Denne artikel er publiceret i det elektroniske tidsskrift
Artikler fra Trafikdage på Aalborg Universitet
(Proceedings from the Annual Transport Conference at Aalborg University)
ISSN 1603-9696
www.trafikdage.dk/artikelarkiv

Den lange vej til deling af anonymiserede Floating Car Data

Pelle Rosenbeck Gøeg, prg@civil.aau.dk, Aalborg Universitet
Irma Kveladze, Aalborg Universitet
Rasmus Øhlenschlæger, Aalborg Universitet
Harry Lahrmann, Aalborg Universitet
Niels Agerholm, Aalborg Universitet

Abstrakt

Flere års forskningsprojekter, hvor Floating Car Data (FCD) er en essentiel del af observationerne, viser at der er en stor værdi i disse. Desværre er erfaringen, at projekterne når at afsluttes før deres potentiale er fuldt udnyttet. Ved at åbne for adgangen til disse data, kan potentialet forhåbentligt realiseres. Før der kan gives adgang, skal centrale informationer relateret til køretøjet og dennes ejer i datasættet anonymiseres, så den videre brug ikke afslører individers færdselsmønstre, da dette er et brud på den enkeltes privatliv. FCD indsamlet i det Nordjyske projekt ITS Platform, er anonymiseret ved en metode der modificerer tidspunktet for afgange, og en trimning af enderne i turene.
I alt er FCD fra 389 biler anonymiseret, svarende til 0,7 milliarder positioner og en turlængde på 9,7 millioner kilometer. De anonymiserede FCD afspejler de oprindelige kørselsmønstre på vejnettet med undtagelse af småture. Antallet af småture ture er reduceret med 27 %, mens de lange ture er bevaret, hvilket har forøget den gennemsnitlige turlængde med 30 %.

Introduktion

I løbet af de sidste par årtier er der sket en kraftig udvikling af projekter og løsninger, der leverer floating car data (FCD) (1,2). I flere projekter, er der indsamlet en betydelig mængde FCD og der indsamles fortsat flere. Den primære anvendelse af FCD i projekterne er typisk opfyldt. Men da omfanget af FCD og analysemulighederne er store, er det sjældent at dataindsamleren selv når at lave en udtømmende analyse af de indsamlede FCD (2). De fleste datasæt er således kun i begrænset omfang blevet analyseret og andres perspektiver på dem kunne have værdi. Derfor ønskes det at give adgang til et indsamlet FCD sæt til tredjeparter (3,4).

Mens nogle analyser i FCD er lavet i de forskellige projekter, så har de enkelte projekter typisk et ensidigt fokus på, hvad disse data skal bruges til. Det indsamlede datamateriale indeholder derfor viden, der ofte ikke udnyttes. Et eksempel kunne være de mange studier af intelligent hastighedstilpasning (5,6), mens få har haft fokus på effekterne af transporttid (7,8). På samme måde, er en del studier lavet for at teste en specifik hypotese, men uden at anden viden i data er blevet brugt (9).

Endvidere har kun få analyser af data kunnet vise resultater fra enkeltture og den værdi, der ligger i viden om kørselsvariation blandt bilisterne og i den enkeltes kørselsadfærd.

I mange tilfælde er denne manglende udnyttelse grundet i mangel på ressourcer eller en undervurdering af tiden det kræver at lave selve analysearbejdet. Omkostningerne for indsamlingen af FCD er høj og tidskrævende. Med implementeringen af EU Persondataforordningen i maj 2018, er det yderligere blevet vitalt, at beskyttelse af eventuelle personfølsomme oplysninger skal sikres og at disse oplysninger skal være beskyttet og kun gemt i det omfang, der er passende til formålet og kun deles, når der er taget tilstrækkelige skridt for at beskytte personfølsomme oplysninger. Hertil skal det understreges, at hvis indsamlede data skal anonymiseres, så de ikke længere er personfølsomme, betyder det, at hverken den dataansvarlige eller databehandleren kan forbinde dem til personfølsomme oplysninger f.eks. ved at genskabe identiteten på føreren via en analyse af data. Kan føreren ikke identificeres er FCD ikke længere at betragte som personfølsomme og kan principielt deles med tredjepart (10, 11, 12).

Et anden perspektiv har været, at alt data, der er indsamlet for offentlige midler, bør være tilgængelige uden omkostninger, bortset fra de, som knytter sig til transaktionen (13). I praksis er denne datadeling en sjældenhed, da få organisationer er villig til at dele deres data. Her er GDPR, korrekt eller ej, en central del af forklaringen på manglende deling.

For at imødekomme udfordringen med at FCD kan være personfølsomme, er det et mål med denne artikel, at dokumentere en metode der anonymiserer FCD, så de kan stilles til rådighed for tredjepart. FCD indsamlet over 2,5 år (2012-2014) fra ITS Platform projektet er anvendt til at teste metoden.

I forbindelse med anonymiseringsarbejdet, er det blevet klart, at tidligere præsenterede principper for anonymisering af FCD som demonstreret i (2,3,10,17) ikke sikrer anonymisering tilstrækkeligt godt. En anonymisering af FCD er ikke en helt simpel opgave.

Det første forsøg med anonymisering af ITS Platforms FCD gennemført 2016-2017 mislykkedes, fordi kompleksiteten i arbejdet blev undervurderet (4).

