Anonymised Floating Car Data – the long path to data sharing

Gøeg, Pelle; Kveladze, Irma; Lahrmann, Harry Spaabæk; Agerholm, Niels; Koskinen, Sami

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

Paper ID SP1722

Anonymised floating car data – the long path to data sharing

Pelle Rosenbeck Gøeg[1], Irma Kveladze[1], Harry Lahrmann[1], Niels Agerholm[1*], Sami Koskinen[2]

1. Division of Transportation Engineering, Aalborg University, Denmark, na@civil.aau.dk

2. VTT Technical Research Centre of Finland, Espoo, Finland

Abstract
It is expected that there exists a significantly unused value while reusing floating car data (FCD) for other purposes than initially anticipated. An open reuse of such data requires that drivers whose driving pattern can be reflected from such data remain anonymised regardless of any later use of the FCD. The FCD collected in the Danish big data project, ITS Platform, have been anonymised. Anonymisation has partly been implemented by removing the time of driving, although the main effort ensured that the starts and stops of any trips do not eliminate anonymisation. FCD from 389 cars have been anonymised. In total, 0.74 billion positions and a distance driven of about 9.7 million km were anonymised. The anonymised FCD reflect the original driving pattern, although the number of trips has been reduced by 26%, while the distance pr. trip has increased by 27%.

Keywords:
floating car data, anonymisation, transport analyses

Introduction

A range of research as well as innovative projects and products in the field of floating car data (FCD) have been conducted during the last two decades (1,2). While significant FCD volumes have been collected and are currently being collected, the value of these data, aside from its initial purpose, is rarely discovered and/or realised (2). Most datasets have only been slightly analysed, and many other perspectives from these data may reasonably be considered productive when applied (3,4).

Many-sided purposes of collected FCD have been fulfilled in several different projects, yet in many cases, the data may be used for other purposes. As an example, the results of speeding effects have been published from many Intelligent Speed Assistance projects (5,6), while few have focused on the effects on transportation time (7,8). Likewise, careful analyses have been conducted in naturalistic driving studies on a few predefined hypotheses, while the abundance of other knowledge has hardly been reached (9). Also, with few exceptions, the conducted data analyses have been unable to present data from individual trips to harvest the value from the knowledge of driving variation occurring among various road users.

In many cases, this unused source of knowledge is associated with a lack of resources and/or significant underestimation of the costs associated with data analyses; however, in other cases, the analyses are limited due to privacy issues (4,10). By implementing the General Data Protection Regulation (GDPR) in May 2018, it was stressed that privacy must be sufficiently provided and personal data must be sufficiently protected, maintained in appropriate volumes, capable of fulfilling their purpose, and only be shared when sufficient data protection steps are taken. However, if data are anonymised and neither the data responsible nor any data practitioner can connect them to the personal information via, e.g., reverse engineering, the data are not considered personal and are hence considered non-sensitive (10,11,12).

Another perspective is that all data collecting financed by public funding must be fundamentally accessible to the public without any further costs apart from those related to the data-sharing procedure (13). In practice, data sharing rarely happens because few organisations are prepared to share their data for various reasons, of which privacy issues - be they right or wrong - have acted as an obstacle for any data sharing.

In order to meet the privacy challenge as well as make data collected by public funding available for data sharing, the target involves being able to share data from the Danish big data project, ITS Platform (14,15,16), such that they will be accessible to a wider audience.

Through the work performed to produce this paper, it was made clear that, although some researchers (2,3,10,17) have demonstrated good principles for anonymisation, the sufficient anonymisation of big FCD under GDPR's full effect is far from a trivial task when conducted successfully. A first attempt to anonymise FCD collected for the ITS Platform from 2016 to 2017 failed due to the complexity of the task (4). However, in order to gain further usability from these data, the FCD were anonymised afterwards. Based on the effort related hereto, the following explorative research question was identified:

> *Can we execute a sound anonymisation of FCD that allows us to both retain central information concerning the driving and make it publicly available?*

