the proportion of OSM participants determined from provided data is surprisingly similar to the expected proportions (Nielsen 2006).

4.3.1. Contributions' span over time

The scatterplot of OSM contributors' lifespan (Figure 4-1)) was used to better understand general patterns of contributors' life cycle, as well as the impact of specific events on the recruitment of withdrawal of contributors.

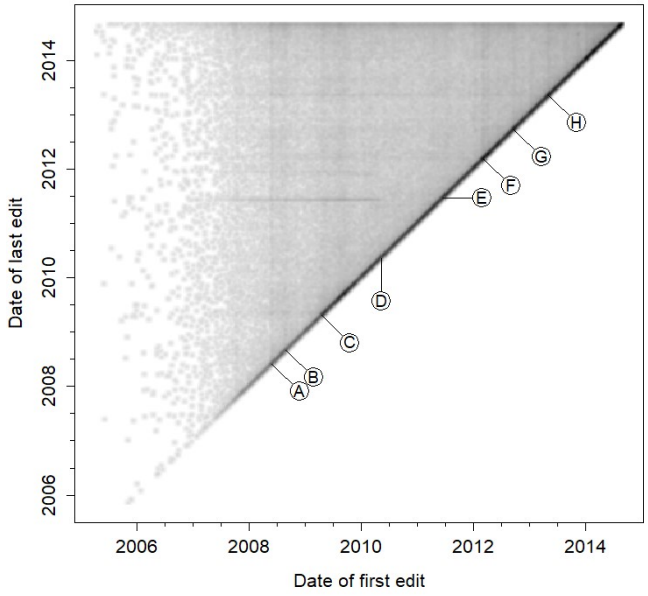


Figure 4-1 Density of OSM contributors' first and last edits over time. Each point represents about a week of contribution. The larger the number of contributors, the darker the colour is. Some noteworthy density variations are identified using labels (see Table 4-2 for details).

Figure 4-1 is characterized by a dark diagonal and variations of density over the vertical and horizontal axes. The diagonal highlights the fact that a large number of participants contributed only for a very short period of time after enrolling in the OSM project. Shading density generally increases from left to right, showing a cumulative growth in the number of OSM contributors over time. The darker line at the top of the graph represents active contributors at the time of the history dump. Vertical lines show specific peaks in OSM recruitment, followed by a gradual withdrawal of participants. Horizontal lines show bursts of participants withdrawing from the project at a specific time. The span of some of these horizontal lines indicates that only older contributors were affected by those events. Some specific density variations in the plot (i.e. labels A-H) were linked to specific events in the project's history that are listed in Table 4-2.

Table 4-2 Events related with notable variations of graph density. The "Effect" column characterizes the effects these events may have had on the number of OSM contributors as being either positive (+), negative (-) or both (*).

Id	Date	Effect	Event description
A	2008-05-30	+	The German journal *Der Spiegel* compares OSM to Wikipedia
B	2008-08-29	+	The BBC[1] quotes the president of the BCS[1] being positive about OSM
C	2009-04-21	*	API v. 0.6 brings changesets and drops anonymous edits
D	2010-05-12	+[2]	New users must now agree to the new ODbL Licence [3] to register
E	2011-06-19	-	Established users who declined the Licence are excluded from OSM
F	2012-03-08	*	ArcGIS Editor for OpenStreetMap is made available
G	2012-09-20	*	Import guidelines require dedicated accounts
H	2013-05-08	*	New ID editor is made available on the OSM website

[1] BBC: British Broadcasting Corporation; BCS: British Cartographic Society
[2] This event did not increase enrollment but has limited subsequent withdrawals after some users were excluded from OSM (E).
[3] OSM switched to an Open Database Licence (ODbL) after a lengthy process that lasted almost four years.

Figure 4-1 also shows that contributors who agreed to the new licence when joining the project (D) did not seem concerned by subsequent collective withdrawals that affected older contributors (E), which continued up to 2013 (i.e. horizontal darker lines ending on D).

4.3.2. Survival analysis

The results of the CCDF analysis are presented in Figure 4-2. Six inflection points were identified on the CCDF, bounding seven periods when contributors showed similar patterns of withdrawals (i.e. relatively constant slopes on the graphs) or changed their behaviour (i.e. slope changes).

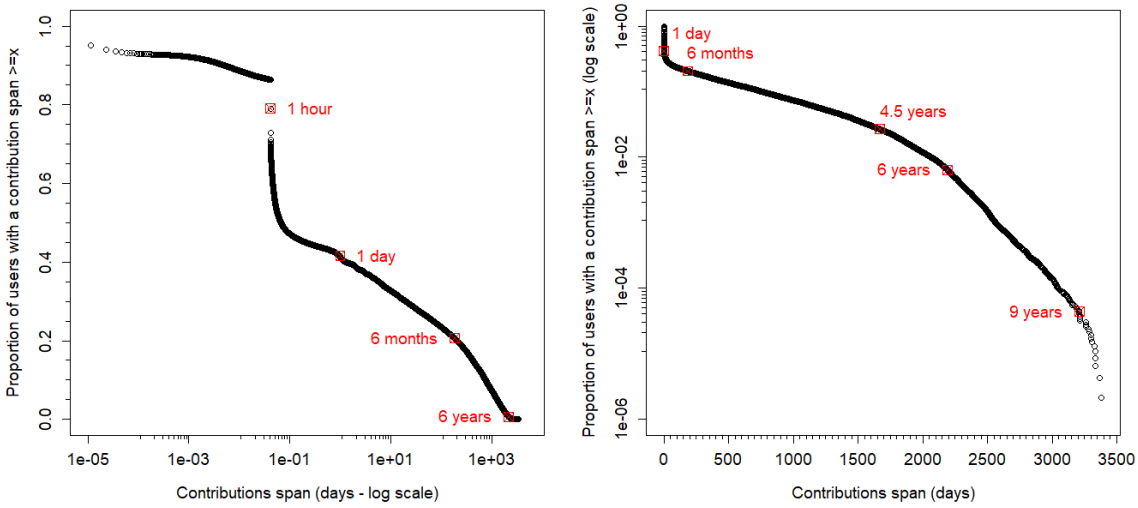


Figure 4-2 Complementary cumulative distribution function (CCDF) of contributions' span. Short contributions were enhanced by applying a logarithmic scale on the X axis (a) while a logarithmic scale on the Y axis enhanced longer contributions (b).

Figure 4-2a shows an abrupt drop of contributors starting exactly one hour after enrollment in the project, continuing for a few hours. The proportion of remaining contributors declined from 80% to 50% during this very short period. After 24 hours, the proportion of contributors that remained active in the project declined to 40% and the slope becomes constant until it reaches six months. Figure 4-2b shows a constant slope from six months to about 4.5 years, after which the slope increases until it reached six years. From this point, the slope remains relatively constant until it starts increasing again after nine years. Results from the Kaplan-Meier analysis are presented in Figure 4-3.

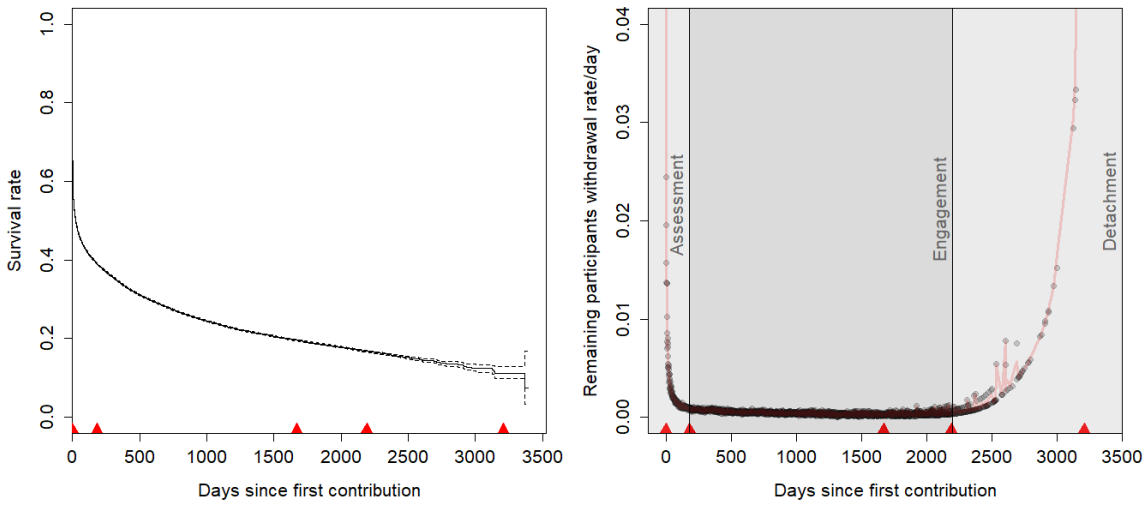
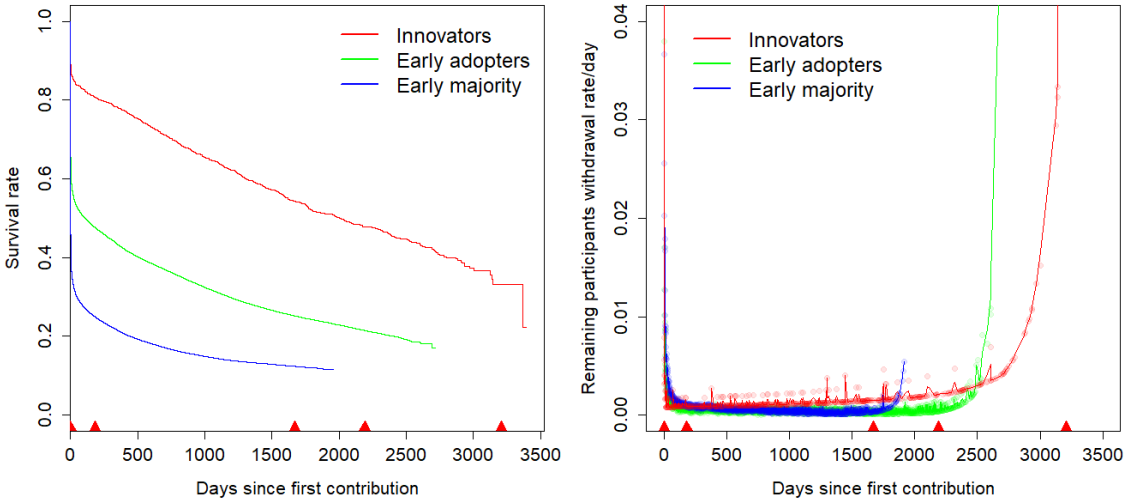


Figure 4-3 Survival curves from Kaplan-Meier estimators on the entire OSM population, with (a) Survival rates over time with confidence intervals, and (b) the daily withdrawal rates (Hazard curve) with the three stages from Bégin *et al.* (2017b). Locations of CCDF inflection points are reported on both X axes.

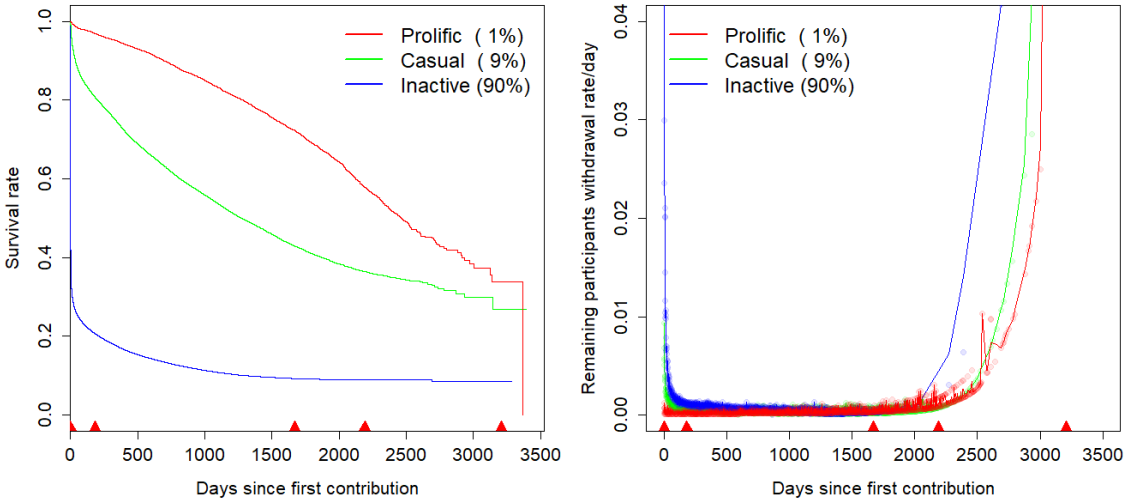
The inflection points of Figure 4-3b match those from the CCDF analysis (reported on the X axis). The hazard curve shows a bathtub shape typical of the life cycle of complex systems. The three overarching stages, described earlier, are identified on Figure 4-3b. The "Assessment" stage includes both the first (1 hour) and second (1 day) inflection points from CCDF (merged in the first symbol). The boundary between the "Assessment" and "Engagement" matches the third inflection point (6 months). During the "Engagement" stage, withdrawal rates are low and almost constant (i.e. about 15% per year), with lowest values found around the 4.5 years inflection point, where contributors switch of behavioral processes (Bégin *et al.* 2017b). The boundary between "Engagement" and "Detachment" fits the location of the next CCDF inflection point (6 years). Finally, the "Detachment" stage contains the last inflection point of the CCDF (9 years), the point

at which long-term contributors leaving the project. The Kaplan-Meier estimator was also stratified according to Nielsen (Figure 4-4) and DoIT categories (Figure 4-5).