Denne artikel beskriver en metode, der vurderes at svare ja til nedenstående spørgsmål:
Kan FCD offentliggøres, hvor en stor del af den indlejrede viden gemmes, samtidig med at anonymiseringen af de deltagende bilisters kørselsdata sikres tilstrækkeligt?

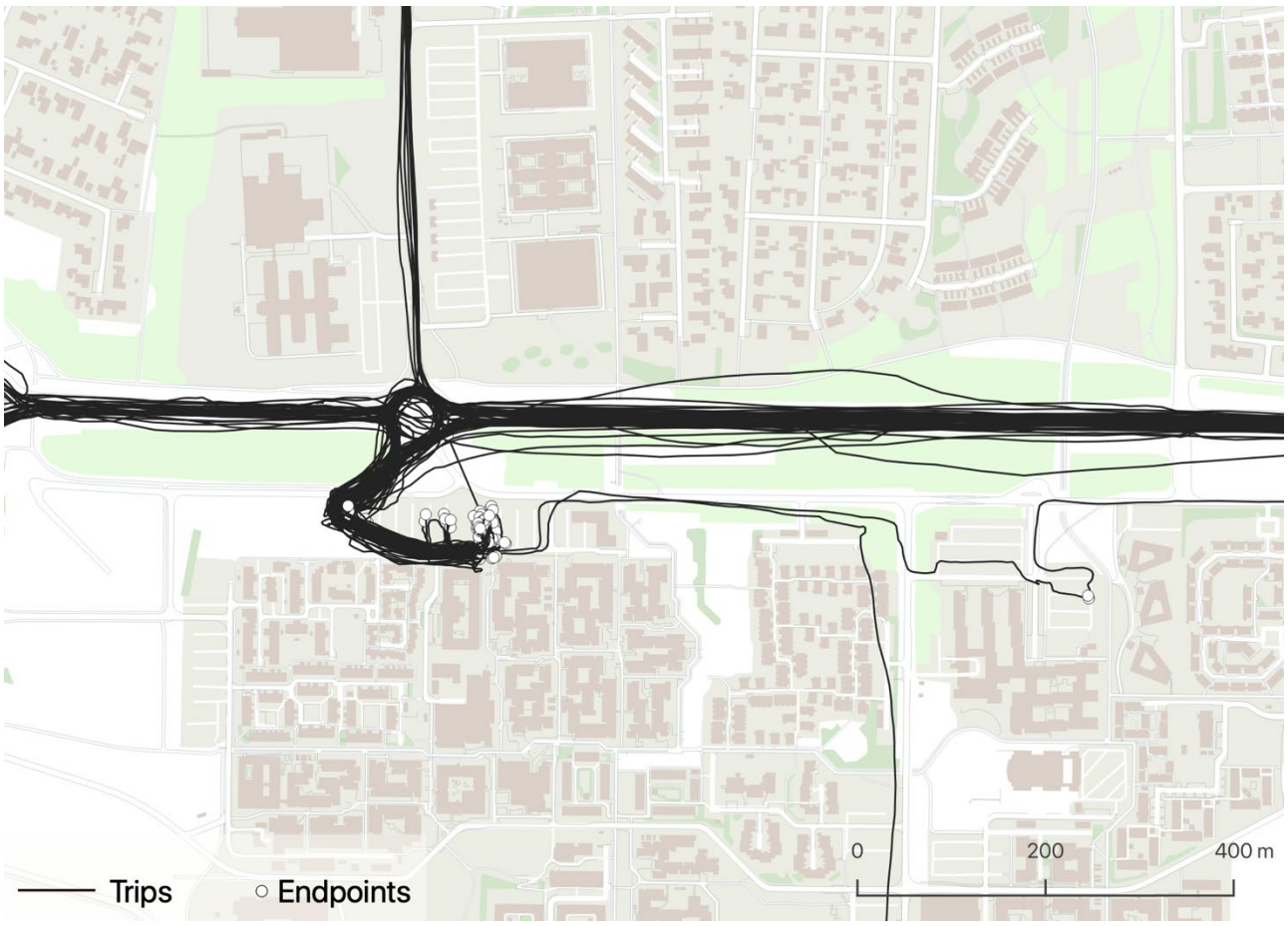
Metode

Et enkelt blik på et køretøjs rumlige færden kan hurtigt afsløre førerens identitet og afsløre dennes privatliv. Både start- og slutpunkter vil afsløre personens interessepunkter (POI) som bopæl og arbejdsplads. Derfor er det nødvendigt med en sløring af disse informationer. En simpel afkortning af køreturene, vil ikke være tilstrækkelig sløring, da POI med lethed vil kunne genskabes. I stedet foreslås følgende metode:

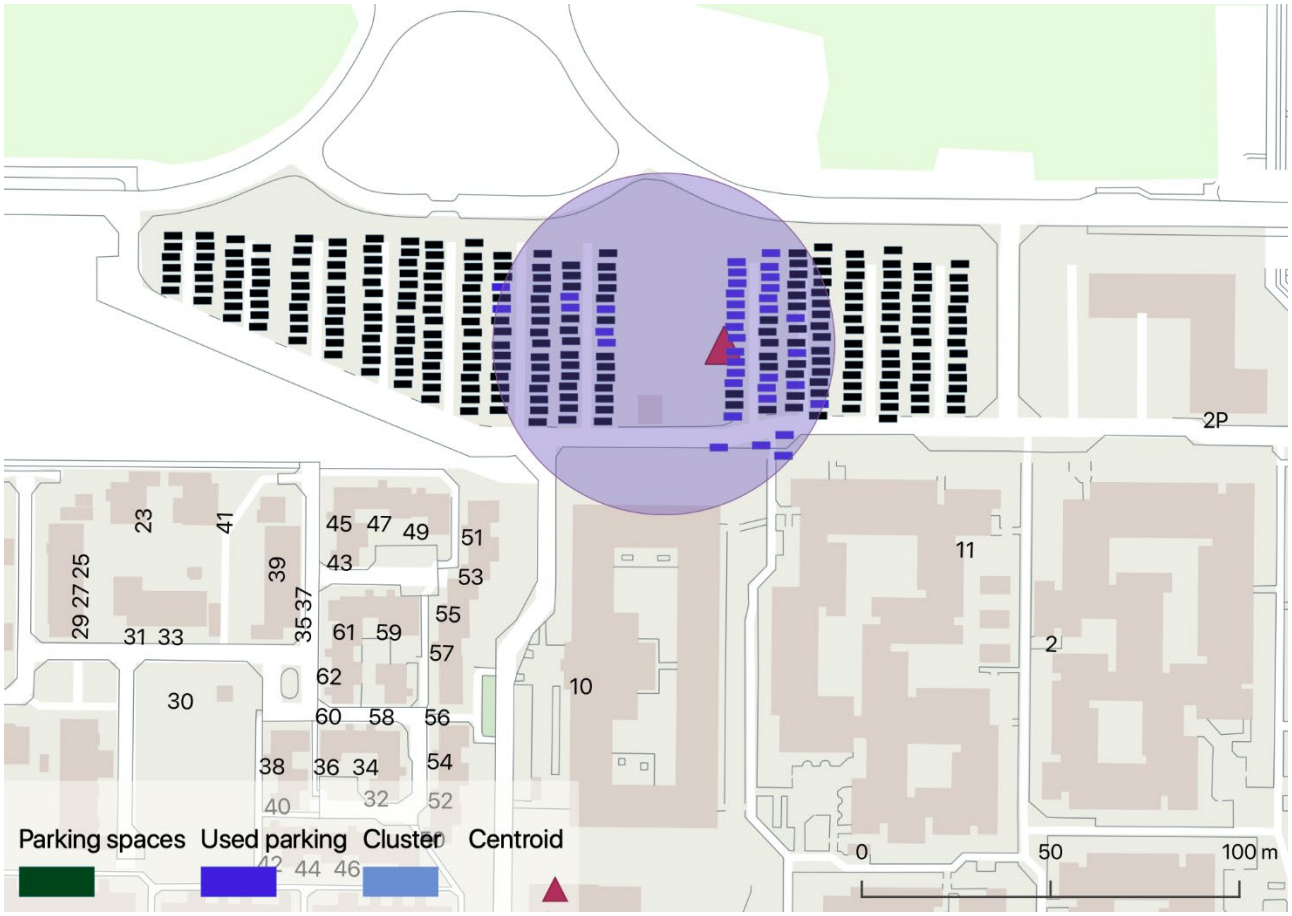
For hvert enkelt køretøj; Først identificeres alle endepunkter for alle ture. Dernæst grupperes endepunkterne i en klynge. En cirkelbuffer skabes omkring klyngen, hvorefter et tilfældigt punkt vælges som centrum for en ny cirkelbuffer, med en radius så stor, at bufferen som minimum inkluderer alle endepunkterne. Alle FCD relateret til endepunkterne i klyngen afkortes så i forhold til bufferens afgrænsning. Til sidst fjernes informationer om køretøjet og turenes begyndelsestidspunkt grupperes i forskellige tidsperioder.