Method

A coherent view of the vehicle's driving pattern might affect the associated driver's privacy. Therefore, it is necessary that the FCD be anonymised. Hence, the following procedure for performing this obfuscation has been followed in order to make any privacy information unavailable. First, the endpoints of all trips for each individual driver have been identified in order to define each individual trip in the FCD. Second, endpoint clusters are identified, a buffer with a random centroid near each endpoint is implemented for each cluster, and FCD related to any endpoints are removed. Finally, all information connecting trips together are removed, and trips are distributed into time periods to determine the type of each trip - e.g., peak period trip. The details of the anonymisation procedures include the following:

1. FCD is aggregated to trajectories as linestring ZM. Data are connected according to temporal value and are maintained with the same IDs depending on their connected on-board units (OBUs). Minor outages are ignored (typically 1–2 seconds), while outages >120 seconds are treated as a completed trip (see Figure 1).
2. FCD are map-matched to the used street dataset (OpenStreetMap (18)).
3. Then, the endpoints (start and stop) are identified for each individual trip.

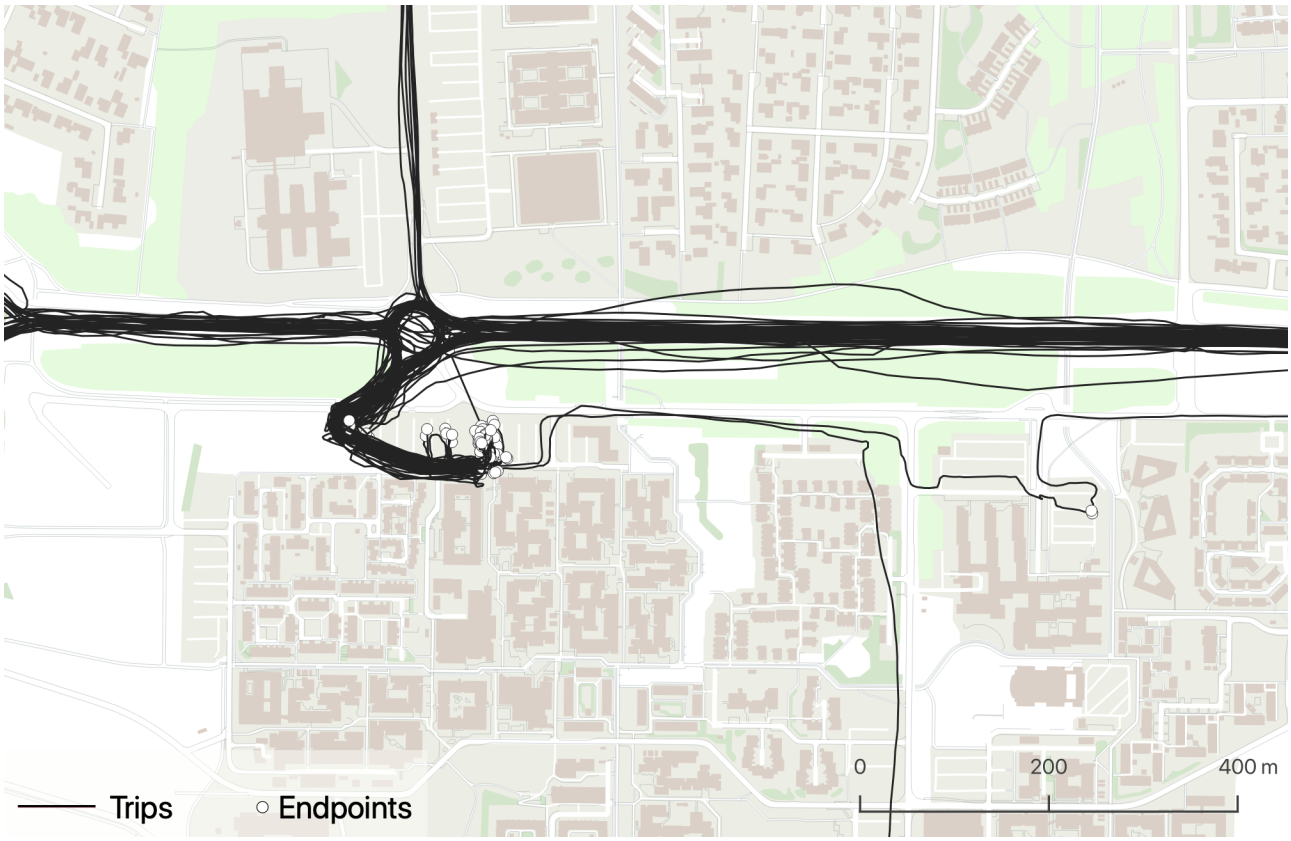

Figure 1 - FCD trajectories' starts and stops in a parking area.