(a) **(b)**

Figure 4-4 Survival curves from Kaplan-Meier estimators stratified by DoIT categories, where (a) illustrates the survival rates, and (b) shows daily withdrawal rates.



(a) **(b)**

Figure 4-5 Survival curves from Kaplan-Meier estimators stratified by Nielsen's categories, where (a) illustrates the survival rates, and (b) shows daily withdrawal rates.

Figure 4-4a (DoIT) and Figure 4-5a (Nielsen) show very distinct survival rates, although the shape of their curves remains similar, except for “Innovators” and “Prolific” contributors. While the survival curve of “Innovators” is relatively linear, the “Prolific” contributors are characterized by a convex survival curve rather than a concave one. The curve eventually converges toward “Casual” mappers around the last inflection point (9 years). Interestingly, the curves of both “Inactive” and “Early majority” contributors stabilize around the fourth CCDF inflection points (4.5 years). The stair case effect visible on both graphs results from the smaller number of contributors near the last inflection point.

The daily rates of withdrawal, from both DoIT (Figure 4-4b) and Nielsen (Figure 4-5b) categories, reproduce a pattern that is very similar to the one of Figure 4-3b with only slight variations between categories. The various extents of the curves from Figure 4-4b (DoIT) were expected by definition. However the truncated “Assessment” stages of “Innovators” (Figure 4-4b) and “Prolific” contributors (Figure 4-5b) were not anticipated.

4.3.3. Identification of contributors’ life cycle phases

According to the different inflection points found in both the CCDF (Figure 4-2) and the hazard curves (Figure 4-3b), six phases covering the life cycle of OSM contributors were identified. Most of the phases were identified without ambiguity since both methods agreed on their approximate location. The curves from Kaplan-Meier analysis on DoIT and Nielsen classification provided clues about the underlying structures of contributors’

life cycle. The resulting phases are presented below (Table 4-3) with the volume and the frequency of contributions made at each phase.

Table 4-3 Detailed description of the phases of the life cycle of the OSM contributors. Phases' name aims at characterizing contributors' lifespan and/or behaviour. Definitions of each column are provided as footnotes.

Phase	End	Span	Rate	Members	Active	PAC	MV	MF
Visitors	1 day	1	65%	263,848	281 [1]	<1%	4	NA
Explorers	6 months	182	39%	105,262	20,088 [1]	25%	30	4
Adopters	4.5 years	1487	20%	83,357	48,604	61%	56	77
Veterans	6 years	520	17%	9411	8338	10%	125	47
Elders	9 years	1019	11%	2957	2751	3%	189	20
Founders	NA	NA	NA	22	21	<1%	242	14

End: Estimated termination of the phase since contributors' first edits.
Span: Duration of the phase (days).
Rate: Survival rate at the end of the phase according to Figure 4-3a.
Members: Number of contributors belonging to the phase according to their lifespan.
Active: Number of active members at the time of the history dump.
PAC: Proportion of all active members belonging to the phase.
MV: Median volume of edits over members' whole lifespan (edits per active day).
MF: Median frequency of edits over members' whole lifespan (days between contributions).
[1] Value adjusted for withdrawal uncertainty over first 591 days (see explanations in the text).

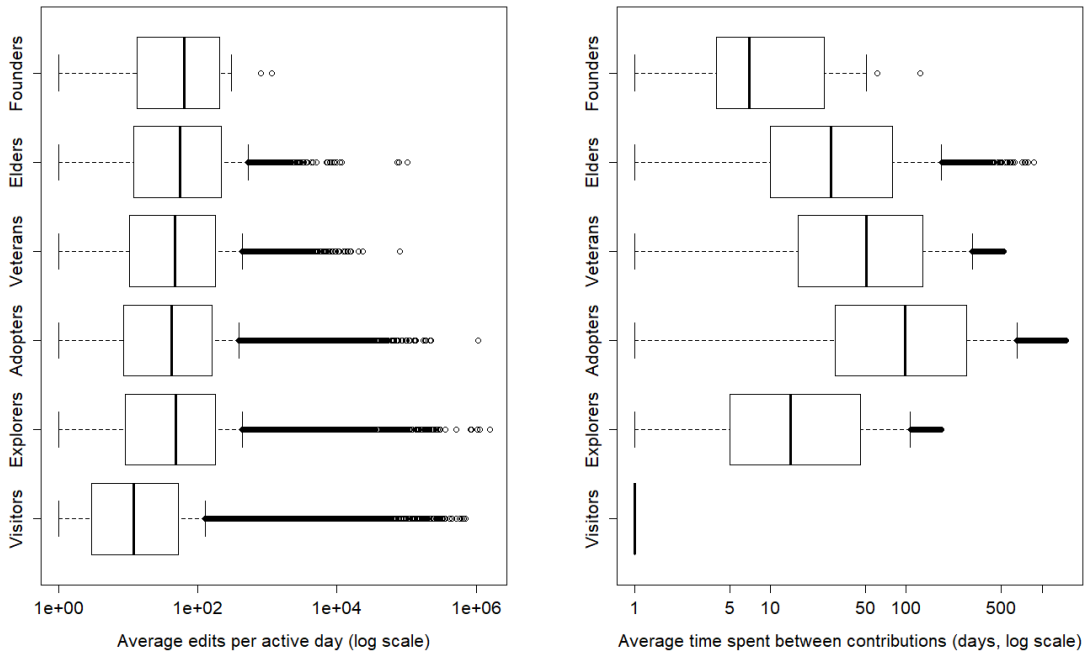
The first phase (Visitors) results from combining the first two segments from the CCDF analysis (Figure 4-2a). The lifespan of OSM participants was measured from the changesets they provided, which in turn depends on the OSM application programming interface (API). The OSM API applies constraints regarding the time over which a changeset has been opened, by automatically closing it either after being inactive for one hour, or after being active for 24 hours. Since the first inflection point of Figure 4-2a is found at exactly one hour, it is most probably a consequence from API operations, and the point was excluded from the analysis.

Since the boundaries of each phase were determined manually, their locations are

approximate, particularly regarding the later phases. In these cases, contributor withdrawal could occur several months before or after the observed dates without significantly changing our results. In Table 4-3, the number of active participants at each phase (Active) is derived from the sum of participants and the number of those who withdrew. These withdrawn participants were identified with a 95% probability by using a time threshold since their last contribution. As a result, one-time contributors were considered active until they reached 591 days without contributing, even if 70% of them will have withdrawn at that time (Figure 4-3a). Consequently, the numbers of active "Visitors," "Explorers" and "adopters" were potentially overestimated since their phases extend beyond that threshold. Equation (4-2) was then considered to provide more realistic estimations of active participants in these categories, using the latest trend in new contributors' enrollment (281 people/day). "Visitors" and "Explorers" phases were found to be overestimated using this evaluation and their values were replaced. The proportion of active contributors (PAC) refers to the sum of all active contributors at the time of the history dump.

4.3.4. Volume and frequency of contributions

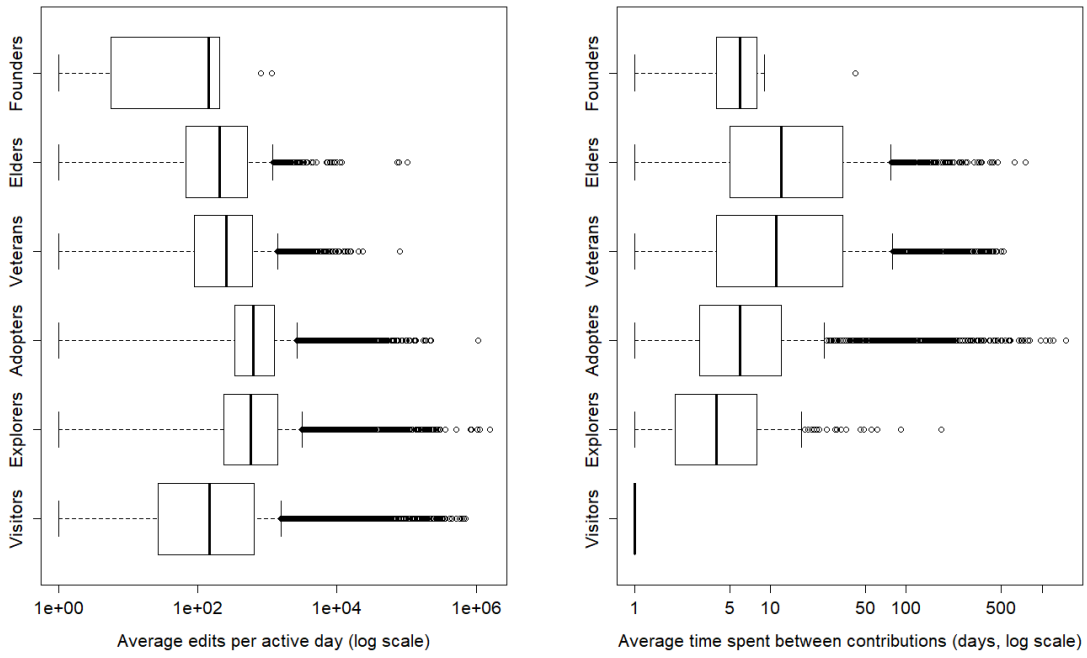
A boxplot analysis looked at the contributions made by participants at each phase of their lifespan in the project. The contributions of each participant were then distributed over each corresponding phase. The first analysis looked at the contributions made by all OSM participants and the results are presented in Figure 4-6.



(a) **(b)**

Figure 4-6 Volume and frequency of contributions made by all OSM participants according to the phase they belong to at the time of contributions, where (a) shows the volume of edits (edits per active day) and (b) the frequency of edits (time spent between active days).

Figure 4-6a shows that the average number of edits per active day is relatively constant over all phases (approximately 52 edits), with the exception of the first day (Visitors; characterized by 12 edits). The range of outlier values decreased over time. Figure 4-6b shows that the average time spent between contributions increased up to the "adopters" phase, before decreasing in later phases. A detailed analysis has shown that the maximum daily rate of edits occurred after about twenty active days, regardless of the contributor category. The same analyses were conducted on a subset made up of "Prolific" contributors (Figure 4-7).



(a) **(b)**

Figure 4-7 Contributions made by "Prolific" participants according to the phase they belong to at the time of contributions, where (a) shows the volume of edits (edits per active day) and (b) indicates the frequency of edits (time spent between active days).

The volume of edits per active day (Figure 4-7a) increased and reached its maximum at the "adopters" phase before decreasing in the following phases. As expected from "Prolific" contributors, the median values were much higher than for other contributors (Figure 4-6a), but the spans of outliers were similar. Figure 4-7b shows that the frequency of edits generally decreased (i.e. the time spent between contributions increases), except over the last phase ("Founders"). The spans of outliers were much larger than in Figure 4-6b, particularly over the "adopters" phase.

In order to better understand the nature of contributions from the participants, the results were also broken down for each phase using Nielsen's categories (Table 4-4).

Table 4- 4 Contributions from participants at each phase of their life cycle. The volume of edits (average edits per active day) and the frequency of edits (average time spent between edits) are broken down using Nielsen's classification.

	Volume of edits			Frequency of edits (days[1])		
Phase	Prolific	Casual	Inactive	Prolific	Casual	Inactive
Visitors	149	140	9	1	1	1
Explorers	569	233	20	4	10	18
Adopters	630	136	14	6	42	188
Veterans	259	45	11	11	57	110
Elders	206	38	10	12	36	72
Founders	144	51	12	6	8	24

[1] Frequency of edits is expressed as the number of days spent between edits (active days) where the larger the number is, the less often the participants contributed.