Algoritmen

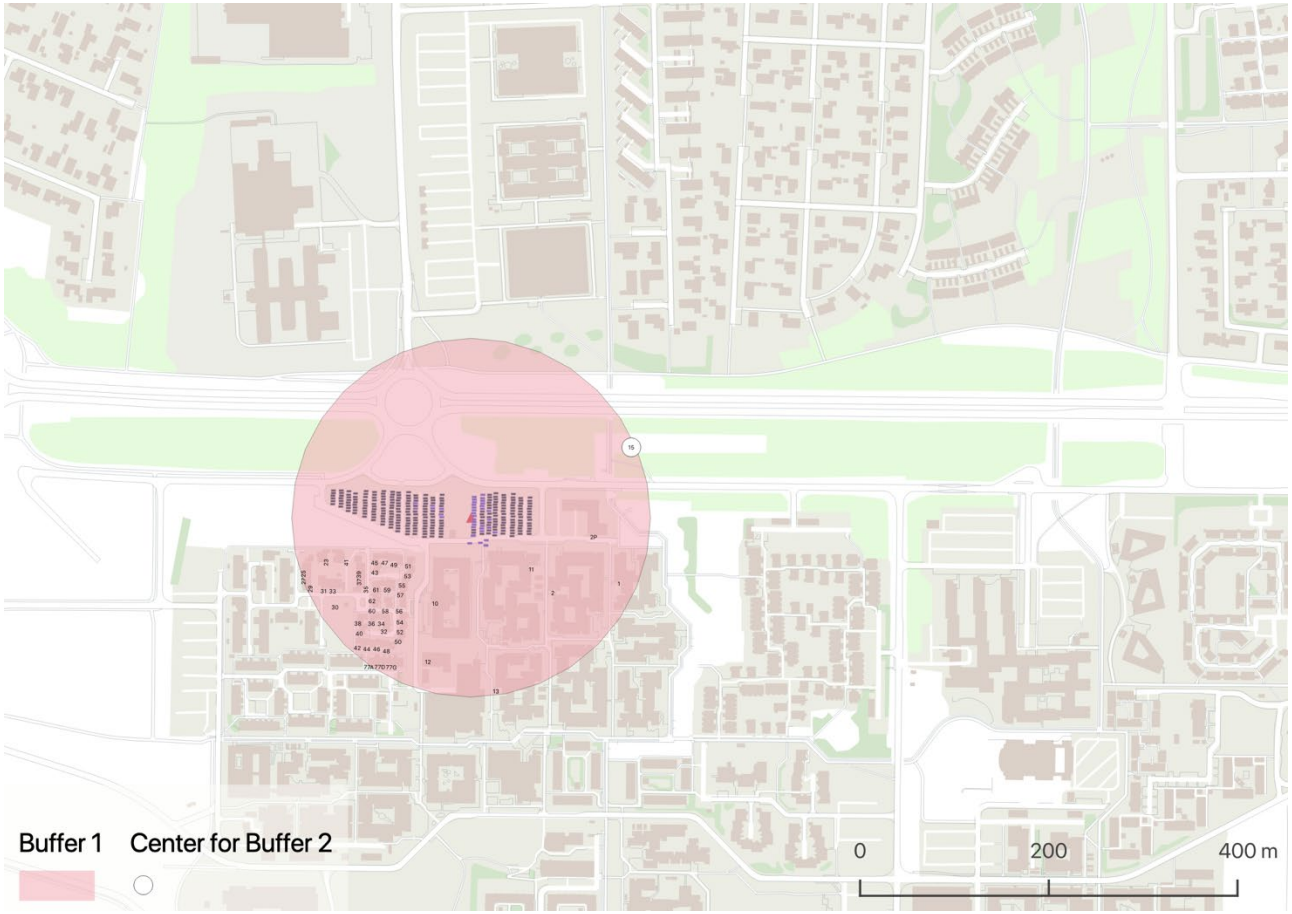
1. Alle FCD aggregeres til ture i vektorformatet Linestring ZM, hvor Z værdien er vores punkt_id og M er positionens tidstempel. Positionerne forbindes på baggrund af tidstemplet og identiteten af det enkelte køretøj. Typiske udfald på 1-2 sekunder ignoreres, mens udfald større end 120 sekunder ses som en afsluttet tur (se figur 1).
2. FCD mapmatches til OpenStreetMap (18).
3. Endepunkter for hvert køretøjs ture identificeres.
4. Endepunkterne repræsenterer, hvor en tur begynder eller slutter. Parkeringen antages at ske tæt på en destination, der kan være personfølsom og udpeges derfor til en POI. Køretøjet parkerer ofte forskellige steder i et område, og endepunkterne klynges derfor sammen. Til dette anvendes DBScan algoritmen (19). Den maksimale indbyrdes afstand mellem endepunkter sættes til 50 meter. Afstanden er valgt, da det ud fra erfaring virker som en fornuftig afstand (se figur 2).
5. En minimumscirkel beregnes om klyngen af punkter (buffer 1), så den indeholder mindst 50 adresser. Derved bliver størrelsen af bufferen større eller mindre, i forhold til tætheden af adresser. Ved meget tyndt bebyggede områder vælges en radius på 2.000 m. Buffer 1's centrum er klyngens centroide. (se figur 3)
6. Af de opnåede adresser, i punkt 5, udvælges én tilfældigt som et nyt centrum for en ny buffer (buffer 2). Buffer 2 har en radius hvis størrelse er afstanden fra buffer 2's centrum til buffer 1's fjerneste periferi. Alle FCD tilhørende den aktuelle klynge jf. punkt 4, fjernes fra buffer 2's område. Således vil ture gennem klyngen bibeholdes, mens FCD tilhørende klyngens ture fjernes (se figur 4 og 5).
7. For at sløre turenes afgangstid, grupperes turenes starttid i følgende perioder; myldretid (07-09 og 14-17), dag (9-14), aften (17-22) og en free flow periode (22-07). På samme måde, simplificeres de til første hverdag, henholdsvis første weekenddag i datoformatet.
8. Informationerne om turenes tilhørsforhold i forhold til det unikke køretøj fjernes. Turene får et unikt id, uafhængigt af de pågældende køretøjer, så positionerne indsamlet i den enkelte tur er kædet sammen.
9. Forskellen mellem data før og efter anonymiseringen er testet ved at sammenholde længden på turene før og efter og ved at analysere mapmatchingen.