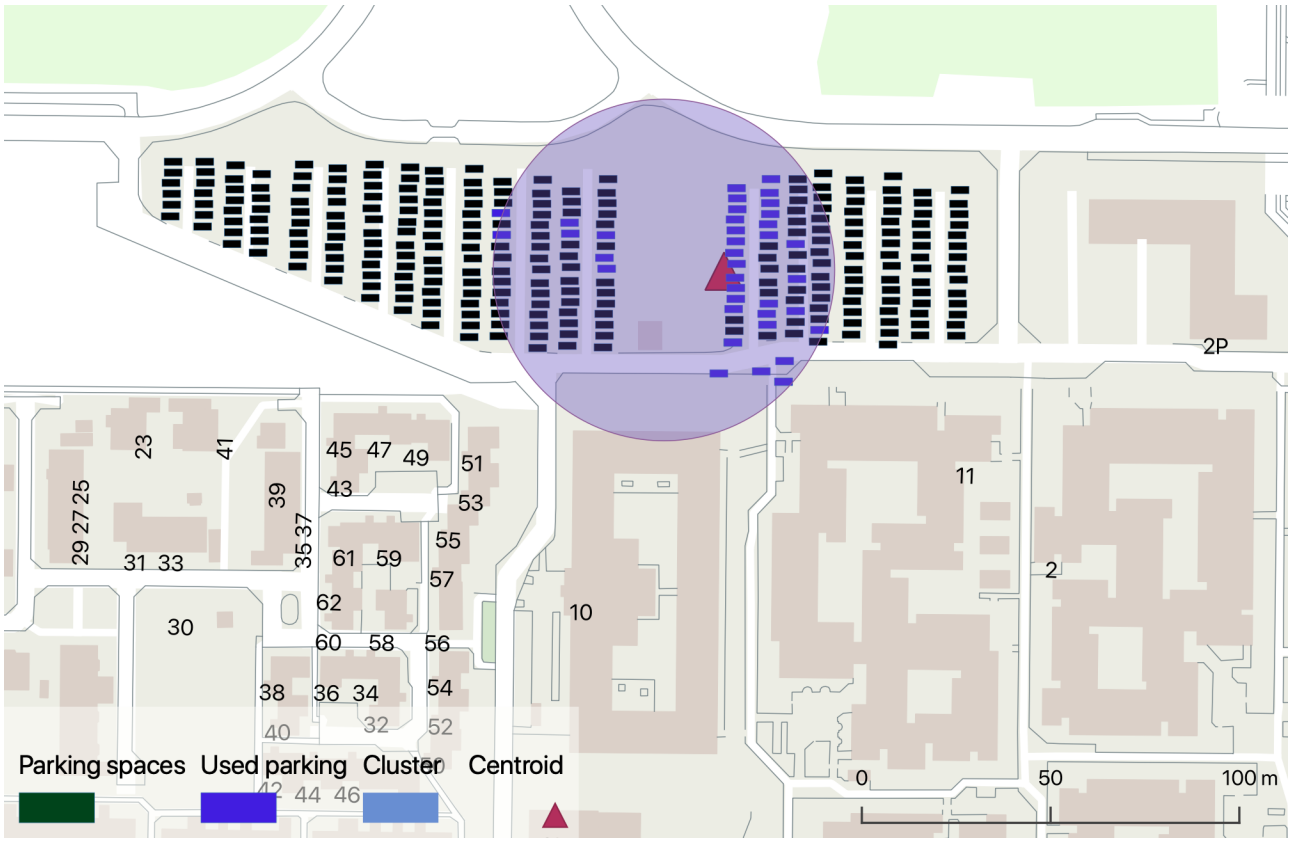

Figure 2 - Parking situation illustrating multiple parking slots used for the same point of interest (POI); blue colour represents used slots. The blue circle is the minimum bounding box for the starting point and endpoint.

