Journal of Transport Geography 48 (2015) 135–144

Contents lists available at ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrg

Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data

Juha Oksanen [a,*], Cecilia Bergman [a], Jani Sainio [b], Jan Westerholm [b]

[a] Department of Geoinformatics and Cartography, Finnish Geospatial Research Institute/National Land Survey of Finland, P.O. Box 84, FI-00521 Helsinki, Finland
[b] Faculty of Science and Engineering, Åbo Akademi University, Joukahainengatan 3-5, FI-20520 Turku, Finland

ARTICLE INFO

Article history:
Received 19 December 2014
Received in revised form 3 September 2015
Accepted 5 September 2015
Available online xxxx

Keywords:
Cycling
Location-based services (LBSs)
Urban planning
Privacy
Big data
GIS

ABSTRACT

Utilization of movement data from mobile sports tracking applications is affected by its inherent biases and sensitivity, which need to be understood when developing value-added services for, e.g., application users and city planners. We have developed a method for generating a privacy-preserving heat map with user diversity (ppDIV), in which the density of trajectories, as well as the diversity of users, is taken into account, thus preventing the bias effects caused by participation inequality. The method is applied to public cycling workouts and compared with privacy-preserving kernel density estimation (ppKDE) focusing only on the density of the recorded trajectories and privacy-preserving user count calculation (ppUCC), which is similar to the quadrat-count of individual application users. An awareness of privacy was introduced to all methods as a data pre-processing step following the principle of k-Anonymity. Calibration results for our heat maps using bicycle counting data gathered by the city of Helsinki are good ($R^2 > 0.7$) and raise high expectations for utilizing heat maps in a city planning context. This is further supported by the diurnal distribution of the workouts indicating that, in addition to sports-oriented cyclists, many utilitarian cyclists are tracking their commutes. However, sports tracking data can only enrich official in-situ counts with its high spatio-temporal resolution and coverage, not replace them.

1. Introduction

Mobile sports tracking applications have become popular among the public audience, and a large number of smartphone users are willing to collect and compare their workouts privately, as well as to share their data within social networks or even publicly, for all application and Internet users. Key factors in this development have been the maturity of sensor technology, such as an accelerometer, digital compass, gyroscope, and GPS (e.g., Lane et al., 2010), available in nearly all recent mid- and top-range smartphones; and well-documented application programming interfaces for third-party developers to create new applications for mobile platforms.

The aim of our work is to develop methods to enrich workout data from a mobile sports tracking application to create privacy-preserving information about the most popular places to do sports. While our case study focuses on public cycling workouts collected using *Sports Tracker* (http://www.sports-tracker.com/), the developed methods can be used for any other sports recorded using any mobile sports tracking application. The approach chosen for this study relies on visual data mining, which utilizes human perception in data exploration, and combines human flexibility, creativity, and knowledge with a computer's storage capacity, computing power, and visualization capabilities (Keim, 2002). When integrated into a location-based service (LBS), the result of our analysis replies to the end-user's question "*Where have most cyclists continued to from here?*" In addition, we investigate the relation of tracking data and in-situ bicycle counting information in order to compare the quality of the derived heat maps, as well as to calibrate the heat maps based on mobile sports tracking application data, for example, for city planning purposes. We use the term "workout" throughout the paper to denote all recorded trajectories, be they recreational/exercise or utilitarian by purpose.

The idea of generating heat maps from mobile sports app data to communicate the popularity of sports is not new (e.g., Garmin, 2013; Lin, 2012; Strava, 2014), but less attention has been paid to the methods for making the calculation, and concerns about the appropriate understanding of the heat maps have been raised due to new application areas of heat maps, such as city planning (Maus, 2014a) and analysis of eye-tracking data (Bojko, 2009). When creating heat maps, an obvious surrogate for the popularity of sports is the density of workout trajectories, but other surrogates, such as the number of different people doing sports, can also be used. According to the limited information available on existing heat maps, the one provided by Strava uses the number of GPS points as a pixel value (Mach, 2014), whereas in the heat map offered by Nike+, the value at each pixel represents the number of users (Lin, 2012). As we will show in this paper, the two methods can locally result in very different patterns of bike riding.