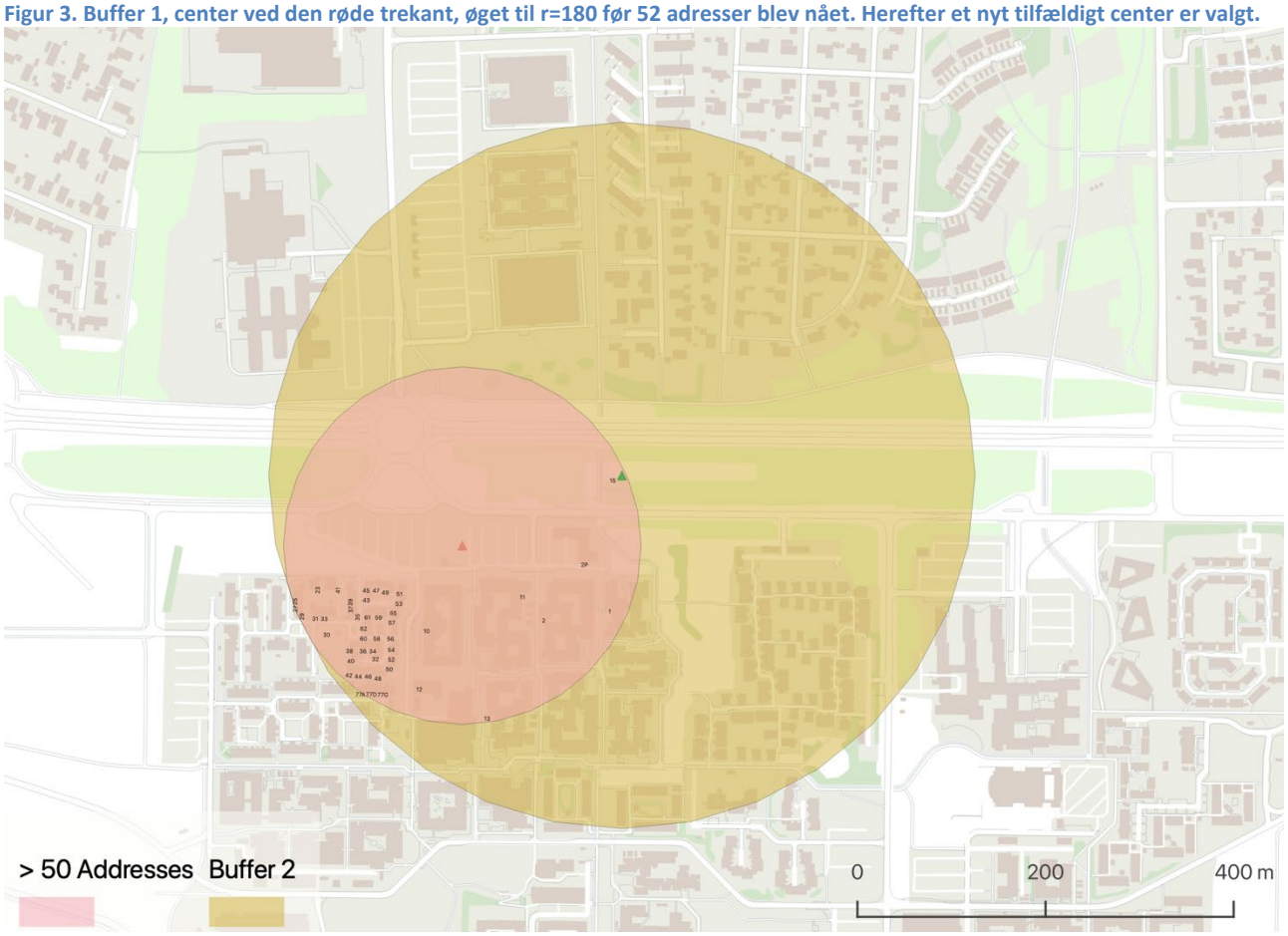
Figur 1. Parkeringsplads med start og stop.



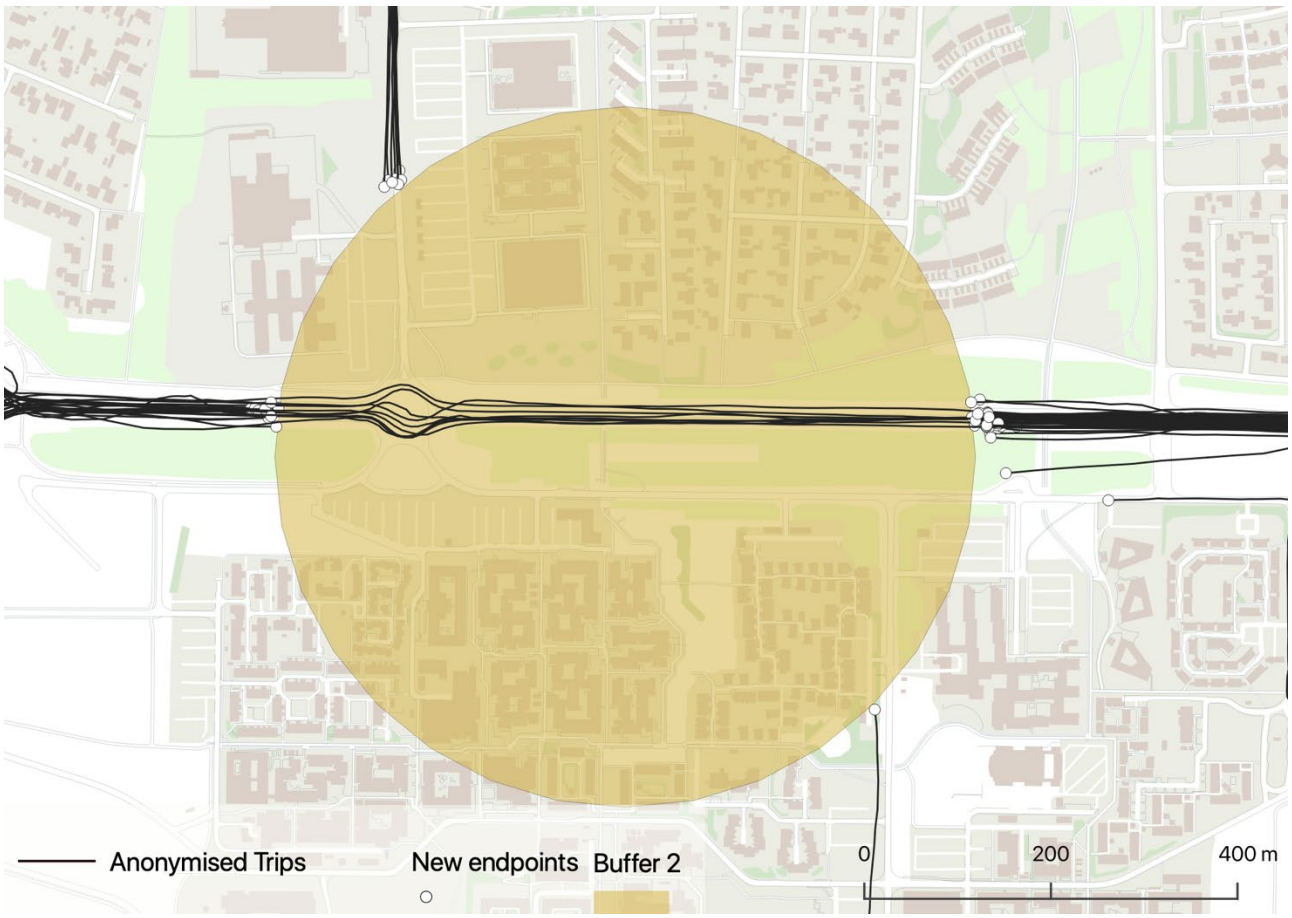
Figur 2. Flere parkeringsmuligheder. De blå markeringer er parkering tilhørende ét køretøj. Den blå cirkel beskriver den i punkt 5 beskrevne mindste cirkel.