4. All parking situations (e.g., starts and stops) are derived from each trip's endpoints, which have been clustered using the DBscan algorithm (19) and wherein the distance has been set to 50 m. All FCD are scanned, and parking locations with less than 50 m between are seen as the same cluster - meaning, in the event that one driver has various parking spaces selected in an area, the cluster may be rather large. The distance of 50 m was chosen because it has been assessed as a reasonable distance drivers, in most cases, consider suitable for walking to/from a parking lot. These clusters are termed 'points of interest' (POI) (see Figure 2).
5. Subsequently, an algorithm calculates the necessary radius of a circle, including a buffer (buffer 1) containing a minimum of 50 addresses or a radius of 2,000 m. The density of addresses near one location defines the size of the circle's buffer, although in low-density areas, the default radius size is 2,000 m. Subsequently, the buffer circle's centroid is defined as the parking cluster's centroid (see Figure 3).
6. One address inside buffer 1 is randomly selected as origo for the new buffer (buffer 2), which is calculated with a distance of at least that to buffer 1's periphery. All FCD from the OBU in question are removed from the buffer area unless they indicate the buffer's passage without a stop >120 seconds (see Figures 4 and 5).

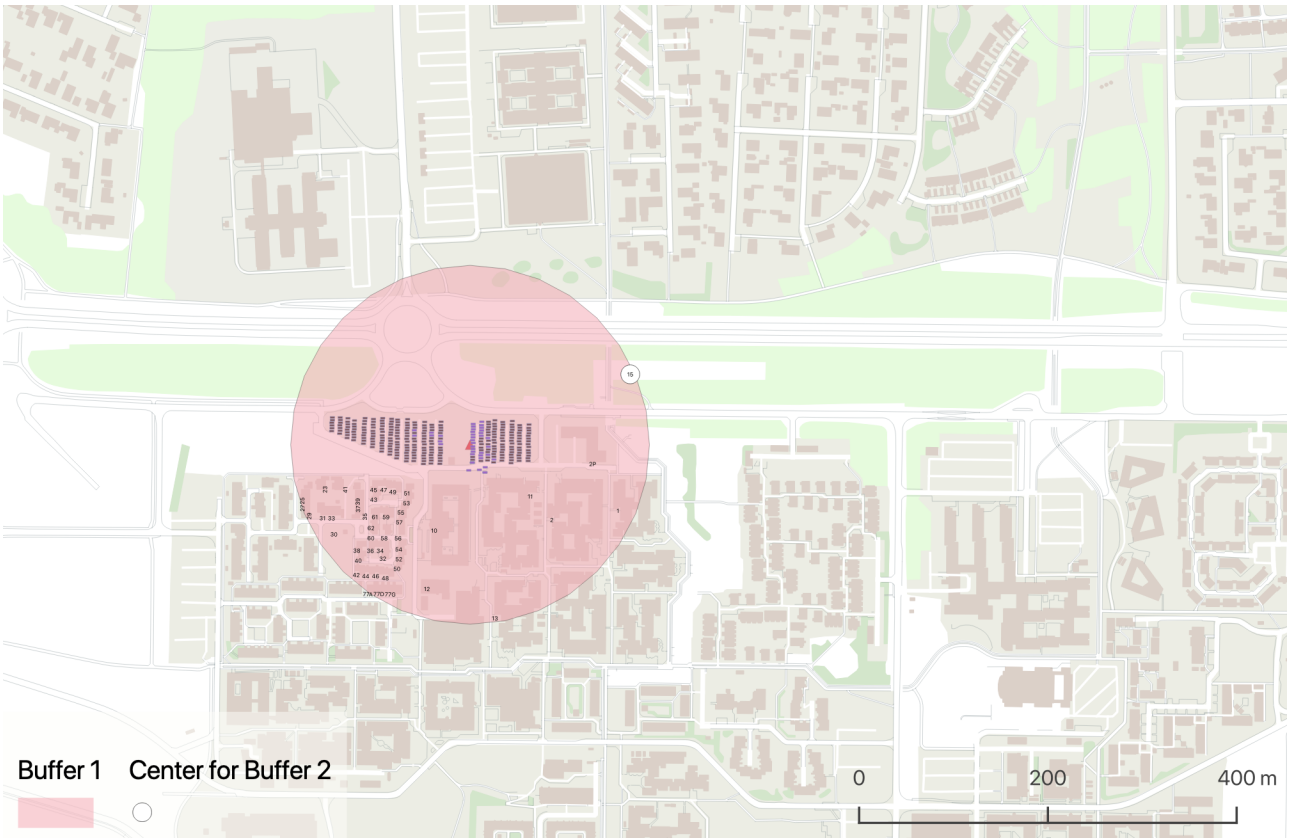

Figure 3 - Buffer, origo depicted as a red triangle, increased to r=180 m before 52 addresses are reached; thereafter, a new, random origo is depicted as a white circle.

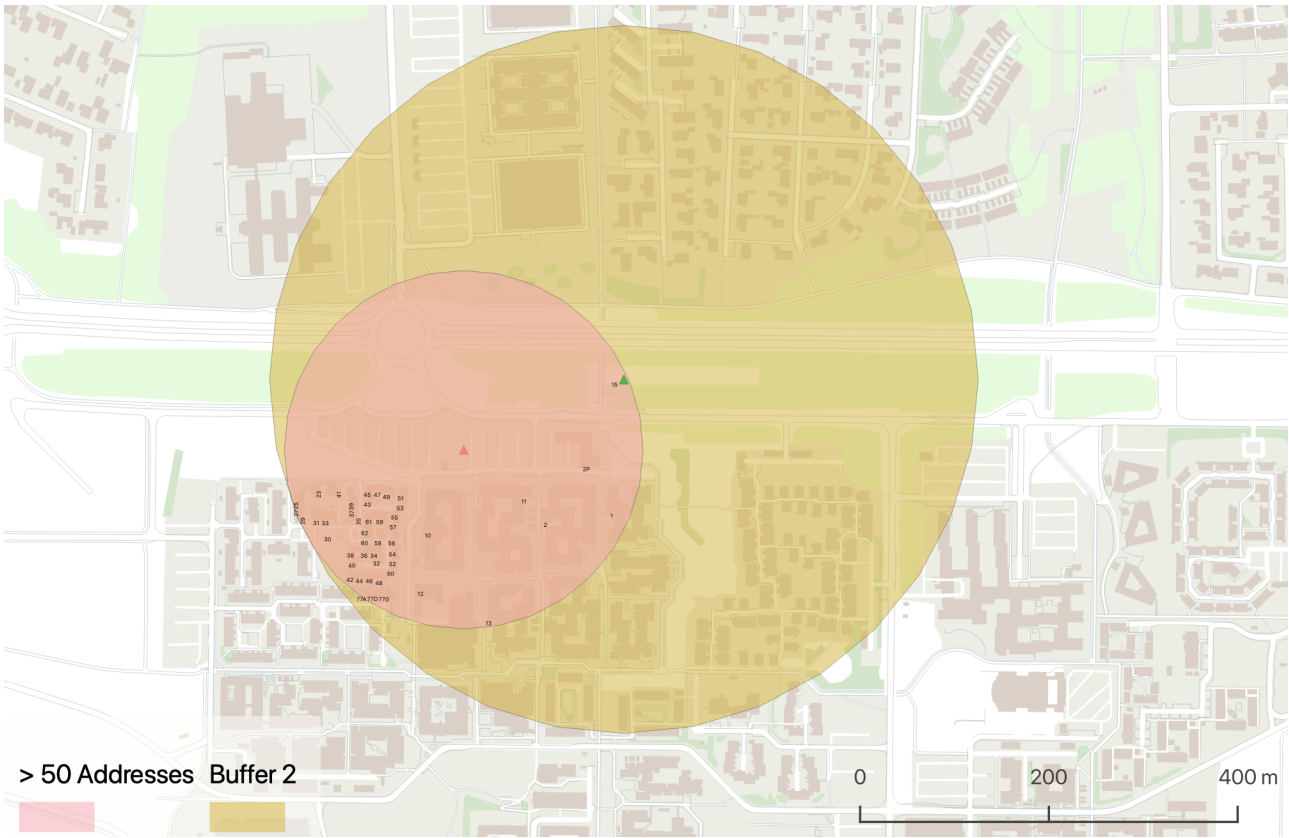

Figure 4 – Buffer created randomly, with a minimum of the first 50 addresses calculated.