* Corresponding author. Tel.: +358 40 831 4092.
E-mail address: juha.oksanen@nls.fi (J. Oksanen).

http://dx.doi.org/10.1016/j.jtrangeo.2015.09.001
0966-6923/

The use of heat maps as a representation of the intensity of a phenomenon has its roots in spectrometry (e.g., Moon et al., 2009) and the generation of isarithm and dot density maps (Slocum et al., 2009), but in the context of cartography, a textbook definition of a heat map is still missing (e.g., Trame and Keßler, 2011). Heat maps are a common visualization technique in many fields of research where large amounts of data are handled. For example, in human–computer interaction, "attention heat maps" are a popular tool of visual analysis of eye-tracking data (e.g., Blascheck et al., 2014; Bojko, 2009). Related to studies utilizing the increasing volumes of volunteered geographic information (VGI; Goodchild, 2007), heat maps have been used, for example, to reveal attractive/popular places in a region using the density-based spatial clustering of geotagged images (Kisilevich et al., 2010; Kurata, 2012) and videos (Mirković et al., 2010), or to visualize spatio-temporal patterns revealed by the distribution of tweets (e.g., Morstatter et al., 2013; Zeile et al., 2012). The coloring of a heat map is typically selected in such way that the interpretation of the intensity differences becomes intuitive. This is expected to happen when warm colors, in terms of color temperature, are used for high intensities of the represented phenomenon and cool colors for low intensities (e.g., Špakov and Miniotas, 2007).

In addition to the public audience, interest in mobile tracking applications, and enriching, especially, cycling data collected with them has emerged among city planners (Albergotti, 2014; Charlton et al., 2011; Hood et al., 2011). One of the biggest challenges in a city-planning context regarding non-motorized traffic is the lack of documentation on the usage and the demand figures for, for example, cyclists and pedestrians (Lindsey et al., 2014; NBPD, 2014). Traditional approaches for monitoring cycling traffic have been the use of surveys for qualitative results and different types of manual and automatic in-situ counting for quantitative results (Griffin et al., 2014; NBPD, 2014; Rantala and Luukkonen, 2014). Mobile tracking of cyclists has been seen as an attractive, inexpensive, and dynamic alternative to traditional bicycle data collection (Hudson et al., 2012). An early approach to tracking was the development of dedicated platforms, such as CycleTracks, which was developed at San Francisco Municipal Transportation Agency and has since been used at a number of agencies and municipalities in the US (Charlton et al., 2011; Masoner, 2014). In the UK, another crowdsourcing-based application, Cycle Hackney, is expected to provide a cost-effective way to find out where, especially, utilitarian cycling is taking place (CycleStreets, 2014). Recently, in the city of Oulu, Finland, there has been a development project aiming to create a "smoothness navigator" for cyclists, based on 1000 recorded tracks of people participating in the pilot phase (Poikola, 2014). The problem in dedicated tracking platforms appears to be the limited group of people interested in using them voluntarily (SFMTA, 2013). To overcome this problem the potential of utilizing mobile sports tracking data has been recognized, reflecting the idea of utilizing humans as sensors (Goodchild, 2007) and big data analytics (e.g., Russom, 2011). For example, Oregon's Department of Transportation paid $20,000 to use data from the mobile sports tracking application Strava for a year, containing 400,000 individual bicycle trips, totaling 8 million bicycle kilometers traveled (Estes, 2014; Maus, 2014b).