Figur 3. Buffer 1, center ved den røde trekant, øget til r=180 før 52 adresser blev nået. Herefter et nyt tilfældigt center er valgt.



Figur 4. Buffer 2 er skabt, og indeholder buffer 1.



Figur 5. Anonymiseret ture, POI kan ikke identificeres da centrum ikke er POI. Således vil en cirkel med tre eller flere kendte punkter ikke resultere i genskabt POI.

Udførelse

Til arbejdet er der skrevet en række programmer i python3.6 til at foretage anonymiseringen. Derudover, er Meili biblioteket i Valhalla (20), en mapmatching algoritme på OpenStreetMap dækkende Danmark (18) brugt til mapmatching og PostgreSQL/Postgis 2.5 er brugt til lagring spatiotemporale analyser på FCD.

Den ovenstående metode adskiller sig fra tidligere beskrevne metoder, da datavolumen er væsentlig større og nødvendigheden af en vandtæt anonymisering på baggrund af GDPR-krav skal opfyldes.

I litteraturen skelnes der grundlæggende mellem ”offline” og ”online” anonymisering.

Jensen et al. (17), der beskriver en offline metode, som nærværende artikel, foreslår kun at fjerne FCD i det faste 2 km x 2 km kvadrat, med størst aktivitet uden for arbejdstid. Dette fjerner for meget FCD i tætte områder og potentielt for lidt i yderområder. Deres arbejde er dog foregået på et meget mindre datasæt (ca. 530.000 punkter), indsamlet over seks uger, hvilket betyder langt færre ture pr. køretøj. Det betyder, at det enkelte køretøjs færdsel ikke i samme grad blotlægger bilistens adfærd, som var data indsamlet over længere tid og dermed flere ture.

Derudover har andre forsøg på anonymisering af FCD haft fokus på hele dataflowet, fra opsamlingen til færdig databehandling – potentielt i bilens computer eller lignende. Den proces vil i mange tilfælde foregå med mindre computerkraft, og derfor kræve en simplere og mindre dataintensiv metode, jf. (21).

Som en konklusion på valget af metode, så har det været højeste prioritet af få en god anonymisering af et stort FCD-sæt, samtidig med at anonymisering er så skånsom som muligt mod FCD, samt kan laves i en opsætning, der kræver mindst muligt manuel behandling.

Data

FCD, der anonymiseres, er indsamlet i ITS Platform projektet fra 2012-2014 (4,14,15). ITS Platform projektet fungerede ved at en enhed blev installeret i hvert køretøj. Denne sendte FCD til en backendserver i realtid. 430 person- og varebiler, primært privatejede leverede FCD til projektet. FCD blev opsamlet med 1 Hz og bestod bl.a. af følgende attributter: id, position, tidstempel, retning og hastighed (14). De centrale karakteristika for FCD er beskrevet i Tabel 1. Datasættet består af ca. 1,3mia. positioner svarende til omkring 10 mio. kørte km.

Tabel 1. Attributter i det anonymiserede datasæt.

Attribut	
Trip_id	En UUID streng, der er unik for hver tur. En tur er defineret ved at der er < 2 minutter mellem punkterne. Trip_id har ingen forbindelse til andre oplysninger.
Timestamp	Tidstempel, hvor hverdage er set under ét og tilsvarende for weekender. Tiden er inddelt i fire perioder over døgnet, repræsentativ for forskellige typer trafik.
Geometry	Projekteret til UTM systemet med datum wgs84.
EPSG	EPSG kode for UTM systemet.
Speed	Målt hastighed i punktet baseret på den indbyrdes afstand mellem observationer med 1 Hz interval.
Direction	Kørselsretning.