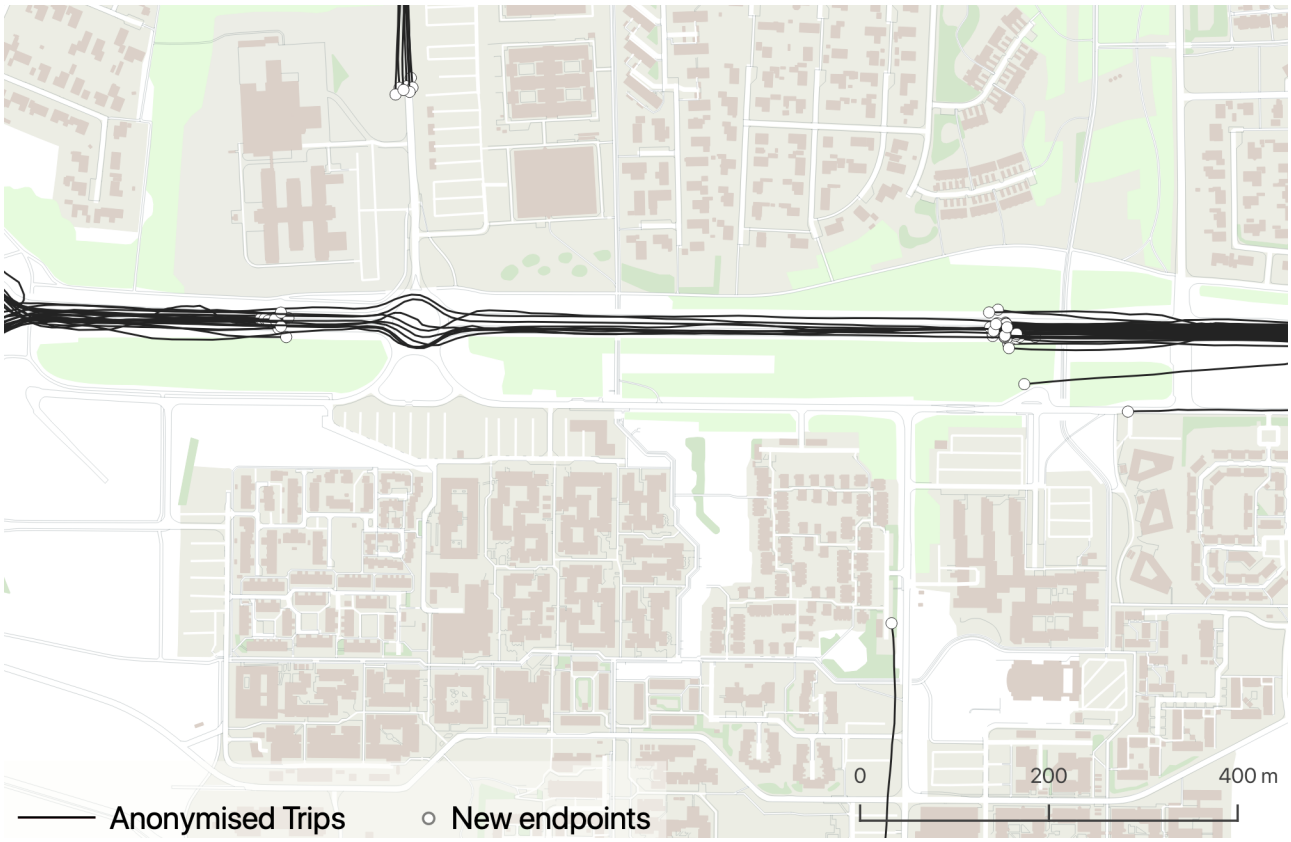

Figure 5 – Anonymised trajectories; the POI is not to be found, as basic geometry with a minimum of three points on the circle's periphery would not necessarily lead to the centroid.