4.4. Discussion

Earlier studies have proposed different classifications to describe phases in the life cycle of online contributors, based on contributors' motivations, knowledge, and skills (Coleman *et al.* 2009) or their roles in the community (Cheung *et al.* 2005, Bryant *et al.* 2005, Preece and Shneiderman 2009). However, none of those classifications clearly linked contributors' behaviours to the time they spent in a project or the number of days they actually contributed. In a previous study (Bégin *et al.* 2017b), we suggested that the life cycle of OSM contributors exhibited three important stages. In this study, we confirmed these three stages and further subdivide them into six distinct phases, providing the first detailed analysis of temporal patterns in OSM contributors' lifespan. Stages and phases of OSM contributors' life cycle are summarized in Figure 4-8.

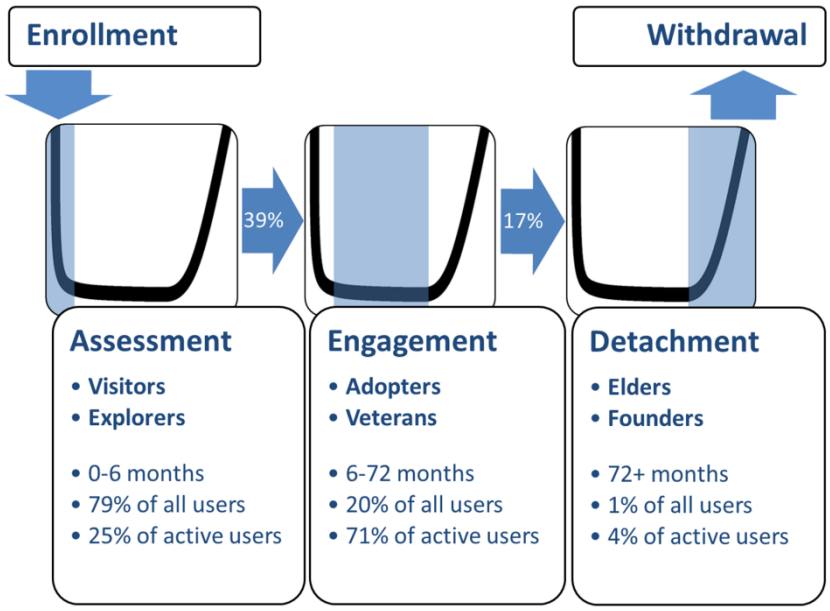


Figure 4-8 The life cycle of OSM contributors, from enrollment to withdrawal. The three stages of Figure 4-3b are presented with associated phases, stage duration and the proportion of contributors associated to the stage. Arrows show the proportion of contributors reaching the next stage.

4.4.1. Phases description

In this study, two distinct analyses have corroborated the different phases of OSM contributors' life cycle. The results show strong evidence that contributors' lifespan follows six distinct phases that reflect the evolution of contributors' motivation and interests in the project. However, some results also suggest that phases' boundaries may continue to evolve over the years.

Metaphorically, the "Visitors" phase could be seen as dipping one's toe in a project (Preece and Shneiderman 2009), a first impression that drives about two third of participants to come back to the project. However, using DoIT categories as temporal stratification, we see that this proportion ranged from 90% for "Innovators" to 46% with

the “Early majority” (Figure 4-4a). As expected from DoIT, “Innovators” enrolled over a period of time when contributing to the project was much more complex than it is for current “Early majority” participants. We previously suggested that while learning to contribute, participants may be less inclined to withdraw from the project (Bégin *et al.* 2017b). This has been seen in communities of practice (Lave and Wenger 1991, Wenger 1998) where new participants slowly move from the periphery to the core of an activity by learning from others. The improvements made to APIs and editing applications over years may have enabled many “Early majority” participants to be autonomous and to reach the core of the activity much faster than previously possible. Consequently, this retention mechanism may no longer apply, resulting in an immediate decision to disengage from the project when expectations are not met.

During the “Explorers” phase, participants assess the fragile equilibrium between engagement and withdrawal (Nov *et al.* 2011, Aknouche and Shoan 2013), balancing the costs (e.g. time invested, learning efforts) with benefits (e.g., pleasure, outcome utility) of contributing to the project. This phase is crucial in determining contributors’ engagement in a project since most of those who go through the phase will stay in the project for years (this forms the “Assessment”/“Engagement” boundary).

For most participants who withdraw at the “Explorers” phase, the decision to quit the project is made over the first few weeks even though the rate of withdrawals stabilizes only after six months (Figure 4-3b). Looking at DoIT categories (Figure 4-4b) we found that the duration of the “Explorers” phase has changed over time. During the early years of

OSM (participants considered “Innovators”) this phase barely existed, while its duration increases for “Early adopters” (2007-2009) and for “Early majority” participants (2009-). Several reasons can be invoked to explain the apparent absence of “Explorers” and “Adopters” phases for “Innovators.” Among others, a recent study shows that early OSM participants (“Innovators”) took on average two years to contribute after having registered to OSM project and suggests they may have experimented their “Assessment” stage otherwise (Bégin *et al.* 2017a, Figure 5). This long delay may also have excluded *de facto* most of those whose life expectancy was shorter according to our results. Globally, the proportion of withdrawal at the end of the phase is about 60%, which includes only 20% of “Innovators” but almost 75% of “Early majority” participants.

The next two phases (i.e. “Adopters” and “Veterans”) capture the long-term “Engagement” of OSM contributors (Figure 4-8). Over almost six years, the daily rates of withdrawal remained low, with less than 17% of remaining contributors quitting the project each year. The boundary between both phases is illustrated in Figure 4-3b where the slope of the curve switches from negative to positive after 4.5 years, as the contribution processes change. Again, phase duration could lengthen over time as illustrated by Figure 4-4b.

The “Elders” phase coincides with the beginning of the “Detachment” stage when the odds that a contributor will withdraw from the project increases exponentially. The upper limit of the phase is expected to increase in the future, as the maximum lifespan of OSM participants has not been reached yet (i.e. Figure 4-3a does not end at 0%).

Finally, the "Founders" phase seems to be an artefact of the recent history of the project and concerns only about twenty contributors. The small appendix at the end of the CCDF curve (Figure 4-2b) may not show the last "survivors" of the project but rather its initiators: people that have a special attachment to the project. As they eventually withdraw from the project, we expect the corresponding segment to disappear from the CCDF curve and from OSM contributors' life cycle phases.

In summary, our results suggest that the different phases of the life cycle apply regardless of the volume of edits the contributors provide (Nielsen's classification) or the phase of the diffusion of the project at the time they enroll (DoIT). The next DoIT category of participants to enter the project should not impact the phases except for their duration. According to DoIT, the next type of contributors that should be interested in the project is the "Late majority." The different personality traits proposed by Rogers (1983) describe "Late majority" participants as conservative people that believe far more in tradition than in progress (Moore 2001, p. 34). In a context were OSM may not be considered as a conventional map provider yet, it suggests that this "Late majority" should not constitute the mainstream of OSM contributors yet.

4.4.2. Nature of contributions over time

The assessment of the contributions made at each stage of contributors' life cycle revealed some interesting findings. Figure 4-6a displays an apparent stability of the volume of edits over each phase which we suggest is a result of a complex combination of edit rates and proportion of contributors from Nielsen's categories at each phase. The

same phenomenon has affected the apparent frequency of contributions presented in Figure 4-6b. The actual variations of volume and frequency show no such stability over the different phases (Table 4).

In this context, the profile of contributions from "Prolific" participants is of particular interest since they provided 90% of OSM data. The average volume of edits (Figure 4-7a) seems to follow the contribution processes described earlier. Volume increases over the first three phases (i.e. over learning-adaptation process), before declining over the last three (i.e. during cumulative-damage process). Such dichotomous behaviour has also been observed in other online communities (Danescu-Niculescu-Mizil *et al.* 2013).

The volume of edits provided over the first phases is characterized by a large participation inequality (outliers). For instance, while most contributors provided a few edits and withdraw over the "Visitors" phase, some participants provided hundreds of thousands of edits on that same day. Similar inequality also applies to the "Explorers" and "adopters" phase. The profiles of the hundred most "prolific" OSM contributors indicate that approximately half of them were dedicated import or bot accounts. However, the participants from the remaining half remained active for longer periods of time and often show mixed content (i.e. imports, personal edits, GPS tracks) at least until 2012, when OSM guidelines on import operations were updated.

The frequency of contributions from "Prolific" participants was expected to match

trends observed in the volume of edits. However the analysis revealed that frequency dropped over time (Figure 4-7b). The other Nielsen categories (Table 4) show that "Casual" and "Inactive" mappers increase the frequency of their contributions over latest phases. Such behaviour is counter-intuitive considering it happens over the "Detachment" stage. A potential explanation is that many of these "Casual"/"Inactive" participants, who did not withdraw after so many years, may not have had the opportunity to contribute at will throughout their lifespan. They may also have changed their objectives over time, bringing new motivations to contribute. The convex/concave shape of survival curves (Figure 4-5a) may be related to this "incomplete" experience, further distinguishing "Prolific" participants from other contributors.

4.4.3. History of contributions to OSM at a glance

Survival analyses and the analysis of contributions at each phase provided an in-depth understanding of participants' life cycle. Our first analysis (Figure 4-1) proved to be a simple yet powerful approach to better understand both contributors' lifespan over years, and the history of a project. Trends identified by the literature regarding the behaviour of contributors in online communities are revealed by this simple graph (Figure 4-1).

The absence of diagonal patterns in the upper left of the graph illustrates findings from our survival analyses. First, it shows that the life cycle of contributors is not affected by sudden changes after initial withdrawals, which would have created diagonal fading of density toward the upper left corner. Second, it demonstrates that the maximum lifespan of contributors has not been reached yet since there is no definite blank triangle on the top

left corner.

The slight variations in density over horizontal and vertical axes tell intimate stories about the project and its participants. These patterns reveal how people are brought to the project following media coverage and illustrate how conflicts between participants' personal values and beliefs can be expressed by collective withdrawals. These conflicts can be openly shared with the community as with the ODbL licence change, for example. Opponents took a public stance and drew a significant number of contributors to their arguments. When actions are taken in relation to the conflict (e.g. blocking opponents, implementing a disputed solution), a significant number of supporters may withdraw from the project in response, even long after the opponents have left the project. Such conflicts may not be public, nevertheless resulting in similar collective withdrawals. For instance, when Esri, a prominent player of the GIS industry, proposed an interface to the project, contributor withdrawal patterns suggest that a portion of the community was offended or reduced their interest to participate in the project. Although this may not be surprising in an environment dominated by free and open-source software (FOSS) enthusiasts (Budhathoki and Haythornthwaite 2013, Elwood *et al.* 2012, Perkins 2011), such an assumption cannot be confirmed based on the available data.

Finally, the graph also reveals an unexpected relationship between withdrawals and application improvements. Among other reasons, participants may have withdrawn from the project when the perceived cost of contributing exceeded the derived benefits. These changes may have appeared as too difficult to cope with, particularly for

contributors who were in the cumulative-damage-dominated phases (i.e. “Veterans” and “Elders”). However, since the graph interpretation is purely qualitative, further analyses are needed to confirm these relationships.

4.5. Conclusions

This paper has shown that two contribution processes combined with three major participation stages among contributors resulted in six phases of the participant life cycle. It has revealed that the volume of edits provided on each active day is driven by two contribution processes illustrating the evolution of contributors’ motivation over time. Surprising increases of the frequency of contributions among non-prolific contributors over the later phases of their life cycle will require further study.

Analyses confirmed that the phases of the life cycle are influenced by contributors’ lifetime, not the number of active days they experienced as considered at the beginning of the study. Although it could not be verified within the scope of this study, the number of active days may have an influence on contributors’ behaviour following project adoption.

We have seen that the proportion of experienced contributors currently represents 75% of all active participants and is expected to continue increasing until their maximum lifespan is reached. Experienced contributors have a deeper knowledge of OSM features and more skills regarding data capture. As the proportion of experienced contributors increases, we can expect the diversity and quality of the data in OSM to increase as well.

The temporal approaches described in this paper offer novel methods for

determining the phases of contributors' life cycle and shed a new light on the nature of VGI contributions that are increasingly valuable to GIScience. Furthermore, since the lifespan of contributors is an objective metric that might be available in most projects, and considering that phases' definition was not linked to contributors' motivations (subjective) or the nature of their contributions (subjective and/or project specific), these phases may apply to a broad range of collaborative online communities.

4.6. References

Aknouche, L. and Shoan, G., 2013, *Motivations for Open Source Project Entrance and Continued Participation.* Thesis (Master). Lund University.

Arsanjani, J.J*., et al.*, 2015. *OpenStreetMap in GIScience.* Cham (CHE): Springer.