While mobile sports tracking data may not qualify as 'big data' regarding its volume – except in the sense "bigger than previously" (Goodchild, 2013) – it shares many characteristics with other social media data, often classified as big data. Many of the characteristics follow from the fact that big data is typically not collected with any specific purpose in mind or not used for its original purpose (Kitchin, 2014). In statistics, random sampling is used to guarantee the representativeness of observations, but in big data analytics, the 'sample' is not randomly chosen at all (Goodchild, 2013). Rather, the aim is to use all the data following the principle of exhaustivity in scope (Harford, 2014; Kitchin, 2014). However, considering that only a small and possibly behaviorally biased subset of cyclists use mobile applications to track their routes, the question is how well they represent the whole population of cyclists (e.g., Maus, 2014a; Rantala and Luukkonen, 2014); i.e., as social media data in general, sports tracking data is affected by self-selection bias (Shearmur, 2015). In addition, mobile tracking applications have their differences with respect to appearance and function, such as the available range of sports (multi-sports or single activity type), and may therefore attract different people. As an example, the cycling and running app Strava has the reputation of being used by more competitive or "serious" cyclists (Zahradnik, 2014) and is also targeting people who identify themselves as "athletes" (Strava, 2014). On the other hand, for example, Sports Tracker (ST, 2014) "want[s] to help people train better, connect through sports, and live healthier, happier lives;" HeiaHeia! focuses on the business-to-business sector and work welfare (Kauppalehti, 2013); and Endomondo (2014) is, in its own words, aiming "to motivate people to get and stay active." It has been estimated that 90% of the cyclists who use Strava are male (Usborne, 2013; Vanderbilt, 2013) and in 2012, 75% of all Endomondo users were men (Endoangela, 2012). Furthermore, participation inequality is a known property of VGI and online communities, according to which 90% of community members are followers and do not contribute to the community, whereas 9% contribute from time to time, and 1% account for most contributions (Nielsen, 2006). Although sports tracking applications do not today represent shared projects where people would track and share their workouts to promote the common good, some typical motivations for contribution in VGI, such as social reward and enhanced personal reputation (Coleman et al., 2009), can be identified within their communities as well. These bias issues introduce a major challenge in using mobile sports app data in a city-planning context.

According to Westin's tenet, privacy is an individual's right to have full control over information about themselves, and to decide when, how, and to what extent this information is shared with others (Agrawal et al., 2002). Guaranteeing privacy in LBSs is extremely important, due to the unique characteristics of moving object data (Fung et al., 2010; Montjoye et al., 2013; Verykios et al., 2008). Topics such as anonymization of the original dataset (e.g., Monreale et al., 2010; Pensa et al., 2008), or de-identifying a given LBS-request location (e.g., Bettini et al., 2005; Gedik and Liu, 2004; Gruteser and Liu, 2004), have gained a great deal of attention in trajectory studies but are beyond the scope of this paper. Instead, we approach privacy-preservation from the standpoint of visualization.

The idea behind preserving privacy in visualizations is to generalize or otherwise obfuscate data in such a manner that the disclosed data is still useful in the particular case (Andrienko et al., 2008; Fung et al., 2010). Various methods of geographical masking, first introduced by Armstrong et al. (1999), have been developed with the aim of protecting the confidentiality of individual locations by adding stochastic or deterministic noise to the geographic coordinates of the original data points without substantially affecting analytical results or the visual characteristics of the original pattern (Kwan et al., 2004). Spatial aggregation of individual-level data for administrative areas or other areal units that have a population greater than a chosen cutoff value is a common procedure of preserving confidentiality, for example, in censuses where disclosure control has long been an integral part of the process (Armstrong et al., 1999; Kwan et al., 2004; Leitner and Curtis, 2006; Young et al., 2009). Because aggregation can hide important spatial patterns in the data, various alternative geo-masking techniques, such as random perturbation and affine transformation (translate, rotate, and scale), have been introduced to preserve the disaggregated, discrete nature of the original data (Armstrong et al., 1999; Kwan et al., 2004). Although they have been used mainly with geo-referenced, sensitive health- and crime-related point data (e.g., Leitner and Curtis, 2006; Kounadi and Leitner, 2015), Krumm (2007) and Seidl et al. (2015) have applied them also to GPS trajectory data. In this study, where it was crucial to prevent re-identification of an individual user and trajectory while providing the heat map viewer accurate information about popular cycling paths in their actual locations on the road network, geo-masking techniques as such were, however, not