Wrocław University
of Environmental
and Life Sciences

Assessing the impact of data processing methods on bias in human mobility science

Kamil Smolak

Supervised by Prof. Witold Rohm

Co-supervised by Dr. Katarzyna Siła-Nowicka

Institute of Geodesy and Geoinformatics

The Faculty of Environmental Engineering and Geodesy

Wroclaw University of Environmental and Life Sciences

March 21, 2022

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy.

Abstract

Without a doubt, mobile devices have impacted society and facilitated many aspects of life, starting as a means of communication and reaching the role of all-in-one powerful computers. Their popularity and high population penetration rate made mobile devices also a great source of geolocated information, providing data on millions of people at almost no cost. Further development of various communication protocols only increased the volume of data collected on human movement, creating new data sources. The abundance of human location data caused the rapid development of human mobility science, transforming it to its new, data-driven form, in the formation of which Geographical Information Science (GIS) played a crucial role. This field of science is highly interdisciplinary. Spatio-temporal characteristics of mobility data involve researchers and analysis methods from GIS but also combines knowledge and solutions from, physics, computer science, and geography.

The utility of human mobility data was proven in many areas, ranging from population distribution estimation, through traffic congestion analyses, to disease spread modelling. Studies of human mobility resulted in many novel algorithms focused on deriving useful information from human mobility data. However, the approach to the extraction of basic information, that is times of visits and important locations in a movement trajectory, is mostly the same across human mobility studies. These data processing methods are controlled by parameters (e.g. temporal sampling interval and magnitude of spatial aggregation), that are not neutral to the following analysis. The change of parameters will also have a profound impact on the conclusions. Human mobility data have high spatio-temporal resolution and are noisy, which increases the probability that inattentive processing will introduce biases into analyses. Such problem is known in areas where spatial and spatio-temporal analyses are involved under the name of Modifable Aeral Unit Problem (MAUP) and Modifiable Temporal Unit Problem (MTUP). It was also noticed in early data-driven works on human mobility. Despite the efforts, an ultimate solution to avoid this bias have not been yet found.

The effect of MAUP and MTUP can be observed in human mobility predictions. In 2010, a work deriving a methodology for mobility predictability limit estimation was published. It measured that human movement is predictable in 93%. This work was followed by numerous studies implementing this theory to their datasets. This limit was found to vary highly across these experiments, ranging from 43% up to 95%. Investigation of these studies reveals that the magnitude of spatio-temporal aggregation in experimental datasets correlates with the value of predictability limit. Later studies confirmed that relationship. This leads to the conclusion that the accuracy of prediction algorithms can be inflated or deflated depending on the parameters of data processing methods, which influence the spatio-

temporal aggregation of mobility data. Moreover, some works indicated contrary results, where sophisticated prediction algorithms surpassed the theoretical limit. This suggests that the predictability limit theory is erroneous.

The goal of this PhD dissertation is to investigate the effects of MAUP and MTUP on the spatio-temporal movement data, with a special focus on human mobility prediction tasks. Specifically, this work analyses the impact of MAUP and MTUP on the predictability of human mobility and identifies the sources of its variations. Furthermore, it validates and confirms the discrepancies between the predictability limit theory and mobility prediction task. To address that issue this work provides a novel pattern-matching-based metric that explains a majority of variability in mobility prediction accuracy. This metric also allows for quick estimation of the potential predictability of a movement trajectory, therefore it is proposed as an alternative to the predictability limit metric. The outputs of this work aim to advance GIS through understanding the significance of MAUP and MTUP presence in spatio-temporal analytics. A general goal of published studies is also to give an overview of human mobility data processing methods, types of next location prediction tasks and mobility modelling approaches.

Keywords: human mobility, movement trajectories, bias, MTUP, MAUP, prediction, modelling

Streszczenie

Zaczynając jako środek komunikacji i osiągając rolę potężnych komputerów, urządzenia mobilne bez wątpienia wpłynęły na społeczeństwo i ułatwiły wiele aspektów życia. Popularność i wysoki współczynnik penetracji populacji sprawiły, że urządzenia mobilne są również znakomitym źródłem danych geolokalizowanych, dostarczającym informacji o milionach ludzi niemalże bez żadnych kosztów. Dalszy rozwój różnych protokołów komunikacyjnych tylko zwiększył objętość gromadzonych z nich informacji. Obfitość danych o lokalizacji ludzi spowodowała szybki rozwój nauki o mobilności ludzi, przekształcając ją w jej nową formę wysoce opartą na danych, w kształtowaniu której Systemy Informacji Przestrzennej grały kluczową rolę. Obecnie, ma ona charakter wysoce interdyscyplinarny. Ze względu na czasowo-przestrzenne właściwości danych mobilnych, dziedzina ta oparta jest w dużej mierze na rozwiązaniach z obszaru Systemów Informacji Przestrzennej, ale adaptuje ona również wiedzę z obszaru fizyki, informatyki oraz geografii, która ma długą tradycję analizowania przemieszczeń ludności.