Bégin, D., Devillers, R. and Roche, S., 2017a, Contributors' Enrollment in Collaborative Online Communities: The Case of OpenStreetMap. *Geo-spatial Information Science,* 19 (3), 282-295.

Bégin, D., Devillers, R. and Roche, S., 2017b, Contributors' Withdrawal from Online Collaborative Communities, the Case of OpenStreetMap. *ISPRS International Journal of Geo-Information,* 6 (11), 340.1-340.20.

Borst, W.A.M., 2010. *Understanding Crowdsourcing—Effects of motivation and rewards on participation and performance in voluntary online activities.* 1st ed. Rotterdam (NLD): Erasmus University of Rotterdam.

Bright, J., De Sabbata, S. and Lee, S., 2017, Geodemographic biases in crowdsourced knowledge websites: Do neighbours fill in the blanks? *GeoJournal,* [Online], 1-14.

Brown, J.J. and Reingen, P.H., 1987, Social ties and word-of-mouth referral behaviour. *Journal of Consumer research,* 14 (3), 350-362.

Bryant, S.L., Forte, A. and Bruckman, A., 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work,* November 6-9 Sanibel Island (USA). New York (USA): ACM, 1-10.

Budhathoki, N.R., 2010, *Participants' motivations to contribute geographic information in an online community.* Thesis (PhD). Graduate College of the University of Illinois.

Budhathoki, N.R. and Haythornthwaite, C., 2013, Motivation for Open Collaboration Crowd and Community Models and the Case of OpenStreetMap. *American Behavioral Scientist,* 57 (5), 548-575.

Budhathoki, N.R., Nedovic-Budic, Z. and Bruce, B., 2010, An interdisciplinary frame for understanding volunteered geographic information. *Geomatica,* 64 (1), 11-26.

Capineri, C.*, et al.*, 2016. *European handbook of crowdsourced geographic information.* 1st ed. London (GBR): Ubiquity Press.

Chacon, F., Vecina, M.L. and Davila, M.C., 2007, The Three-Stage Model of Volunteers' Duration of Service. *Social behaviour and personality,* 35 (5), 627-642.

Cheung, K.S., *et al.*, 2005, The development of successful on-line communities. *International Journal of the Computer, the Internet and Management,* 13 (1), 71-89.

Clary, E.G., 1998, Understanding and assessing the motivations of volunteers: A functional approach. *Journal of personality and social psychology,* 74 (6), 1516-1530.

Coleman, D.J., Georgiadou, Y. and Labonté, J., 2009, Volunteered geographic information: The nature and motivation of producers. *International Journal of Spatial Data Infrastructures Research,* 4, 332-358.

Danescu-Niculescu-Mizil, C., *et al.*, 2013. No country for old members: User lifecycle and linguistic change in online communities. *Proceedings of the 22nd international conference on World Wide Web,* ACM, 307-318.

Elwood, S., Goodchild, M.F. and Sui, D.Z., 2012, Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers,* 102 (3), 571-590.

Hemetsberger, A. and Pieters, R., 2003. *When consumers produce on the internet: the relationship between cognitive-affective, socially-based, and behavioral involvement of prosumers* [online]. CiteSeerX. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.9299&rep=rep1&type=pdf.

Houle, B.B.J., 2005, A Functional Approach to Volunteerism: Do Volunteer Motives Predict Task Preference? *Basic and applied social psychology,* 27 (4), 337-344.

Kimura, A.H. and Kinchy, A., 2016, Citizen Science: Probing the Virtues and Contexts of Participatory Research. *Engaging Science, Technology, and Society,* 2, 331-361.

Lave, J. and Wenger, E., 1991. *Situated learning: Legitimate peripheral participation.* 1st ed. Cambridge (GBR): Cambridge university press.

Ma, D., Sandberg, M. and Jiang, B., 2015, Characterizing the Heterogeneity of the OpenStreetMap Data and Community. *ISPRS International Journal of Geo-Information,* 4, 535-550.

Michelucci, P. and Dickinson, J.L., 2016, The power of crowds. *Science,* 351 (6268), 32-33.

Mooney, P. and Corcoran, P., 2012. Who are the contributors to OpenStreetMap and what do they do? *Proceedings of the GIS Research UK 20th Annual Conference,* April 11-13 Lancaster (GBR). Lancaster (GBR): Lancaster University, 355-360.

Moore, G.A., 2001. *Crossing the chasm—marketing and selling high-tech products to mainstream customers.* Revised ed. New York (USA): HarperCollins.

Napolitano, M. and Mooney, P., 2012, MVP OSM: A Tool to identify Areas of High Quality Contributor Activity in OpenStreetMap. *The Bulletin of the Society of Cartographers,* 45 (1), 10-18.

Neis, P. and Zielstra, D., 2014, Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap. *Future Internet,* 6 (1), 76-106.

Neis, P. and Zipf, A., 2012, Analyzing the Contributor Activity of a Volunteered Geographic Information Project—The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information,* 1 (2), 146-165.

Nielsen, J., 2006. *The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities* [online]. Nielsen Norman Group. Available from: http://www.useit.com/alertbox/participation_inequality.html [Accessed 2012-10-26].

Nov, O., 2007, What motivates wikipedians? *Communications of the ACM,* 50 (11), 60-64.

Nov, O., Arazy, O. and Anderson, D., 2011. Technology-Mediated Citizen Science Participation: A Motivational Model. *Proceeding of the Fifth International AAAI Conference on Weblogs and Social Media,* July 17-21 Barcelona (ESP). Menlo Park (USA): The AAAI Press, 249-256.

Ochoa, X. and Duval, E., 2008. Quantitative analysis of user-generated content on the web. In: D. De Roure and W. Hall, eds. *Proceedings of the First International Workshop on Understanding Web Evolution (WebEvolve2008): A prerequisite for Web Science,* April 22 Beijing (CHN). 19-26.

OpenStreetMap contributors, 2014a. *Complete OSM Data History* [online]. OpenStreetMap Wiki. Available from: http://planet.openstreetmap.org/planet/full-history/ [Accessed 2014-07-03].

OpenStreetMap contributors, 2014b. *Main Page* [online]. OpenStreetMap Wiki. Available from: http://wiki.openstreetmap.org/wiki/Main_Page [Accessed 2017-06-19].

Penner, L.A., 2002, Dispositional and organizational influences on sustained volunteerism: An interactionist perspective. *Journal of Social Issues,* 58 (3), 447-467.

Perkins, C., 2011, Researching mapping: methods, modes and moments in the (im) mutability of OpenStreetMap. *Global Media Journal-Australian Edition,* 5 (2).

Preece, J. and Shneiderman, B., 2009, The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction,* 1 (1), 13-32.

R Core Team, 2016. *R: A Language and Environment for Statistical Computing.* Vienna (AUT): R Core Team.

Rogers, E.M., 1983. *Diffusion of Innovations.* 3rd ed. New-York (USA): The Free Press.

Schneider, A., Von Krogh, G. and JäGer, P., 2013, "What's coming next?" Epistemic curiosity and lurking behaviour in online communities. *Computers in Human Behavior,* 29 (1), 293-303.

Sui, D.Z., Elwood, S. and Goodchild, M.F., 2013. *Crowdsourcing geographic knowledge.* 1st ed. New York (USA): Springer.

Sun, N., Rau, P.P. and Ma, L., 2014, Understanding lurkers in online communities: A literature review. *Computers in Human Behavior,* 38, 110-117.

Therneau,T.M. and Lumley,T., 2017. *R Survival Package—Survival Analysis.* Fermanagh (IRL): CRAN.

Wang, K.S., Hsu, F. and Liu, P., 2002, Modeling the bathtub shape hazard rate function in terms of reliability. *Reliability Engineering & System Safety,* 75 (3), 397-406.

Wenger, E., 1998. *Communities of practice: Learning, meaning, and identity.* 1st ed. Cambridge (GBR): Cambridge university press.

Weon, B.M., 2016, Tyrannosaurs as long-lived species. *Scientific reports,* 6 (srep19554), 1-5.

Chapter 5: Discussion and Conclusions

Over the past two decades, Web 2.0 has allowed for the development of large collaborative online communities, including some built for the sharing of geographic information (VGI). Since the nature of VGI contributions relies on contributors' motivation, interests, knowledge and skills, several attempts were made to link contributors and their contributions through different typologies.

Although previously proposed typologies were directly or indirectly related to the time spent by the contributors in a project, none used contributors' actual lifespan to understand their behaviours. In order to fill this knowledge gap, I examined VGI contributors' lifespan and its evolution over time by using tools and concepts generally dedicated to the study of demographic data. I aimed to understand the evolution of enrollments in, and withdrawals from, a VGI community by identifying significant variations in both phenomena over time. Variations of enrollments and withdrawals were used as proxy measures of a project's capability to meet contributors' needs, desires or aspiration (i.e. motivation). By linking the most significant variations to events throughout the history of a project, I identified those that most likely affected the motivation and the lifespan of project's participants.

Our research assessed enrollments in the large user-led VGI project (OSM) to identify the different events that changed people's awareness or perception about the project, either encouraging or preventing participation (Chapter 2). By assessing withdrawals from the OSM community (Chapter 3) I aimed to identify the events that may

have reduced contributor motivation. Furthermore, in order to assess withdrawals, I developed a formal approach to differentiate temporary absences from contributors who have left the project permanently. Using survival analyses on contributors' lifespan (Chapter 4), I identified the different phases in a contributors' life cycle from a temporal perspective. An analysis of the volume and the frequency of contributions from participants at each phase highlighted potential relationships between these phases and the nature of contributions.

This research applied robust statistical analyses to the assessment of contributors' behaviours. Confidence in my results relies on the rigorous statistical approach developed to identify withdrawn contributors, and to assess each participant's contribution and circadian cycles.

5.1. Key research findings

This thesis is based upon the hypothesis that the life cycle of VGI contributors follows a series of predictable phases characterized by distinct patterns of behaviour. Chapters 2 and 3 demonstrate that a population of VGI contributors is subject to mortality rates (withdrawals) that change over the contributors' lifespan in the project. The volume of edits appears to be related to underlying processes (learning-adaptation, cumulated-damages), while the frequency of contributions is related to the volume of contributions (Nielsen).

5.1.1 Answers to initial questions

1- What are the different phases of the life cycle of VGI contributors?

Our analyses showed that there are currently six phases in the life cycle of OSM contributors. During the first two phases ("Visitors" and "Explorers") more than half of contributors quit the project within six months of enrollment. Over the following six years ("Adopters" and "Veterans") this proportion drops by twenty percent. Finally, the last phases ("Elders" and "Founders") see an exponential increase of withdrawals. The maximum lifespan of contributors in OSM project has not yet been reached and some early participants are still active.

2- Is there a relationship between the different phases and the nature of provided contributions and if so, what are these relationships?

I found that the volume of contribution increases over the first three phases before decreasing over the last ones. The frequency of contributions shows a more complex pattern that seems related to both the different phases and the volume of contributions provided by each participant (Nielsen). The frequency of contributions tends to lower over time before increasing over last phases depending on Nielsen's categories.

3- Are there any events, or factors that have changed contributors' lifespan throughout the development of VGI communities?

I did not find any specific event that altered substantially the lifespan of contributors. However, I found events that affected the project's enrollments and withdrawals.

4- What events or factors affected enrollments in a community over years?

Improvements to the contribution environment have had a positive impact on enrollment, as has the increasing recognition of the project by authoritative individuals and credible organizations. Active internal conflicts seem to prevent participants from registering or starting contributing.

5- What events or factors affected withdrawals from a community over the same period?

Internal conflicts proved to be the main cause of collective withdrawals from the community, but it affected only a small proportion of contributors. Interestingly, some indications show that even improvements to the contributing environment drive some people to withdraw from a project. I suspect that concerned people may be affected by a cumulative-damage process but it will require further research to confirm this.

While answering these questions, I also found environmental factors affected the number of contributors in the OSM project and the phases of their life cycle. I identified underlying structures that seem to determine phase duration as well as some aspects of their contributions during these phases.

5.1.2 Collaborative and Technical Environmental Factors

Time series analyses (Chapters 2 and 3) provided a detailed view of the variations in both enrollments and withdrawals over time. Linking their trend and random components to the events that dotted the history of the project has shown different mechanisms by which a project grows, stabilizes, or declines. The events identified as bringing new contributors

to the community were found to change over time, in parallel with the project's development.

