

Modelling Spatio-Temporal Human Behaviour with Mobile Phone Data: A Data Analytical Approach.

DIETER OOSTERLINCK

Supervisors: Prof. Dr. Dries F. Benoit
Prof. Dr. Philippe Baecke

Typeset in L^AT_EX.

Copyright © 2021 by Dieter Oosterlinck (dieter.oosterlinck@ugent.be)

All rights are reserved. No part of this publication may be reproduced or transmitted in any form or by any means electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the author.

EXAMINATION BOARD

Prof. dr. Patrick Van Kenhove (Dean, Ghent University)

Prof. dr. Dirk Van den Poel (Secretary, Ghent University)

Prof. dr. Dries F. Benoit (Supervisor, Ghent University)

Prof. dr. Philippe Baecke (Co-supervisor, Vlerick Business School)

Prof. dr. Nico Van de Weghe (Ghent University)

Prof. dr. Wouter Verbeke (KU Leuven)

Prof. dr. Vera Miguéis (Universidade do Porto)

Voorwoord

Eerst en vooral wil ik mijn promotoren, Dries en Philippe bedanken. Voor de kans die ze mij boden, voor de mooie jaren, de goeie samenwerking, de ruimte die ze mij gaven, de constructieve feedback en hun positieve ingesteldheid. Zonder hen was dit doctoraat nooit tot stand gekomen. Bij deze een tip voor iedereen in ons vakgebied, maar ook aan iedereen daarbuiten: praat eens met deze twee unieke mensen, ga er eens een pint mee drinken, je zal niet ontgoocheld zijn.

Secondly, I would also like to sincerely thank the jury of my doctoral defense. Thank you for the valuable comments and for challenging me with critical questions. Thank you for agreeing on awarding me with the title of PhD, it means a lot to me to receive this honour from such an internationally esteemed jury.

Dankjewel aan de vele collega's die ik de voorbije jaren heb mogen leren kennen op de vakgroep. Thank you very much to all colleagues that I've had the pleasure of getting to know over the past years. I will continue in English, to make sure that each and every one in this international and diverse group fully understands the following words. It's been a true pleasure to work in such a nice environment and although the main divide in the department has been the one between 'the modellers' and the 'CB-ers', this division was only based on research and not on any personal level what so ever. I've really missed the great atmosphere during the last year of my PhD as we were all exiled from the faculty due to the corona pandemic.

Nog een extra dankjewel aan mijn bureaugenoten. Met jullie heb ik de voorbije jaren de meeste tijd gependend en jullie invloed op dit PhD (en mezelf) mag zeker niet onderschat worden. Dankjewel Steven voor jouw rol als officiële peter, dankjewel Gudrun voor jouw rol als onofficiële, zelfbenoemde, meter.

Dankjewel ook aan alle collega's buiten de UGent, die op conferenties waardevolle input gaven en ook daar voor een heel aangename sfeer zorgden. Ondanks het competitieve aspect in de academische wereld, heb ik deze wereld vooral als een wereld van 'peers' mogen ervaren.

Naast alle mensen die ik rechtstreeks dankzij de academische wereld heb mogen leren kennen, wil ik uiteraard ook nog heel wat andere mensen bedanken, die evenmin een niet te onderschatten invloed op dit PhD hebben gehad. Dankjewel aan mijn vrienden en familie om mijn agenda tijdens de twee vrije dagen in de week steeds op een leuke manier in te vullen. Dankjewel aan mijn (schoon)ouders, (schoon)broer(s), schoonzussen en (schoon)grootouders. Een aanzienlijk stuk van mijn motivatie voor dit doctoraat komt voort vanuit jullie allemaal die steeds in mij geloven en trots zijn op wat ik doe. Merci!

Kim, sinds we afgestudeerd zijn, heb je niet alleen altijd aan mijn zijde gestaan bij dit doctoraat, maar hebben we al heel wat andere mooie projectjes verwezenlijkt. De coronapandemie is geen pretje, de sfeer op de vakgroep missen is dat ook niet, maar ik ben jou ongelooflijk dankbaar dat ik het privilege heb gehad om dit doctoraat te kunnen finaliseren in ons huis, 'locked down' met jou en Elias. Onze zuidgerichte ramen zijn magnifiek, maar de echte zonnetjes in huis, dat zijn jullie!