7. On the basis of their initiation time, the trips are distributed into the following time periods: peak period (07-09 and 14-17), shoulder period (09-14), evening period (17-22), and free flow period (22-07). This is done to obfuscate the exact time of each trip, yet at the same time to maintain information about peak period speed, free flow speed, etc. This means that, if a trip is initiated at 8.30 AM, the first registration will be depicted as 7.00. Likewise, date, year, and weekday are substituted exclusively with workday/weekend.
8. Finally, the ID related to the unique OBU is removed - i.e., the individual trip can be followed, although it is not possible that different trips be related to one another.
9. The anonymised FCD are assessed by measuring the lengths of trajectories and utilising map matching to quantify the individual points as well as uncover the differences within the data.

The anonymisation procedures are conducted in Python 3.6 alongside the Meili library in Valhalla (20), a map-matching engine on OpenStreetMap, as well as data covering Denmark (18) and PostgreSQL/PostGIS 2.5 for storage and spatiotemporal analyses of the FCD.

The above-mentioned procedure differs from other described procedures for anonymization as the data volume is different and the needs for watertight anonymisation due to the GDPR requirements have increased markedly. At the same time, it is desirable to keep most possible information in the FCD in order to keep the value of the anonymised version of this information.

Jensen et al. (17), which we owe a lot, suggest only removing FCD in a fixed 2 x 2 km square with most activities outside working time, it partly removes too much FCD in dense areas and too little in very remote areas. Also, they worked with a smaller data set (approximately 530,000 positions in total) collected in a short-term period (six weeks). This means that the pattern in driving behaviour, which might endanger anonymization due to months or years of comparable trips, would not be a challenge in their study.

Furthermore, in some previous efforts for anonymization of FCD, the aim was among others to protect anonymization during the entire data flow from collection to completed data treatment. This online process will in many cases be carried out with less data processing power and hence more simplified and "one-fits-all" approaches as e.g. described in (21). An approach like this, being off- or online, would to the best knowledge of the authors devaluate the value of the anonymised ITS Platform FCD too much.

As a conclusion on the choice of methods, it has been paramount to reach sufficient anonymization with long-term collected FCD and, at the same time, make the procedures sufficiently agile to minimise the removal of valuable data where it is not needed. More manual approaches are not viable due to data volume.

Data

The utilised FCD were collected in the ITS Platform project from 2012 to 2014 (4,14,15). The ITS Platform consisted of an OBU, a backend server, and several applications. 425 mainly privately owned cars (of which 389 provided data) were included in the project. The OBU was a mobile unit that was tailor-made specifically for mobile ITS services and was operated via GPRS. In order to run selected applications and obtain knowledge of driving patterns for research purposes, the OBUs collected FCD with 1 Hz among additional attributes, including ID, position, timestamp, direction, and speed (14). FCD from some of these OBUs are removed, while others are transferred as part of the anonymisation procedure. The most central attributes in the anonymised data appear in Table 1. FCD consisted of about $0{,}57*10^9$ positions, equalling a distance driven of about 9.7 million km.

Table 1 - The attributes in the anonymised FCD.

Attribute	
Trip - ID	a unique number that is connected to each trip. A trip is defined if there are > 2 minutes from the prior data registration; there are no connections between trip number, time, or car.
Timestamp	Date is converted to month, but start time is modified to fit into the four time periods.
Longitude - X	a 6-digit number, which describes the eastward–westward distance **in metres** to Greenwich according to WGS89 UTM32N.
Latitude - Y	a 7-digit number, which describes the distance **in metres** to the equator according to WGS89 UTM32N.
Speed	measured in **metres/seconds** based on the Global Global Navigation Satellite System (GNSS) registrations.
Direction	recorded driving direction in **degrees** (0–359) based on the GNSS registrations.

Results

By comparing original and anonymised FCD, it is demonstrated that 28% of FCD are removed, thus implementing an average increased trip length of 27%, as explained in Table 2.

Table 2 - Characteristics for the original and anonymised FCD for one random vehicle and in total, respectively.