During the establishment of OSM and until the infrastructure and applications matured, contributors were limited to a small group of people with advanced knowledge and technical skills. Once a certain threshold was reached in available tools, project growth accelerated and subsequent system improvements usually generated bursts of new contributors. However, I also found that major improvements were followed by smaller, but nevertheless important, number of withdrawals. This unexpected effect on some contributors' motivation will need to be explored further. Another unexpected result found following an improvement to the project is the rapid increase of the proportion of "lurkers" after 2009. Although this increase seemed related to a linguistic barrier resulting from the asynchronous translations of registration and contribution interfaces, the time at which it happened also correlated with a change in contributors' DoIT categories, which will be discussed later.

Regarding the impact of human factors on contributors' life cycle my findings are twofold. First, I found that internal conflicts may have important impacts on people's perception about the capability of a community to meet their needs, desires and aspirations. An internal conflict not only affects actual contributors, as might be expected, but it may have an even greater effect on the number of potential new participants. In Chapter 3, I concluded that the conflict about the OSM licence change was likely connected to thousands of withdrawals from the project. However, by examining the drop

in the daily rates of enrollment over the same period (Chapter 2) the number of people that might have, but did not, enroll in the project is on the order of tens of thousands of people. Even newly registered participants refrained from making first edits during that period.

Second, I found that the profile of participants who enroll changed over time. For instance, the proportion of contributors who have withdrawn from the project is 70% higher among those who have enrolled in recent years compared to early contributors. Similarly, the nature of the events that brought these early contributors to the project (personal communications) is quite different from the one that brought the latest contributors (recognition of the project by large online communities).

These findings led me to consider the development phases of a project (DoIT) as a determining factor of the behavioural profile of the participants who enroll.

5.1.3 Diffusion of innovation theory

In previous chapters, I proposed that new OSM participants' behaviours have evolved over years according to early phases of DoIT. I have found that "Innovators" (2005–2007) do not seem to go through an "Assessment" stage, have a high retention rate in the project (50%) and enrolled to the project after being introduced to OSM by other participants. The "Early majority" (2009+) shows a well-marked "Assessment" stage, a low retention rate (15%), and adopted the project after large communities had migrated to OSM, giving credibility to the project. My results present the "Early adopters" (2007–2009) as a transition stage between "Innovators" and "Early majority". These participants were attracted to the project by authoritative members of the GIS and/or OSM

communities.

Diffusion of innovation theory was originally developed by assessing behaviours of adopters of new ideas or technology. These innovations were usually offered to potential adopters after R&D cycles completed. However, in the case of OSM, "Early adopters" had access to the innovation (the project) long before main R&D cycles were completed. As a "user-led" project, many of early OSM participants were involved in its development and, accordingly, a majority of these participants were found to be open-source software (OSS) developers (Budhathoki 2010).

A technology adoption model derived from DoIT was proposed by Moore (2001), which seems more adapted to users-led communities where participants are involved right from the start in project's development (the innovation), is presented in Figure 5-1.

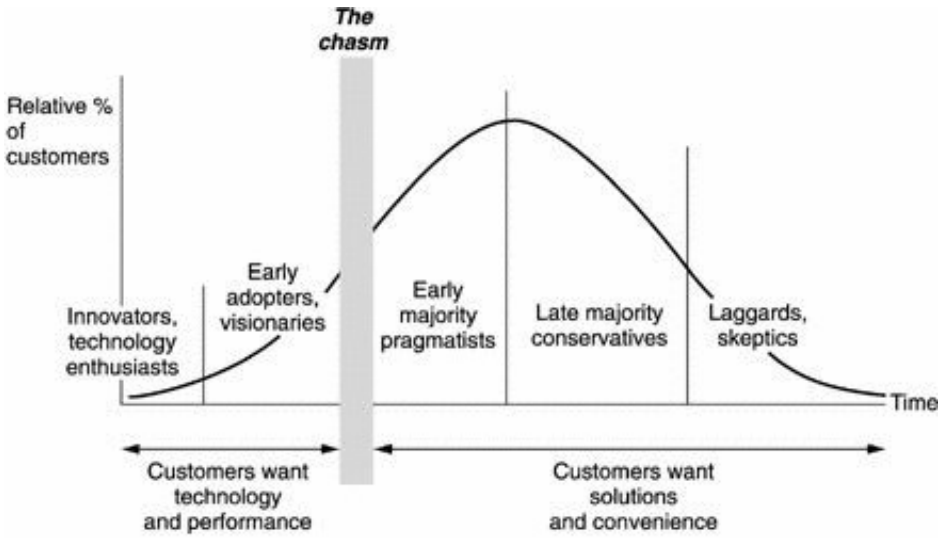


Figure 5-1 Technology adoption model adapted from Moore (Searls 2003)

The main contribution of Moore's proposal is the existence of a gap ("the chasm")

between “Early adopters” and “Early majority” stages where many technological innovations plateau or decline because they never reach a critical mass of adopters.

This chasm results from a lack of buy-in from the “Early majority” which derives from a conflict of objectives, motivation, or even personality profiles between the “Early majority” and “Early adopters.” “Early adopters” are described as visionaries that seek for game-changer innovations if it can bring them to advantage. They are willing to “bear with the inevitable bugs and glitches that accompany any innovation” (Moore 2001). “Early majority” participants are described as pragmatic people looking only at innovations for productivity improvements. They prefer to see the bugs cleaned up by others and they want to see an endorsement of the technology by others before investing time on it. By contrast, the visionaries (“Early adopters”) are more likely to implement an innovation and then move on to the next project (Moore 2001).

Interestingly, I located the boundary between OSM “Early adopters” and “Early majority” in mid-2009, a few months before the two-year conflict about the licence change broke out. This raises questions about the actual origin of the conflict and the stagnation of enrollments over that period. On the one hand, most innovations show such stagnation or even declines of acceptance rates at this stage. Consequently, it is possible that the observations I made in the previous section and in Chapter 2 (i.e. 2010–2012 enrollment stagnation) could have happened even without this conflict. On the other hand, this conflict could be the expression of the tensions resulting from these two different visions of project’s future, one idealist and the other pragmatic.

A similar question arises concerning the increasing proportion of lurkers after 2009. In this case, instead of resulting from a linguistic barrier, it may result from an increasing proportion of pragmatic participants ("Early majority") who were assessing the project. Many of these participants could have registered to the project in order to assess available online tools and infrastructure, and judged they were sufficient at that time. We found that the rate of lurkers has stabilized since 2014 which, interestingly, fits the replacement of the default online OSM editor (Potlatch 2) by a new one (ID) that is still used today. Among numerous improvements, this one may have significantly changed the perceptions of these new pragmatic participants.

5.1.4 Underlying Structures to Phase Determination

A major contribution of this thesis is to have identified underlying structures to the phases of contributors' life cycle. The basic concepts behind hazard curve's interpretation were derived from reliability engineering (Wang *et al.* 2002) and modelled surprisingly well contributors' behaviours.

The first underlying structure was revealed by the shape of the hazard curve obtained from the survival analysis (Chapters 3 and 4). Its bathtub shape shows with striking evidence that contributors are going through three important stages in their life cycle. The first one is an "Assessment" stage during which more than 50% of participants quit the project within a few months. The second one is an "Engagement" stage during which contributors engage in the project for years. The last one is a "Detachment" stage during which contributors withdraw from the project at an increasing rate. Reliability

engineers refer to these three stages as “infant mortality,” “useful life” and “wear out” stages. This underlying structure is simple and refers to general behaviours that I expect to apply to most collaborative projects and leisure activities.

The second underlying structure is defined by two processes associated with different behaviours. When they enroll in a community, the contributors enter a learning-adaptation-dominated process. During this period, they adapt to prevailing norms and rules, learn how to contribute with provided tools, and adjust to fit perceived community requirements. This process requires continuous effort that must be fuelled by the satisfaction derived from contributing (motivation). This process corresponds to the first three phases of contributors’ life cycle (“Visitors”, “Explorers” and “Adaptors”) and I found that the volume of edits made per active days increases during these phases.

After years of contributing and adapting to changing tools, norms and rules (particularly in users-led communities), motivation seems to wear off in a cumulative-damage-dominated process. This process corresponds to the last three phases of contributors’ life cycle (“Veterans”, “Elders” and “Founders”); during this period I found that the volume of edits made per active days decreases.

In reliability engineering, the switch from first to second process is arbitrarily located in the middle of the useful life of a system. In this case, I located this change in the Engagement stage, where the daily rate of withdrawal reached its minimum. Other authors made similar findings about contributors’ behaviour (Danescu-Niculescu-Mizil *et al.* 2013) and located this shift at about one third of contributors’ lifespan. The authors were

even able to predict when participants would withdraw from the project from the time at which they were adopting a more conservative attitude.

5.2. Practical implications

Early in this research, it became clear that the characterization of the VGI data (i.e. its content and quality) necessarily involved the characterization of their contributors (Bégin *et al.* 2013). However, our knowledge of contributors' behaviour was too limited to move forward, considering that more fundamental knowledge was still missing about contributors' basic behaviour. This research has addressed some of these knowledge gaps by identifying factors that must be taken into account when studying the behaviour of online contributors, whether from a VGI community or not.

I found that automation processes (OSM API, editing tools setup) may alter perceived contributors' lifespan over the first hours of contribution. I also found that in a context where contributions come from all around the world, it is necessary to consider the actual circadian cycle of contributors when assessing the frequency of contributions from UTC timestamps. The practical implication of these findings is that even before analyzing their data, one should consider the implicit and explicit impacts the data storage (UTC) and automation processes (i.e., API) may have on raw data and their derived results.

During the literature review, I noticed that all studies of contributors' lifespan were using arbitrary criteria to differentiate those who had quit a project from those who temporarily absent between contributions. The result is that these studies could not be

compared in order to derive trends in contributors' behaviours. I decided to develop a rigorous statistical approach to identify withdrawn contributors from their historical frequency of contributions using a given level of certainty (probability). The approach used to assess contributors' lifespan in an online project enables comparisons between the behaviours of contributors from online communities. This is a significant step for online communities' studies since it provides for the first time a tool that can help standardize the concepts of "active," "inactive" and "withdrawn" contributors.

I have found that the stage of diffusion of a project (innovation) has an impact on the profile of the contributors that enroll at a given time. This is of prime importance since it determines the proportion of contributors that will engage with the project for the long term. It also defines contributors' expectations regarding project's development, and their decision to enroll. This is particularly true when a project crosses the "chasm" which would determine if a project engages an "Early majority." Based on these findings, assessment of contributors' behaviour requires an understanding at which phase of DoIT each contributor enrolled. Furthermore, it also suggests that changes in new contributors' profiles may result, as predicted by the theory, in tensions in the community, conflicts and withdrawals, which could bring a shift in community's values and beliefs, and then in community's behavior. Literature has identified such shifts in communities' behaviour, like in Wikipedia, something that should be examined from this perspective.

When changes are made to the environment in which participants contribute, the potential impact of these changes needs to be adequately addressed. I found that even

improvements to the contribution environment may result in the withdrawal of many contributors. Given the voluntary nature of their participation, changes imposed to contributors' habits may be enough to make them quit the project, particularly if they are in a phase affected by the cumulative-damage process.

The underlying three-stage structure and the two processes that determined the phases of contributor life cycle influence contributor behaviour. Although I identified only one clear correlation between the volume of contributions and the hazard curves underlying processes (learning-adaptation and cumulative damage), one should expect these curves and their underlying structures to have an effect on a large number of contributors' behaviours.

The criterion used to determine OSM contributor' life cycle phases (i.e., contributors' lifespan) is not specific to a project (e.g., constraints, goals, requirements) but only relates to the time contributors freely contribute to an unsupervised leisure activity. The method proposed in this research provides a new analysis framework for unsupervised leisure activities. The findings of this research are then likely apply to most UGC communities, not only VGI ones. In order to assess the potential scope of the method, I presented the results (Chapter 4) to Dr. R.A. Stebbins from the University of Calgary, a pioneer in serious, casual and project-based leisure studies. Looking at the results from a serious leisure perspective (SLP)[26], he considered that the approach "gives considerable substance to the leisure careers of those who participate in the OSM project, much more than anyone approaching amateur science careers from a purely SLP point of

[26]https://www.seriousleisure.net/

view." Consequently, the proposed approach might be used to understand the structure of the lifespan of participants in serious leisure activities as well.

Online community managers could use the different stages or phase of the life cycle of their contributors to identify when and which retention techniques can have the most impact. For example, techniques that aim to attract and retain new contributors (learning-adaptation phase) should be different from those deployed to keep experienced contributors (cumulative damage phase). The nature of the techniques to apply at each phase was however not in the scope of this research. Online community managers must also be very careful when applying changes to the contribution environment (e.g., rules, norms, applications). The results show that most changes that aimed at improving the retention of contributors also brought withdrawals from the project. Introducing such changes should be presented first as alternatives, the time for experienced participants to assess these alternatives and to adapt, while new participants could be oriented directly toward the proposed alternatives. Finally, I have shown that using contributors' circadian cycle is and effective way to aggregate on a daily basis contributions made from all around the world. Without this approach, online community managers may obtain biases results when assessing the daily frequency of contributions.