Dieter Oosterlinck

Table of Contents

List of Figures	xi
List of Tables	xiii
Nederlandstalige Samenvatting	xv
Summary	xvii
1 General Introduction	1
1.1 Data Sources for Tracking Spatio-Temporal Human Behaviour . .	2
1.2 From Tracking Data to Modelling Behaviour	6
1.3 Tracking and Modelling Applications	8
2 Bluetooth Tracking of Humans in an Indoor Environment: An Appli- cation to Shopping Mall Visits	13
2.1 Introduction	14
2.2 Literature Review	15
2.2.1 Customer Tracking	15
2.2.2 Methods for Tracking	16
2.2.2.1 GPS	17
2.2.2.2 RFID	17
2.2.2.3 Bluetooth	18
2.2.2.4 Wi-Fi	19
2.2.3 Indoor Positioning	20
2.3 Experimental Design	22
2.3.1 Equipment	22
2.3.2 Description of Study Area	23
2.3.3 Test Cases	24
2.4 Results and Analysis	30
2.4.1 Data	30
2.4.2 Applications	31
2.5 Discussion and Conclusion	37
2.6 Limitations and Future Research	38
2.7 Acknowledgments	39
3 Home Location Prediction with Telecom Data: Benchmarking Heuris-	

tics with a Predictive Modelling Approach	41
3.1 Introduction	42
3.2 Literature Review	44
3.2.1 CDR Data for Human Mobility and the Need for Home Detection	44
3.2.2 Categories of Home Detection Algorithms	45
3.2.2.1 Decision Rules and Heuristics Based on Activ- ity (Activity Heuristics)	46
3.2.2.2 Decision Rules and Heuristics Based on Inactiv- ity (Inactivity Heuristic)	48
3.2.2.3 Two-Step Clustering Approaches	48
3.2.3 Validation	50
3.2.4 Social Network	52
3.3 Methodology	54
3.3.1 Data	54
3.3.2 Validation Metrics	55
3.3.3 Benchmarks	56
3.3.4 Predictive Modelling Approach	56
3.3.4.1 Binary Dependent Variable: Home Tower . . .	56
3.3.4.2 Independent Variables	57
3.3.4.3 Binary Classification Algorithms	58
3.4 Results	60
3.4.1 Benchmarks	60
3.4.2 Optimisation of Benchmarks	62
3.4.3 Predictive Modelling Approach	65
3.4.4 Exploratory Performance Analysis	67
3.4.5 Combined Inactivity Activity Heuristic Method	70
3.4.6 Summary of Results	71
3.5 Conclusion and Future Research	74
 4 From One-Class to Two-Class Classification by Incorporating Expert Knowledge: Novelty Detection in Human Behaviour	 79
4.1 Introduction	80
4.2 Literature Review	81
4.2.1 Novelty Detection	81
4.2.2 One-Class Novelty Detection	83
4.2.3 Two-Class Novelty Detection	85
4.3 Method Development	88
4.3.1 Expert Knowledge	88
4.3.2 Expert Scenarios	89
4.4 Case Study: Telecom Subscription Fraud	91
4.4.1 Business Problem	91
4.4.2 CDR Data	92
4.4.3 Incorporating Expert Knowledge	93
4.4.4 Benchmark Model 1: One-Class Classification	95

4.4.4.1	Benchmark Model 1a: One-Class Probabilistic	96
4.4.4.2	Benchmark Model 1b: One-Class k-Nearest-Neighbours	97
4.4.4.3	Benchmark Model 1c: One-Class SVM	97
4.4.5	Benchmark Model 2: Two-Class Artificial Data Generation Models	99
4.4.5.1	Benchmark Model 2a & 2b: Probabilistic Artificial Data Generation	99
4.4.5.2	Benchmark Model 2c: Distance-based Artificial Data Generation	100
4.4.6	Two-Class Expert Model	101
4.4.7	Real-Life Post-Launch Implementation and Validation	105
4.4.8	Manual Checks on Post-Launch Predictions	106
4.5	Discussion	109
4.6	Conclusion	110
4.7	Appendix	111
4.7.1	Benchmark Model 1b: Optimal Value for k	111
4.7.2	5-Fold Cross-Validation: Performance Measures of the Different Folds	112
4.7.3	Overview of Variables	112
5	Conclusion	117
5.1	Discussion	118
5.1.1	The Single Best Method	118
5.1.2	Privacy and GDPR	120
5.2	Conclusion and Implications	121
5.2.1	Benefits of Bluetooth Tracking and Call Detail Records	123
5.2.2	Methodological Contributions	124
5.2.3	Numerical Results of Applications and Implications	125
5.3	Limitations and Future Research	126
	Bibliography	131