Użyteczność danych o mobilności ludzi została udowodniona wielokrotnie, poczynając od estymacji gęstości zaludnienia, przez analizy natężenia ruchu, po modelowanie rozprzestrzeniania się chorób zakaźnych. Badania nad mobilnością ludzi zaowocowały wieloma nowatorskimi algorytmami, które dedykowane są ekstrakcji użytecznej informacji z danych o mobilności. Jednakże, podejście do pozyskania podstawowych informacji, takich jak lokalizacja istotnych miejsc w trajektorii ruchu i czasy wizyt w nich, są takie same w większości badań. Te metody przetwarzania danych są kontrolowane przez parametry (na przykład częstotliwość próbkowania i wielkość agregacji przestrzennej), które wpływają na dalsze analizy. Zmiany tych parametrów mają też duży wpływ na wyciągnięte wnioski. Dane o mobilności ludzi mają wysoką rozdzielczość czasowo-przestrzenną i często zawierają błędy pomiarowe, co zwiększa prawdopodobieństwo, że ich nieuważne przetwarzanie wprowadzi obciążenie statystyczne do analiz. Taki problem znany jest jako problem zmiennej jednostki odniesienia przestrzennej (MAUP) i czasowej (MTUP). Występuje on w analizach gdzie stosowane są metody przetwarzania danych czasowo-przestrzennych. Został on też zauważony już we wczesnych pracach analizujących dane o mobilności ludzi. Pomimo wysiłków, ostateczne rozwiązanie pozwalające na uniknięcie tych problemów nie zostało znalezione.

Wpływ MAUP i MTUP może być zaobserwowany w również predykcjach mobilności. Opublikowana w 2010 roku praca opracowująca metodykę estymacji granicy przewidywalności, oszacowała, że ruch ludzi jest przewidywalny w 93%. Metodyka ta została wielokrotnie zastosowana w późniejszych pracach do oszacowania przewidywalności innych zbiorów danych. Okazało się, że wartość tej granicy bardzo różniła się pomiędzy tymi pracami, wahając się od 43% do 95%. Analiza tych badań ujawnia, że wielkość agregacji czasowo-przestrzennej

zbiorów danych koreluje z wartością limitu przewidywalności. Późniejsze badania potwierdziły ten związek. Prowadzi to do wniosku, że dokładność algorytmów predykcyjnych może zostać zawyżona lub obniżona w zależności od parametrów metod przetwarzania danych, które wpływają na ich czasowo-przestrzenną agregację. Ponadto, niektóre prace wykazują sprzeczne rezultaty, kiedy to zaawansowane algorytmy predykcyjne przekraczają teoretyczny limit przewidywalności. Sugeruje to, że teoria granicy przewidywalności jest błędna.

Celem niniejszej rozprawy doktorskiej jest zbadanie wpływu MAUP i MTUP na naukę o mobilności ludzi. W szczególności, praca ta analizuje wpływ MAUP i MTUP na przewidywalność mobilności ludzi oraz określa źródła jej zmienności. Ponadto, praca ta bada i potwierdza rozbieżności pomiędzy teorią granicy przewidywalności oraz zadaniem predykcji ruchu. Niniejsza praca dostarcza nową miarę opartą na dopasowaniu wzorców, która wyjaśnia większość zmienności w dokładności przewidywania mobilności. Miara ta pozwala również na szybką estymację potencjalnej przewidywalności trajektorii ruchu, a zatem stanowi ona alternatywę do opublikowanej wcześniej granicy przewidywalności. Rezultaty tej pracy mają zapewnić rozwój nauki of danych przestrzennych poprzez lepsze zrozumienie znaczenia MAUP i MTUP w analizach przestrzenno-czasowych. Ogólnym celem tej pracy jest również przedstawienie przeglądu metod przetwarzania danych o mobilności ludzi, różnych podejść do zadania przewidywania następnej lokalizacji oraz do modelowania mobilności.

Słowa kluczowe: mobilność ludzi, trajektorie ruchu, obciążenie statystyczne, MTUP, MAUP, predykcja, modelowanie