5.3. Limitations and Future work

The procedures and methods used or developed throughout this research can be further improved, although I have taken great care to limit the impact of remaining uncertainties.

In my opinion, the greatest source of uncertainty in this thesis is the manual

assignment of boundaries between life cycle phases. Although pivotal points of two curves from distinct analyses were used, the location of latest phases could have been placed months away on both sides. We also observed that the duration of some phases has varied over the years as predicted by DoIT stages, something that could have affected boundaries' location. Finally, phases will continue to evolve as long as the maximum lifespan of OSM contributors has not been reached.

Considering that tools and procedures have been proposed to allow comparative results, my findings about underlying structures and processes that define the phases of contributors' life cycle must be examined and confirmed with other online communities. The approach shows a great potential of application way outside the VGI realm. However, I expect the duration of each phase to change between communities as contributors' maximum timespans may change as well. Although not present in OSM history, the many events that affect contributors' lifespan in a community may result in very distinct populations in a same project (e.g., before and after a given event). A major disruption in either contributors' enrollments, withdrawals or both could create two or more distinct life cycles. Similarly, the structure of the activity may bring supplementary phases as milestones may be imposed to contributors, particularly when using gamification process to motivate them (e.g., badges or privileges). Some results suggested that early phases of contributors' life cycle ('Visitors" and "Explorer") were not present when "Innovators" (DoIT) were enrolling. At the same time, they show that 50% of early contributors are still active. Consequently, using a temporal approach to analyze the life cycle of contributors in a new project may not be adequate.

Our findings regarding the two processes that seem to drive contributors' behaviours were supported by a paper that also proposed that contributors' withdrawal could be predicted based on the nature of their contributions (Danescu-Niculescu-Mizil *et al.* 2013). This is of great interest since it suggests that some aspects of OSM contributions could potentially provide us with similar information. This highlights the fact that the metrics used to assess contributors' behaviours (volume of edits and frequency of contributions) were very limited. This discussion opens a vast field of research on the identification of metrics which would make it possible to characterize the behaviors of VGI contributors, anticipate their withdrawal, adjust the participatory environment to maintain their commitment, or even to increase it, particularly for early phases.

Our understanding of the impact of DoIT on online communities must be deepened. It would be interesting whether current OSM participants still have profiles that correspond to the "Early majority," or if they drifted toward the "Late majority" with more conservative profiles. Comparisons with communities would also be valuable; has crossing the chasm also generated conflicts beyond OSM, or was the timing at which the licence conflict happened in OSM coincidental? Similarly, does the drop of the contribution ratio that coincided with the advent of the "Early majority" also due to chance?

The accurate measurement actions made within an online community constitute a valuable source of proxy measures about general human behaviour. The value of such data comes from the fact that participants freely contribute, prioritize and take actions with

little or no constraints, to answer their needs, desires and aspirations (i.e. motivation). This thesis has highlighted different behaviours among an online community that mimic human behaviour at a much larger scale. For instance, the events that affected contributors' lifespan could mimic the events that affected people during mankind history (epidemics, wars and technological improvements). Exploring enrollments, I shed light on behaviours described by the diffusion of innovation theory. We saw how participants reacted to a conflict within the OSM community. By exploring the withdrawals of participants, I discovered a sequence of behaviours that seem so intuitive that it may apply broadly to online communities and even general leisure activities. We were able to assess participants' circadian cycle and at one point, I even modelled a behaviour that mimicked patterns seen in addictive processes.

Assessing the contributions from people who freely engage with an online community not only shed light on their contributions but also reveals different aspects of human behaviours.

5.4. References

Bégin, D., Devillers, R. and Roche, S., 2013. Assessing Volunteered Geographic Information (VGI) Quality Based On Contributors' Mapping Behaviours. In: B. Wu, W.J. SHI and E. Gilbert, eds. *8th International Symposium on Spatial Data Quality,* May 30—June 1 Hong Kong (CHN). Hannover (GER): ISPRS, 149-154.

Budhathoki, N.R., 2010, Participants' motivations to contribute geographic information in an online community. Thesis (PhD). Graduate College of the University of Illinois.

Danescu-Niculescu-Mizil, C., *et al.*, 2013. No country for old members: User lifecycle and linguistic change in online communities. *Proceedings of the 22nd international conference on World Wide Web,* ACM, 307-318.

Moore, G.A., 2001. Crossing the chasm—marketing and selling high-tech products to mainstream customers. Revised ed. New York (USA): HarperCollins.

Searls, D. 2003, *Closing the Chasm*, 109th edn, Linux Journal LLC, Denver (USA).

Wang, K.S., Hsu, F. and Liu, P., 2002, Modeling the bathtub shape hazard rate function in terms of reliability. *Reliability Engineering & System Safety,* 75 (3), 397-406.

Bibliography and References

Ahmouda, A. and Hochmair, H.H., 2017, Using Volunteered Geographic Information to measure name changes of artificial geographical features as a result of political changes: a Libya case study. *GeoJournal,* 82 (1), 1-19.

Aknouche, L. and Shoan, G., 2013, *Motivations for Open Source Project Entrance and Continued Participation.* Thesis (Master). Lund University.

Amichai-Hamburger, Y., et al., 2016, Psychological factors behind the lack of participation in online discussions. *Computers in Human Behavior,* 55, 268-277.

Arsanjani, J.J., et al., 2013. Assessing the Quality of OpenStreetMap Contributors together with their Contributions. *The 15th AGILE International Conference on Geographic Information Science—Short Papers,* May 14-17 Leuven (BEL). Technische Universität Dresden, 1-4.

Arsanjani, J.J., et al., 2015, An exploration of future patterns of the contributions to OpenStreetMap and development of a Contribution Index. *Transactions in GIS,* 19 (6), 896-914.

Arsanjani, J.J*., et al.,* eds., 2015. *OpenStreetMap in GIScience.* Cham (CHE): Springer.

Beaulieu, A., Bégin, D. and Genest, D., 2010. Community Mapping and Government Mapping: Potential Collaboration? *Symposium of Commission I, ISPRS,* 16-18 June 2010 Calgary (CAN). 1-3.

Bégin, D., 2012. Towards Integrating VGI and National Mapping Agency Operations—A Canadian Case Study. *Role of Volunteer Geographic Information in Advancing Science: Quality and Credibility Workshop,* 18 September 2012 Columbus (USA). 1-2.

Bégin, D., Devillers, R. and Roche, S., 2013. Assessing Volunteered Geographic Information (VGI) Quality Based On Contributors' Mapping Behaviours. In: B. Wu, W.J. SHI and E. Gilbert, eds. *8th International Symposium on Spatial Data Quality,* May 30—June 1 Hong Kong (CHN). Hannover (GER): ISPRS, 149-154.

Bégin, D., Devillers, R. and Roche, S., 2017, Contributors' Enrollment in Collaborative Online Communities: The Case of OpenStreetMap. *Geo-spatial Information Science,* 19 (3), 282-295.

Bégin, D., Devillers, R. and Roche, S., 2017, Contributors' Withdrawal from Online Collaborative Communities, the Case of OpenStreetMap. *ISPRS International Journal of Geo-Information,* 6 (11), 340.1-340.20.

Benkler, Y., 2002, Coase's Penguin, or, Linux and" The Nature of the Firm". *Yale Law Journal,* , 369-446.

Borst, W.A.M., 2010. *Understanding Crowdsourcing—Effects of motivation and rewards on participation and performance in voluntary online activities.* 1st ed. Rotterdam (NLD): Erasmus University of Rotterdam.

Bright, J., De Sabbata, S. and Lee, S., 2017, Geodemographic biases in crowdsourced knowledge websites: Do neighbours fill in the blanks? *GeoJournal,* [Online], 1-14.

Brown, J.J. and Reingen, P.H., 1987, Social ties and word-of-mouth referral behaviour. *Journal of Consumer research,* 14 (3), 350-362.

Bruns, A., 2006. Towards Produsage: Futures for user-led content production. In: F. Sudweeks, H. Hrachovec and C. Ess, eds. *Proceedings of Cultural Attitudes towards Communication and Technology,* June 28—July 01 Tartu (EST). 275-284.

Bruns, A., 2008. *Blogs, Wikipedia, Second life, and beyond: from production to produsage.* 1st ed. New York (USA): Peter Lang.

Bryant, S.L., Forte, A. and Bruckman, A., 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work,* November 6-9 Sanibel Island (USA). New York (USA): ACM, 1-10.

Budhathoki, N.R., 2010, *Participants' motivations to contribute geographic information in an online community.* Thesis (PhD). Graduate College of the University of Illinois.

Budhathoki, N.R. and Haythornthwaite, C., 2013, Motivation for Open Collaboration Crowd and Community Models and the Case of OpenStreetMap. *American Behavioral Scientist,* 57 (5), 548-575.

Budhathoki, N.R., Nedovic-Budic, Z. and Bruce, B., 2010, An interdisciplinary frame for understanding volunteered geographic information. *Geomatica,* 64 (1), 11-26.

Capineri, C*., et al.*, 2016. *European handbook of crowdsourced geographic information.* 1st ed. London (GBR): Ubiquity Press.

Chacon, F., Vecina, M.L. and Davila, M.C., 2007, The Three-Stage Model of Volunteers' Duration of Service. *Social behaviour and personality,* 35 (5), 627-642.

Chakraborty, M., et al., 2016, Recent developments in social spam detection and combating techniques: A survey. *Information Processing & Management,* 52 (6), 1053-1073.

Cheung, K.S., et al., 2005, The development of successful on-line communities. *International Journal of the Computer, the Internet and Management,* 13 (1), 71-89.

Clark, J.O.E., 2005. *100 maps: The science, art and politics of cartography throughout history.* Sterling Publishing Company, Inc.

Clary, E.G., 1998, Understanding and assessing the motivations of volunteers: A functional approach. *Journal of personality and social psychology,* 74 (6), 1516-1530.

Clauset, A., Shalizi, C.R. and Newman, M.E., 2009, Power-law distributions in empirical data. *SIAM Review,* 51 (4), 661-703.

Coast, S., 2007. *The Pragmatic Mapper (part deux)* [online]. blog.openstreetmap.org. Available from: https://blog.openstreetmap.org/2007/03/27/the-pragmatic-mapper-part-deaux/ [Accessed 2017-05-21].

Coast, S., 2011. How OpenStreetMap Is Changing the World. In: K. Tanaka, P. Fröhlich and K. Kim, eds. *International Symposium on Web and Wireless Geographical Information Systems,* March 3-4 Kyoto (JPN). Berlin (GER): Springer-Verlag, 4-4.

Coleman, D.J., 2010. Volunteered geographic information in spatial data infrastructure: An early look at opportunities and constraints. *Proceedings of GSDI 12th conference,* 19-22 October Singapore (SGP). 1-18.

Coleman, D.J., Georgiadou, Y. and Labonté, J., 2009, Volunteered geographic information: The nature and motivation of producers. *International Journal of Spatial Data Infrastructures Research,* 4, 332-358.

Comber, A., et al., 2016. A moan, a discursion into the visualisation of very large spatial data and some rubrics for identifying big questions. *International Conference on GIScience Short Paper Proceedings,* .

Corcoran, P., Mooney, P. and Bertolotto, M., 2013, Analysing the growth of OpenStreetMap networks. *Spatial Statistics,* 3, 21-32.

Danescu-Niculescu-Mizil, C., et al., 2013. No country for old members: User lifecycle and linguistic change in online communities. *Proceedings of the 22nd international conference on World Wide Web,* ACM, 307-318.

DiBiase, D*., et al.*, 2006. *Geographic Information Science &Technology—Body Of Knowledge.* 1st ed. Washington (USA): Association of American Geographers.

Dorn, H., Törnros, T. and Zipf, A., 2015, Quality evaluation of VGI using authoritative data—A comparison with land use data in Southern Germany. *ISPRS International Journal of Geo-Information,* 4 (3), 1657-1671.

Downs, R.M. and DeSouza, A., 2006. *Learning to think spatially: GIS as a support system in the K-12 curriculum.* 1st ed. Washington (USA): The National Academies Press.

Elwood, S., Goodchild, M.F. and Sui, D.Z., 2012, Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers,* 102 (3), 571-590.

Estes, J.E. and Mooneyhan, D.W., 1994, Of maps and myths. *Photogrammetric Engineering and Remote Sensing,* 60 (5).