List of Figures

2.1	Bluetooth Scanner	23
2.2	Test Case 1: Detection of Bluetooth Signal.	25
2.3	Test Case 2: Positioning of Scanners in Three Neighbouring Stores.	26
2.4	Test Case 3a: Interpolation of RSSI Values.	27
2.5	Test Case 3b: Interpolation of RSSI Values, Wrapped Scanners.	28
2.6	Test Case 3c: The Relation between RSSI and Distance.	29
2.7	Number of Stores Visited (Relative Frequencies)	32
2.8	Time Measurement of Visits	32
2.9	Major Flows on First Floor of Shopping Mall	34
2.10	Migration Plot of Major Flows	35
2.11	Cluster Dendrogram of Hierarchical Clustering.	36
3.1	Overview Home Detection Methods.	47
3.2	Example of Adapted Hartigan Leader Algorithm.	51
3.3	Inactivity Method	63
3.4	Distribution of Total Inact_7 Counts	70
3.5	Combined Inactivity Activity Heuristic	72
3.6	Reversed Cumulative Distance Error for Best Performing Models in Different Categories	73
3.7	Zoomed Reversed Cumulative Distance Error	74
4.1	Novelty Detection	82
4.2	Positioning of Expert Data Generation Method in the Novelty De- tection Literature.	89
4.3	Expert Data Generation Method	90
4.4	Network Dyad Selection; Positive and Negative Class.	94
4.5	Robustness Check	103
4.6	Comparison of Predictions on <i>Post</i> -Launch Dataset.	105
4.7	Results of Manual Labelling of Predicted Fraud Suspects by Fraud Team of the Company	106
4.8	Identified Fraud Case: CDR Network Visualisation	107
4.9	Identified Fraud Case: Location Plot	108

List of Tables

1.1	Main Characteristics of Different Spatio-Temporal Data Collection Methods.	4
1.2	Structure of Dissertation	9
2.1	Bluetooth Classes	22
3.1	Structure of the Base Table for the Predictive Modelling Approach.	57
3.2	Variables in Predictive Model	59
3.3	Benchmark Results	60
3.4	Percentage of Individuals without Prediction.	62
3.5	Inactivity Method	63
3.6	Scoring the Hartigan Leader Based Methods with Known Activity/Inactivity Heuristics	64
3.7	Predictive Method Results	65
3.8	Correlations between Explanatory Factors and Performance Measures of Best Performing Methods.	69
3.9	Results Home Detection Methods	75
4.1	Expert Fraud Scenarios	95
4.2	Pre-Launch Data	96
4.3	Confusion Matrix One-Class Benchmark Models	98
4.4	Average Performance (5-fold cv) $P(+ X)$ on Artificially Generated Data.	99
4.5	Confusion Matrix Two-Class Probabilistic Artificial Benchmark Models	100
4.6	Two-Class Artificial Model (Benchmark 2c): Confusion Matrix on the Artificial Test Set	101
4.7	Two-Class Artificial Model (Benchmark 2c): Confusion Matrix on the Expert Test Set	101
4.8	Two-class Expert SVM Confusion Matrix per Scenario on the Test Set	102
4.9	Robustness Check	103
4.10	Fraud Detection Rate (FDR) of Three Novelty Detection Methods Evaluated on Two Datasets.	104
4.11	Predictions of the Different Classes of Models for a Fraud Example	108

4.12	Fraud Detection Rate for Different k -Values	111
4.13	5-Fold Cross-Validation: Performance Measures of the Different Folds.	112
4.14	Overview of Variables	115
5.1	Overview of Main Contributions and Findings of this Dissertation	122

Nederlandstalige Samenvatting

De analyse van tijdruimtelijk menselijk gedrag is waardevol voor verschillende applicaties in zowel een academische als een bedrijfscontext. De traditionele aanpak die gebaseerd is op enquêtes heeft meer en meer plaats gemaakt voor *tracking* methodes, methodes die toelaten om het menselijk tijdruimtelijk gedrag effectief te observeren en te registreren. Mobiele telefoons (GSM en smartphone) zijn hiervoor een zeer geschikt middel. Het hoofddoel van dit proefschrift is onderzoeken hoe data van deze mobiele telefoons gebruikt kan worden om het tijdruimtelijk menselijk gedrag te modelleren. Daartoe werden twee databronnen geselecteerd, het Bluetooth signaal en het gewone telefoonsignaal. De eerste bron resulteert in *Bluetooth tracking* data, de tweede bron is gekend als *call detail record (CDR)* data. Beide databronnen zijn geschikt voor tijdruimtelijke analyses, door middel van het nabijheidsprincipe. In het geval van Bluetooth tracking data geeft de registratie van de mobiele telefoon door de Bluetooth scanner aan dat het individu in de nabijheid van de scanner is geweest. Voor CDR data geeft de registratie door een telefoonmast aan dat het individu in de nabijheid van die mast geweest is. Aangezien de locaties van de scanners en de masten gekend zijn, kunnen deze gebruikt worden als goeie benaderingen voor de effectieve locatie van het individu. De ruwe CDR en Bluetooth data is op zich onvoldoende om tot betekenisvolle inzichten te komen. Daarom werd een data-analytische aanpak toegepast om waarde te creëren op basis van deze databronnen.