Resultater

Sammenlignes de originale og anonymiserede FCD, ses det jf. tabel 2, at 28 % af FCD er fjernetog at den gennemsnitlige turlængde er steget med 30 %.

Tabel 2. Karakteristika for de originale og anonyme FCD for en tilfældig bil og det samlede datasæt.

	Tilfældigt køretøj			Total datasæt		
	Før	Efter	Ændring	Før	Efter	Ændring
Vehicles	1	1		389	389	0%
Trip counts	3.114	2.422	-22%	741.559	541.910	-27%
Length (km)	42.298	36.938	-13%	9.706.929	9.242.218	-5%
Point counts	2.233.133	1.615.489	-28%	569.222.165	408.338.433	-28%
Max. points removed	-	1.383		-	27.395	
Avg. point removed	-	210		-	177	
Avg. trip length (km)	13,6	15,3	+12%	13,1	17,1	30%
Max. trip length (km)	421	416	-1%	832	821	-1%

Adgang til data

De anonymiserede data vil blive tilgængelig for offentligheden via en RESTful Web Service (webservice, se https://fcd-share.civil.aau.dk/ for flere detaljer). Adgang gives til alle, der oplyser navn, email, organisation og formål med brugen af data. Samtidig accepteres det at publicerede artikler på baggrund af disse data, citerer *Anonymised Floating Car Data – the long path to data sharing* (22). Adgangen kontrolleres via en adgangsnøgle. Via webservicen vil brugeren kunne forespørge i datasættet for specifikke dele. Webservicen leverer som udgangspunkt data i GeoJSON formattet. De specifikke dele er blandt andet; interval for tid, dag eller weekend, geografisk område og OpenStreetMap vej identitet. For eksempel vil https://fcd-share.civil.aau.dk/points?when=weekend&osm_id=8149020 returnere alle weekend FCD mapmatchet på OpenStreetMap vej identitet 8149020, hvilket er id for stykket af Universitetsboulevarden nord for AAU Hovedcampus og fremgår af figur 1-5.

Diskusion og sammenfatning

Diskusion

Anonymiseringen af FCD er en balancegang mellem, hvor meget eller lidt information, der skal fjernes, før førerens privatliv ikke kompromitteres, mens værdien af FCD så vidt muligt skal bevares.
Fjernes for lidt information, vil kompromitteringen være mulig, f.eks. kunne det ske ved en simpel identifikation af bilejerens bopæl og arbejdssted. Hvis FCD fjernet i et fast radius fra POI, vil det for lokaliteter med mere end to udkørsler nemt kunne fastlægges, da de nye endepunkter alle vil ligge på en cirkels periferi, hvor centrum er POI.

På samme måde vil en bevarelse af det originale tidsstempel, gøre det muligt at se et reelt turmønster og dermed afsløre potentiel færdsel, der kan kompromittere førerens privatliv. Hvis en pseudonymisering fandt sted ved en bevarelse og omkodning af køretøjets id, ville tre års færdsel nemt kunne sammenstilles og blotlægge personfølsomme oplysninger.

Ved valg af en fast radius uden at tage højde for adressetætheden, vil endnu et stort antal små ture blive fjernet. En fuldstændig sletning af tid, vil gøre analyser, hvor andet end lokaliteten er interessant, umulig. Værdien af FCD vil således blive stærkt reduceret.

Resumé

Datasæt med FCD er i de seneste årtier blevet indsamlet i flere storskalaforsøg. Mens nogle af de initierende analyser er lavet, er den store mængde information og værdi i FCD i de fleste tilfælde ikke udnyttet fuldt ud. Intentionen i både EU- og national lovgivning er, at data indsamlet for offentlig midler, så vidt muligt skal offentliggøres, uden yderligere omkostninger end den der lægger i omkostningerne til at facilitere delingen. Indenværende arbejde har til formål, at dele FCD indsamlet i ITS Platform projektet i anonymiseret form. Data fra 422 biler samlet over tre år (2012-2014) er anonymiseret ved: 1. Fjernelse af FCD i en dynamisk radius fra et interessepunkt; 2. Radius er sat i forhold til af adressetætheden; 3. Tidstemplet er erstattet med tid i fire perioder på dagen, samt hverdag/weekend; 4. Fjernelse af informationer, der kan sammenstille ture til det enkelte køretøj.

Denne procedure reducerede antallet af ture med 27 %, og de tilbagevendende tures gennemsnitlige længde i tid er forkortet fra 12:28 til 11:54 minutter, mens deres gennemsnitlige kørselslængde er forøget fra 13,1 til 17,1 km – sidstnævnte fordi de korteste ture er fjernet.

Med FCD inddelt i tidsperioder i løbet af dagen, kan de anonymiserede FCD stadig analyseres og kørselsmønstre kan identificeres. Mens de længste ture næsten er uændrede, så er de kortere ture blevet ændret betydeligt.

De anonyme FCD vil snarest blive tilgængelige via https://fcd-share.civil.aau.dk.