Fan, J., Han, F. and Liu, H., 2014, Challenges of big data analysis. *National science review,* 1 (2), 293-314.

Fishbein, M. and Ajzen, I., 1975. *Belief, attitude, intention, and behaviour: An introduction to theory and research.* Reading (USA): Addison-Wesley.

Flanagin, A.J. and Metzger, M.J., 2008, The credibility of volunteered geographic information. *GeoJournal,* 72 (3-4), 137-148.

Gandomi, A. and Haider, M., 2015, Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management,* 35 (2), 137-144.

Garling, C. 2012, *Google workers caught 'vandalizing' open source maps*, Wire.com (Business), January 17 [online]. Available from: https://www.wired.com/2012/01/osm-google-accusation/.

Girres, J. and Touya, G.G., 2010, Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS,* 14 (4), 435-459.

Goodchild, M.F., 2007, Citizens as sensors: the world of volunteered geography. *GeoJournal,* 69 (4), 211-221.

Goodchild, M.F., 2007, Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0—Editorial. *International Journal of Spatial Data Infrastructures Research,* 2, 24-32.

Google, 2011. *Google search and search engine spam* [online]. Available from: https://googleblog.blogspot.ca/2011/01/google-search-and-search-engine-spam.html.

Google, 2012. *Fighting Spam* [online]. Available from: https://www.google.ca/insidesearch/howsearchworks/fighting-spam.html.

Gyongyi, Z. and Garcia-Molina, H., 2005. Web spam taxonomy. *First international workshop on adversarial information retrieval on the web (AIRWeb 2005),* May 10-14 Chiba, Japan.

Haklay, M., 2010, How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design,* 37 (4), 682-703.

Haklay, M., et al., 2010, How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal,* 47 (4), 315-322.

Haklay, M. and Weber, P., 2008, OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing,* 7 (4), 12-18.

Heckhausen, J. and Heckhausen, H., 2008. *Motivation and action.* 1st ed. New York (USA): Cambridge University Press.

Heckhausen, J., Wrosch, C. and Schulz, R., 2010, A motivational theory of life-span development. *Psychological review,* 117 (1), 32-60.

Hemetsberger, A. and Pieters, R., 2003. *When consumers produce on the internet: the relationship between cognitive-affective, socially-based, and behavioral involvement of prosumers* [online]. CiteSeerX. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.9299&rep=rep1&type=pdf.

Horita, F.E.A., et al., 2013. The use of volunteered geographic information (VGI) and crowdsourcing in disaster management: a systematic literature review. *Nineteenth Americas Conference on Information Systems,* August 15 − 17 Chicago, Illinois, USA.

Houle, B.B.J., 2005, A Functional Approach to Volunteerism: Do Volunteer Motives Predict Task Preference? *Basic and applied social psychology,* 27 (4), 337-344.

Howe, J. 2006, *The rise of crowdsourcing*, June 14th edn, Condé Nast Publications, New York (USA).

Hristova, D., et al., 2013. The Life of the Party: Impact of Social Mapping in OpenStreetMap. *International Conference On Web And Social Media Papers,* July 8–11 Cambridge, Massachusetts, USA. Palo Alto, California, USA: The AAAI Press, 234-243.

Hyndman, R.J. and Athanasopoulos, G., 2014. *Forecasting: Principles and Practice.* 1st ed. Melbourne (AUS): OTexts.

Keßler, C. and de Groot, René Theodore Anton, 2013. Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap. In: D. Vandenbroucke, B. Bucher and J. Crompvoets, eds. Geographic Information Science at the Heart of Europe. Springer International Publishing., 21-37.

Kimura, A.H. and Kinchy, A., 2016, Citizen Science: Probing the Virtues and Contexts of Participatory Research. *Engaging Science, Technology, and Society,* 2, 331-361.

Laney, D., 2001, 3D data management: Controlling data volume, velocity and variety. *META Group Research Note,* 6, 70.

Lave, J. and Wenger, E., 1991. *Situated learning: Legitimate peripheral participation.* 1st ed. Cambridge (GBR): Cambridge university press.

Limpert, E., Stahel, W.A. and Abbt, M., 2001, Log-normal Distributions across the Sciences: Keys and Clues. *Bioscience,* 51 (5), 341-352.

Ma, D., Sandberg, M. and Jiang, B., 2015, Characterizing the Heterogeneity of the OpenStreetMap Data and Community. *ISPRS International Journal of Geo-Information,* 4, 535-550.

Mashhadi, A., Quattrone, G. and Capra, L., 2015. The impact of society on volunteered geographic information: The case of OpenStreetMap. In: J. Jokar Arsanjani*, et al.*, eds. OpenStreetMap in GIScience. Berlin: Heidelberg: Springer, 125-141.

McLeod, A.I., Yu, H. and Mahdi, E., 2011. Time series analysis with R. In: C.R. Rao, ed. Time Series Analysis: Methods and Applications. Oxford (GBR): Elsevier, 661-707.

Metzger, M.J., 2010, Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of communication,* 60 (3), 413-439.

Michelucci, P. and Dickinson, J.L., 2016, The power of crowds. *Science,* 351 (6268), 32-33.

Mitzenmacher, M., 2004, A brief history of generative models for power law and lognormal distributions. *Internet mathematics,* 1 (2), 226-251.

Mooney, P. and Corcoran, P., 2012. How social is OpenStreetMap? In: J. Gensel, D. Josselin and D. Vandenbroucke, eds. *The 15th AGILE International Conference on Geographic Information Science,* April 24-27 Avignon (FRA). Springer, 1-6.

Mooney, P. and Corcoran, P., 2012. Who are the contributors to OpenStreetMap and what do they do? *Proceedings of the GIS Research UK 20th Annual Conference,* April 11-13 Lancaster (GBR). Lancaster (GBR): Lancaster University, 355-360.

Mooney, P. and Corcoran, P., 2013, Analysis of Interaction and Co-editing Patterns amongst OpenStreetMap Contributors. *Transactions in GIS,* 18 (5), 633-659.

Mooney, P., Corcoran, P. and Winstanley, A.C., 2010. A study of data representation of natural features in OpenStreetMap. *Proceedings of GIScience,* 14-17 September Zurich (CHE). 150-156.

Moore, G.A., 2001. *Crossing the chasm—marketing and selling high-tech products to mainstream customers.* Revised ed. New York (USA): HarperCollins.

Napolitano, M. and Mooney, P., 2012, MVP OSM: A Tool to identify Areas of High Quality Contributor Activity in OpenStreetMap. *The Bulletin of the Society of Cartographers,* 45 (1), 10-18.

Neis, P. and Zielstra, D., 2014, Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap. *Future Internet,* 6 (1), 76-106.

Neis, P. and Zipf, A., 2012, Analyzing the Contributor Activity of a Volunteered Geographic Information Project—The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information,* 1 (2), 146-165.

Nielsen, J., 2006. *The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities* [online]. Nielsen Norman Group. Available from: http://www.useit.com/alertbox/participation_inequality.html [Accessed 2012-10-26].

Nov, O., 2007, What motivates wikipedians? *Communications of the ACM,* 50 (11), 60-64.

Nov, O., Arazy, O. and Anderson, D., 2011. Technology-Mediated Citizen Science Participation: A Motivational Model. *Proceeding of the Fifth International AAAI Conference on Weblogs and Social Media,* July 17-21 Barcelona (ESP). Menlo Park (USA): The AAAI Press, 249-256.

Ochoa, X. and Duval, E., 2008. Quantitative analysis of user-generated content on the web. In: D. De Roure and W. Hall, eds. *Proceedings of the First International Workshop on Understanding Web Evolution (WebEvolve2008): A prerequisite for Web Science,* April 22 Beijing (CHN). 19-26.

OpenStreetMap contributors, 2013. *Stats* [online]. OpenStreetMap Wiki, . Available from: http://wiki.openstreetmap.org/wiki/Stats [Accessed 2012-01-21].

OpenStreetMap contributors, 2014. *Complete OSM Data History* [online]. OpenStreetMap Wiki. Available from: http://planet.openstreetmap.org/planet/full-history/ [Accessed 2014-07-03].

OpenStreetMap contributors, 2014. *Main Page* [online]. OpenStreetMap Wiki. Available from: http://wiki.openstreetmap.org/wiki/Main_Page [Accessed 2017-06-19].

OpenStreetMap contributors, 2017. *History of OpenStreetMap* [online]. OpenStreetMap Wiki. Available from: http://wiki.openstreetmap.org/w/index.php?title=History_of_OpenStreetMap&oldid=1425869 [Accessed 2017-04-07].

OpenStreetMap Foundation, 2017. *OpenStreetMap blog* [online]. Available from: https://blog.openstreetmap.org/ [Accessed 2017-04-07].

O'Reilly, T., 2005. *What is web 2.0: Design patterns and business models for the next generations software* [online]. O'Reilly Media, inc. Available from: http://oreilly.com/web2/archive/what-is-web-20.html [Accessed 2017-12-04].

O'Reilly, T. and Battelle, J., 2009, Web squared: Web 2.0 five years on. *Web 2.0 Summit,* .

Penner, L.A., 2002, Dispositional and organizational influences on sustained volunteerism: An interactionist perspective. *Journal of Social Issues,* 58 (3), 447-467.

Perkins, C., 2011, Researching mapping: methods, modes and moments in the (im) mutability of OpenStreetMap. *Global Media Journal-Australian Edition,* 5 (2).

Poiani, T.H., et al., 2016. Potential of collaborative mapping for disaster relief: A case study of OpenStreetMap in the Nepal earthquake 2015. *49th Hawaii International Conference on System Sciences (HICSS),* January 5-8 Koloa, HI, USA. IEEE, 188-197.

Preece, J., Nonnecke, B. and Andrews, D., 2004, The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior,* 20 (2), 201-223.

Preece, J. and Shneiderman, B., 2009, The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction,* 1 (1), 13-32.

R Core Team, 2016. *R: A Language and Environment for Statistical Computing.* Vienna (AUT): R Core Team.

Raymond, E., 1999, The cathedral and the bazaar. *Knowledge, Technology & Policy,* 12 (3), 23-49.

Rehrl, K. and Gröchenig, S., 2016, A Framework for Data-Centric Analysis of Mapping Activity in the Context of Volunteered Geographic Information. *ISPRS International Journal of Geo-Information,* 5 (3), 37.

Rehrl, K.*, et al.*, 2013. A conceptual model for analyzing contribution patterns in the context of VGI. In: J.M. Krisp, ed. Progress in Location-Based Services, Lecture Notes in Geoinformation and Cartography. Berlin (DEU): Springer-Verlag, 373-388.

Riesch, H. and Potter, C., 2014, Citizen science as seen by scientists: Methodological, epistemological and ethical dimensions. *Public Understanding of Science,* 23 (1), 107-120.

Roche, S., 2012. *Should VGI map space or places? Geographic Information Science Workshop Presentation (GIScience 2012)* [online]. Oak Ridge National Laboratory. Available from: http://www.ornl.gov/sci/gist/workshops/2012/vgi_documents/Roche.pdf [Accessed 2012-11-23].

Rogers, E.M., 1983. *Diffusion of Innovations.* 3rd ed. New-York (USA): The Free Press.

Ryan, R.M. and Deci, E.L., 2000, Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology,* 25 (1), 54-67.

Ryan, R.M. and Deci, E.L., 2000, Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist,* 55 (1), 68.

Sabone, B. 2009, *Assessing Alternative Technologies for Use of Volunteered Geographic Information in Authoritative Databases*.

Schneider, A., Von Krogh, G. and JäGer, P., 2013, "What's coming next?" Epistemic curiosity and lurking behaviour in online communities. *Computers in Human Behavior,* 29 (1), 293-303.

Searls, D. 2003, *Closing the Chasm*, 109th edn, Linux Journal LLC, Denver (USA).

See, L., et al., 2013, Comparing the quality of crowdsourced data contributed by expert and non-experts. *PloS one,* 8 (7), e69958.

Shalizi, C.R., 2007. *So You Think You Have a Power Law—Well Isn't That Special?* [online]. bactra.org. Available from: http://bactra.org/weblog/491.html [Accessed 2017-12-21].

Singleton, A.D., Spielman, S. and Brunsdon, C., 2016, Establishing a framework for Open Geographic Information science. *International Journal of Geographical Information Science,* 30 (8), 1507-1521.

Soden, R. and Palen, L., 2014. From crowdsourced mapping to community mapping: The post-earthquake work of OpenStreetMap Haiti. *COOP 2014-Proceedings of the 11th International Conference on the Design of Cooperative Systems,* May 27-30 Nice (FRA). Springer, 311-326.