	Random vehicle			Total dataset		
	Before	After	Change	Before	After	Change
Vehicles	1	1		389	389	0%
Trip counts	3,114	2,422	-22%	741,559	541,910	-26%
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%	12.8	16.2	27%
Max. trip length (km)	421	416	-1%	832	821	-1%

Data access

The anonymised data is available through the RESTful Web Service (webservice - see https://fcd-share.civil.aau.dk/ for detailed information). In order to obtain data access, it is mandatory that the user sign up with details regarding name, e-mail, organisation, and purpose for data extraction. After submitting these details, an access token will be provided such that the user may gain access to the webservice. Via this webservice, the user may query the dataset for specific parts. The default data format is GeoJSON. A query can be built only to return data that obey - among other specific attributes - time range, working day or weekend trips, area of interest, and OpenStreetMap road identity (for an example, visit http://fcd-share.civil.aau.dk/points?when=weekend&osm_id=8149020).

Discussion and Summary

Discussion

Anonymising FCD is a balanced task: too much removed information results in too little value obtained from the data. On the other hand, too little information removal may result in an increased risk of recognising confidential data, such as trips' starting points and endpoints in combination with residence information. If we had decided to exclusively remove data in a certain radius from a starting point/POI, the centroid, and hence the real starting point, would be obvious for a number of trips shown in the data. The situation would have been the same had we decided to place a unique centroid at the same distance from the actual POI for each trip, although additional trips would have been required before the POI had been made obvious. Likewise, an inclusion of the full-time information or even the actual date would have only made observing a clear driving pattern from the anonymised data possible. With sufficient trips, it would have become likely that a specific trip with any inappropriate privacy information would be exposed.

Furthermore, pseudonymisation rather than anonymisation would have allowed a data analyst to draw both an explicit connection between all POIs in question and, if combined with other sources of information, a clear clue for identifying any driver and hence weakening the anonymity.

On the other hand, if an unnecessary, careful anonymisation were conducted and if an increased area's FCD were removed to a radius of 2 km, many small trips would be removed and the data's value would decrease equivalently. In addition, full removal of time information would result in low data value because little may be extracted from the effects of peak traffic, weekend traffic etc.

The selected anonymisation level and procedure are expected to ensure sufficient anonymisation as well as simultaneously provide information from the FCD such that more of their expected values may be utilised.

Summary
FCD sets have been collected on a large scale during the last two decades. While some of the initiating research/study results have been utilised, most knowledge embedded in FCD has not been utilised sufficiently. It is the intention of the EU statutes that data collected on the basis of public funding be made available to the public at a price covering only the actual data sharing costs. This work aims to utilise the FCD collected from the ITS Platform project anonymously. FCD from 389 cars collected over three years (2012-2014) have been anonymised via: 1: removal of FCD inside a scalable radius with a displaced centroid from the POIs in the dataset; 2: radius scaled to the proportion of the address density such that a densely built-up area requires a much smaller radius than a rural area with few addresses; 3: replacement of time and date with workday/weekend and four time periods scattered during the day; and 4: anonymisation of the individual information of the registered car such that any connection between trips are removed.

The anonymisation procedure reduced the number of trips by 26%, and the trips' average lengths in time and distance were altered from 12:28 to 11.54 minutes 12.8 and to 16.2 km, respectively. Due to the distribution in time periods, the anonymised FCD allow any data analyst to measure driving behaviour depending on the time of the trip. While the characteristics of the longest trips reasonably hardly changed, the shorter trips were less precise. The anonymised FCD are available for further research at a specially designed homepage.

Acknowledgements
This work was possible due to the support provided from beyond the university. Therefore, we would like to thank the Danish Innovation Fund via the DiCyPS Center as well as the European Union's Research and Innovation funding programme, FP7, for their prior support via the FOT-Net Data project. In addition, we wish to thank the original project that collected the raw data, the ITS Platform, as well as the European Regional Development Fund and the North Denmark Region for their support. Finally, we must highlight that Anita Graser's blog (22), 'Movement data in GIS', has been a substantial inspiration to this project.

References

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. In 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. Graser A. *Free and open source GIS ramblings - Movement data in GIS series* [Internet]. [cited 2019 Jan 2]. Available from: https://anitagraser.com/