Tak til

Vi vil gerne takke Den Danske Innovations Fond via DiCyPS centret så vel som den Europæiske Unions Forskning og Innovations Fond, FP7, for deres tidligere støtte via FOT-Net dataprojektet. Derudover takkes det oprindelige projekt, ITS Platform, så vel som den Europæiske Regionale Udviklingsfond og Region Nordjylland for støtte. Til sidst skal det bemærkes, at Anita Grasers blog (23) ”Movement data i GIS” har givet væsentlig inspiration til dette projekt.

Litteraturliste

1. Jamson, S., Carsten, O., Chorlton, K., Fowkes, M. (2006). *Intelligent speed adaptation - Literature Review and Scoping Study*. The University of Leeds and MIRA Ltd.

2. FOT-Net. (2017). *FOT Catalogue* [Internet]. [cited 2019 Jan 2]. Available from: http://wiki.fot-net.eu/index.php/FOT_Catalogue
3. Gellerman, H., Svanberg, E., Kotiranta, R., Heinig, I., Val, C., Koskinen, S., Innamaa, S., Zlocki, A., Bakker, J. (2017). *FOT-Net Data Field Operational Test Networking and Data Sharing Support*. FOT-Net Data.
4. Kveladze, I., Agerholm, N., Lahrmann, H.S. (2017). Opening up Danish FCD - Exploration of movement FCD data: A case study of GeoVisual analytics for four-legged intersections. In Proceedings *12th ITS European Congress*, Strasbourg. ERTICO (ITS Europe), pp. 1-13.
5. Adell, E., Várhelyi, A., Hjälmdahl, M. (2008). Auditory and haptic systems for in-car speed management - A comparative real life study. *Transportation Research Part F: Traffic Psychology and Behaviour*. vol. 11 issue 6. pp. 445–458.
6. Adell, E. Várhelyi, A., Alonso, M., Plaza, J. (2010). Developing HMI components for a driver assistance system for safe speed and safe distance. *Advances in Transportation Studies*. vol. 2, issue 21. pp. 5-14.
7. Young, K. L., Regan, M. A., Triggs, T.J., Tomasevic, N., Stephan, K., Mitsopoulos, E. (2007). Impact on car driving performance of a following distance warning system: Findings from the Australian transport accident commission SafeCar project. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*. vol. 11 issue 3. pp. 121-131.
8. Lahrmann, H., Agerholm, N., Tradisauskas, N., Berthelsen, K.K., Harms, L. (2012). Pay as You Speed, ISA with incentives for not speeding: Results and interpretation of speed data. *Accident Analysis & Prevention*. vol. 48. pp. 17-28.
9. FOT-Net. *Data Catalogue* [Internet]. [cited 2019 Jan 2]. Available from: http://wiki.fot-net.eu/index.php/Data_Catalogue
10. Gellerman, H., Kotiranta, R., Koskinen, S., Val, C., Bakker, J., Agerholm, N. (2017). *FOT-Net Data - Data protection recommendations*. FOT-Net Data.
11. Lu, R., Zhu, H., Liu, X., Liu, J., Shao, J. (2014). Toward efficient and privacy-preserving computing in big data era. *IEEE Network*, vol. 28. issue 4. pp. 46-50.
12. The European Parliament; The Council of the European Union. (2016). *Regulation (EU) 2016/679 of The European Parliament and the Council of 27 April 2016*. 2016/679. pp. 1–88.
13. The European Parliament; The Council of the European Union. (2013). *Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information*. The European Union. pp. 1–8.
14. Lahrmann, H., Agerholm, N., Juhl, J., Bech, C., Tøfting, S. (2013). The development of an open platform to test ITS solutions. In Proceedings *9th ITS European Congress*, Dublin: ERTICO (ITS Europe). pp. 1-5.
15. Lahrmann, H., Agerholm, N., Juhl, J., Araghi, B.N., Højgaard-Hansen, K., Bloch, A.-G. et al. (2012). ITS Platform North Denmark: Idea, content, and status. In Proceedings *19th ITS World Congress*, Vienna. pp. 1-12.
16. Agerholm, N., Lahrmann, H., Jørgensen, B., Simonsen, A.K. (2014). Full-automatic parking registration and payment - in principle GNSS-based road pricing. In Proceedings *10th ITS European Congress*, Helsinki. ERTICO (ITS Europe). pp. 1-12.
17. Jensen, C.S., Lahrmann, H., Pakalnis, S., Runge, J. (2004). *The Infati Data*. Timecenter.
18. Open Street Map. *OpenStreetMap latest data* [Internet]. [cited 2019 Jan 9]. Available from: https://www.openstreetmap.org.
19. Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996). CiteSeerX - A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings *2nd International Conference on Knowledge Discovery and Data Mining*. Portland, Oregon. pp 1-6.
20. Valhalla. (2019). *Valhalla - Open Source Routing Engine for OpenStreetMap* [Internet]. [cited 2019 Jan 11]. Available from: https://github.com/valhalla
21. Gidófalvi, G., Huang, X., Pedersen, T. B. (2007) Privacy-Preserving Data Mining on Moving Object Trajectories. In Proceedings *2007 International Conference on Mobile Data Management*. IEEE.

22. Gøeg P, Kveladze I, Lahrmann HS, Agerholm N. (2019). *Anonymised Floating Car Data – the long path to data sharing.* Afhandling præsenteret på 13th ITS Europe Conference 2019, Eindhoven, Holland.
23. Graser A. *Free and open source GIS ramblings - Movement data in GIS series* [Internet]. [cited 2019 Jan 2]. Available from: https://anitagraser.com/