Stebbins, R.A., 2015. *Serious leisure: A perspective for our time.* 2nd ed. (USA): New Brunswick: Transaction Publishers.

Steinmann, R., et al., 2013. Contribution Profiles of Voluntary Mappers in OpenStreetMap. *Online proceedings of the International Workshop on Action and Interaction in Volunteered Geographic Information (ACTIVITY) at the 16th AGILE Conference on Geographic Information Science,* May 14 Leuven (BEL).

Sui, D.Z., Elwood, S. and Goodchild, M.F., 2013. *Crowdsourcing geographic knowledge.* 1st ed. New York (USA): Springer.

Sun, N., Rau, P.P. and Ma, L., 2014, Understanding lurkers in online communities: A literature review. *Computers in Human Behavior,* 38, 110-117.

Taylor, S. and Todd, P.A., 1995, Understanding information technology usage: A test of competing models. *Information systems research,* 6 (2), 144-176.

Therneau,T.M. and Lumley,T., 2017. *R Survival Package—Survival Analysis.* Fermanagh (IRL): CRAN.

Tichenor, P.J., Donohue, G.A. and Olien, C.N., 1970, Mass media flow and differential growth in knowledge. *Public opinion quarterly,* 34 (2), 159-170.

Utrilla, P., et al., 2009, A palaeolithic map from 13,660 calBP: engraved stone blocks from the Late Magdalenian in Abauntz Cave (Navarra, Spain). *Journal of human evolution,* 57 (2), 99-111.

Venkatesh, V., et al., 2003, User acceptance of information technology: Toward a unified view. *MIS quarterly,* 27 (3), 425-478.

Von Krogh, G., et al., 2012, Carrots and rainbows: Motivation and social practice in open source software development. *MIS Quarterly,* 36 (2), 649-676.

Wang, K.S., Hsu, F. and Liu, P., 2002, Modeling the bathtub shape hazard rate function in terms of reliability. *Reliability Engineering & System Safety,* 75 (3), 397-406.

Weait, R., 2011. *OSM Licence Upgrade—Phase 4 coming soon* [online]. Blogs.OpenStreetMap.org. Available from: https://blog.openstreetmap.org/2011/06/14/osm-license-upgrade-phase-4-coming-soon/ [Accessed 2016-05-08].

Wenger, E., 1998. *Communities of practice: Learning, meaning, and identity.* 1st ed. Cambridge (GBR): Cambridge university press.

Weon, B.M., 2016, Tyrannosaurs as long-lived species. *Scientific reports,* 6 (srep19554), 1-5.

Yamak, Z., Saunier, J. and Vercouter, L., 2016. Detection of Multiple Identity Manipulation in Collaborative Projects. *Proceedings of the 25th International Conference Companion on World Wide Web,* International World Wide Web Conferences Steering Committee, 955-960.

Zielstra, D. and Zipf, A., 2010. A comparative study of proprietary geodata and volunteered geographic information for Germany. *13th AGILE International Conference on Geographic Information Science,* June 4 Guimarães (PRT).

Zook, M.A., et al., 2010, Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. *World Medical & Health Policy,* 2 (2), 7-33.

Appendix A: Big Data Management and Analysis

With currently more than 4.5 million registered users (OpenStreetMap contributors 2013), the OSM project has become the largest user-led content VGI projects of the Web. All edits provided by OSM contributors are made freely available through history dump files (OpenStreetMap contributors 2014). In addition to the edits, the files include the virtual containers (i.e. changesets) in which the edits were provided. These changesets identify the contributors who submitted edits, the temporal extent of each editing session, and a minimum bounding rectangle covering all the features edited during the session.

Retrieving, manipulating and analyzing OSM data fall in the realm of the "Big Data" (Laney 2001, Gandomi and Haider 2015). The term "Big Data" refers to datasets having particular characteristics that imply unusual technological and procedural challenges. The overall process of extracting insights from big data was summarized by Gandomi and Haider (Gandomi and Haider 2015) and adequately illustrates the approach used in this research (Figure A-1).

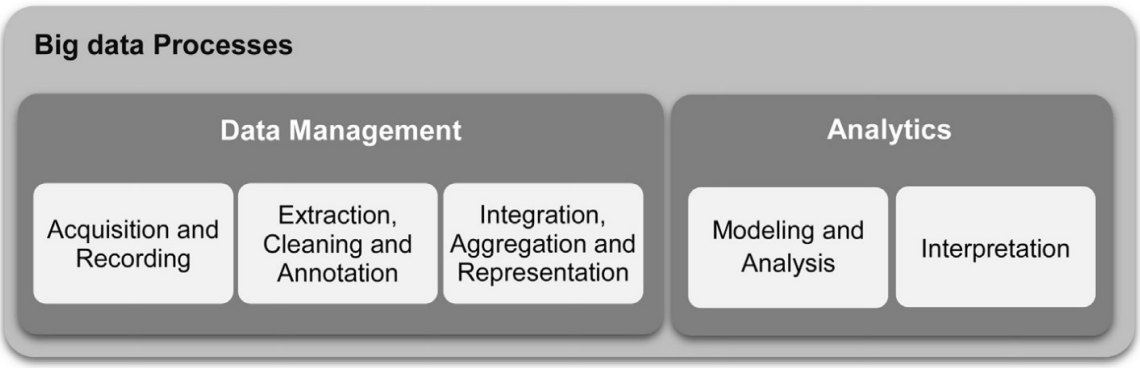


Figure A-1 Insights extraction processes from big data (Gandomi and Haider 2015).

A.1. Big data management

Many criteria have been proposed to characterize these very large datasets (Comber *et al.* 2016) but only the most frequently used are described below in relation to the data used in this research and the challenges it represented.

The first characteristic defining big data is their *volume,* which should be above one terabyte (1 TB). The history dump file downloaded from the OSM website was almost 1 TB once decompressed. After having uploaded all its content in a PostgreSQL database, the resulting tables and indexes required more than 2 TB of disk space. The largest table held more than 2.5 billion records. The "changesets" table alone contained 25 million records describing contributions from about 450 thousand participants. Two challenges were met in relation to the volume of data. The first one was to extract the information from the history file. Reading the file was time consuming and when a problem occurred, the process had to be restarted from the beginning. Big data often requires parallel processing, which cannot be done easily by reading a single file. In order to improve the file reading by using parallel processes, I have been able to split the original file into 1 GB chunks and to recover pieces of information that had been cut off by the splitting process. The second challenge was to optimize the logical and physical architecture of the database to speed up data uploading and aggregation queries. Using parallel uploading processes over multiple tables were demanding to the database because of the write access on the hard drive. Similarly, aggregation queries were going to use joins between tables which would also be demanding on the hard drive (read/write). To avoid such read/write delays, tables were eventually distributed over three external hard drives which increased

processing speed by at least a factor 10. However, even using a high-end DELL desktop computer (8 CPU, 16 GB RAM), several aggregation queries took days to complete.

The second characteristic defining big data is their *variety*. Data variety refers to the structural heterogeneity of datasets (Gandomi and Haider 2015) in which data usually do not conform to strict standards (i.e. unstructured to semi-structured data). The OSM history dump file I retrieved file was in XML format (Extensible Markup Language). XML is a textual language that may contain user-defined tags (key=value tuples) and is a typical example of semi-structured data. Transforming this semi-structured data into a tabular format that can be used by a relational database was a complex task. For instance, the tags provided by contributors have often exceeded their expected range of values, creating processing interruptions due to constraints initially imposed in the database. Interestingly, the main source of unexpected tags keys/values has found to be the OSM related applications (API, map editors), particularly over early years of project's development (e.g. empty changesets, invalid bounding boxes, objects' ID not increasing monotonically).

The third characteristic defining big data is their *velocity*. Data velocity refers to the rate at which the data is generated. For instance, hundreds of incomplete changesets were found in OSM data simply because their owners were editing at the time the history dump was closed. The history dump file used in this research covered the whole project history until September 1, 2014. Using a static history dump has prevented us from continuously updating the database. Given the resources available, the speed at which the

data are produced would not have allowed us to complete a similar study in the years that followed.

Extracting, cleaning and uploading OSM data in the database took almost half a year. However, the process could have taken about a month using 24/7 processes, without having to learn how to handle big data through trial and error, which at that time was about, and may still be, the only way.

A.2. Big data analysis

Attempting to link VGI contributors and contribution patterns through typologies is a form of predictive analysis. These analyses aim at predicting outcomes from the patterns found in historical contributions. However, in an era of big data, caution must be taken with conventional statistical tools (Singleton *et al.* 2016, Comber *et al.* 2016).

Big data statistics are most of the time conducted over whole populations, instead of samples, and they often show long tail distributions (Mitzenmacher 2004, Limpert *et al.* 2001). These long tail distributions seem to origin from the large number of very diversified sources from which the data is collected (i.e. contributors). This characteristic is referred to as the data *heterogeneity* (Gandomi and Haider 2015) because it introduces multiple dimensions in population's segmentation. Consequently, meaningful statistical parameters generally used to describe normal distribution, such as mean and standard deviation, get meaningless with these distributions. In addition, most predicting tools were built around the statistical significance of small samples and require normal (or normal

transformed) distributions. Furthermore, “long tail distribution” is a generic term that describes several types of skewed distributions that are difficult to unravel and use appropriately to normalize or to model the data in predictive analysis (Shalizi 2007, Clauset *et al.* 2009, Limpert *et al.* 2001).

Challenges emerge right from the exploratory data analysis phase. The search for explanatory variables or relationships between variables therefore required special attention and adapted techniques. For example, means and standard deviations were of little use in characterizing our data, and using histograms consistently exhibited L-shaped profiles. We therefore had to resort to quantiles and to complementary cumulative distribution functions to assess and understand the data. Similarly, the search for relationship between variables posed significant challenges. The volume of data made clogged or fuzzy most of the scatter plots I generated between variables. We then relied on density representations to decide which variables seemed to show core relationships or not. Even then, the massive size of these datasets is known to create spurious correlations, even between independent random variables (Fan *et al.* 2014).

A.3. Data analysis results

After examining dozens of variables and trying to establish relationships between them, I finally limited our choice to the following variables for each contributor.

- The date of enrollment in the project and the date of each contribution.
- Participant contribution timespan determined from the dates of the first and last contribution.
- The dates of active and inactive days throughout the duration of its contribution.

- The volume of contributions the participant provided during each active day.

The results were used to assess their potential explanatory/predictive power on contributors' behaviour. In Chapters 2 and 3, I attempted to anchor our results in the history of the project by associating events that have dotted its history to significant variations in the number of participants involved. In Chapter 4, I have associated contributor life cycle phases with the volume and frequency of their contributions.

The results I obtained initially surprised us, but a posteriori, they now seem predictable. We believe the research has provided important pieces of information that were missing from our knowledge of OSM contributors' behaviour. We think our results have shed a new light not only on OSM contributors' behaviour but potentially on that of the contributors of most collaborative projects online.

A. 4. References

Clauset, A., Shalizi, C.R. and Newman, M.E., 2009, Power-law distributions in empirical data. *SIAM Review,* 51 (4), 661-703.

Comber, A., et al., 2016. A moan, a discursion into the visualisation of very large spatial data and some rubrics for identifying big questions. *International Conference on GIScience Short Paper Proceedings, .*

Fan, J., Han, F. and Liu, H., 2014, Challenges of big data analysis. *National science review,* 1 (2), 293-314.

Gandomi, A. and Haider, M., 2015, Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management,* 35 (2), 137-144.

Laney, D., 2001, 3D data management: Controlling data volume, velocity and variety. *META Group Research Note,* 6, 70.

Limpert, E., Stahel, W.A. and Abbt, M., 2001, Log-normal Distributions across the Sciences: Keys and Clues. *Bioscience,* 51 (5), 341-352.

Mitzenmacher, M., 2004, A brief history of generative models for power law and lognormal distributions. *Internet mathematics,* 1 (2), 226-251.

OpenStreetMap contributors, 2013. *Stats* [online]. OpenStreetMap Wiki, . Available from: http://wiki.openstreetmap.org/wiki/Stats [Accessed 2012-01-21].

OpenStreetMap contributors, 2014. *Complete OSM Data History* [online]. OpenStreetMap Wiki. Available from: http://planet.openstreetmap.org/planet/full-history/ [Accessed 2014-07-03].

Shalizi, C.R., 2007. *So You Think You Have a Power Law—Well Isn't That Special?* [Online]. bactra.org. Available from: http://bactra.org/weblog/491.html [Accessed 2017-12-21].

Singleton, A.D., Spielman, S. and Brunsdon, C., 2016, Establishing a framework for Open Geographic Information science. *International Journal of Geographical Information Science,* 30 (8), 1507-1521.