Het eerste centrale hoofdstuk in dit proefschrift onderzoekt de waarde van Bluetooth tracking om het tijdruimtelijk gedrag van bezoekers in een winkelcentrum te modelleren. Deze applicatie vereist een trackingmethode die geschikt is voor indoor gebruik en die tegelijk de nodige precisie op winkelniveau biedt. Een experiment met 56 Bluetooth scanners toonde aan dat Bluetooth tracking geschikt is voor dit doel. 9.81% van de bezoekers werden effectief getraceerd. Deze eerder hoge detectieratio zorgt ervoor dat Bluetooth tracking snel een grote set aan respondenten genereert. Hoewel het experiment aantoonde dat Bluetooth tracking een geschikte methode is, blijven een aantal manuele aanpassingen noodzakelijk om vals positieve registraties te vermijden; bijvoorbeeld bezoekers van een aanpalende winkel die ten onrechte geregistreerd worden. Dit zorgt ervoor dat de methode met een zekere opstartkost te kampen heeft. Desalniettemin is de totale kost zeer laag, zeker in vergelijking met de traditionele enquêtes.

Het tweede centrale hoofdstuk behandelt een grotere, outdoor omgeving. Hiervoor is Bluetooth tracking minder geschikt, waardoor CDR data gebruikt wordt. Heel wat tijdruimtelijke analyses vereisen de thuislocatie als startpunt voor verdere

analyses. Desondanks zijn de methodes om de thuislocatie te identificeren door middel van CDR data niet voldoende gevalideerd in de literatuur, door een gebrek aan validatiedata op een voldoende fijn niveau. Daartoe werden de bestaande heuristische methodes in dit proefschrift gebenchmarkt. De benchmarkstudie toonde aan dat de gemiddelde fout 4,4 kilometer bedroeg voor de beste heuristische methode. Dit proefschrift introduceert daarenboven een nieuwe heuristische methode en een methode gebaseerd op predictieve modellering. Het beste predictieve model reduceert de fout tot 2,8 kilometer. Dit resultaat toont aan dat een gelabelde, predictieve methode de voorkeur geniet van zodra het praktisch mogelijk is om deze toe te passen. Daarenboven werd ook aangetoond dat het inbrengen van variabelen gebaseerd op het sociale netwerk, de performantie verhogen. Dit onderzoek biedt een meer solide basis voor de vaak cruciale eerste stap om de thuislocatie te identificeren.

Het derde en laatste centrale hoofdstuk past de locatiegegevens die vervat zitten in CDR data, toe in een bedrijfscontext. Deze fraudeapplicatie toont de variëteit aan toepassingen voor locatiegegevens aan. De succesvolle business case toonde aan dat CDR data kan gebruikt worden om klanten te identificeren die het nieuwe telecom product op een oneigenlijke wijze gebruiken. Daarenboven vereiste het bedrijfsprobleem de ontwikkeling van een nieuwe analytische methode, aangezien er geen historische fraudedata beschikbaar was. Het methodologische probleem situeert zich daardoor in het gebied van *novelty detection*, het ontdekken van nieuwigheden en anomalieën in de data. Het onderzoek toonde aan dat de traditionele *one-class novelty* methodes niet toereikend waren in deze context van menselijk tijdsruimtelijk gedrag. Daarom werd de kennis van experts gebruikt om het *one-class* probleem te transformeren naar een meer standaard *two-class* probleem. Op deze manier kunnen traditionele, meer performante predictieve modellen gebruikt worden. Dit resulteerde in een sterke stijging van de fraude detectieratio van 8,56% naar 48,72%.

Summary

The analysis of spatio-temporal human behaviour is valuable both for research purposes and for applications in business. The analysis has strongly shifted from survey-based methods to actual tracking methods. Mobile phones provide a very well suited means for tracking humans. The main goal of this dissertation is to investigate how mobile phone data can be used to model spatio-temporal human behaviour. Two data sources are selected, the Bluetooth signal and the standard mobile phone signal. The first results into Bluetooth tracking data, the latter is known as call detail record (CDR) data. Both data sources are adequate for spatio-temporal analysis, by means of the proximity principle. In case of Bluetooth data, the registration of the mobile phone by a Bluetooth scanner indicates that the individual has been in the proximity of the scanner. For CDR data, the registration by a mobile phone tower indicates that the individual was in the proximity of the tower. The locations of scanners and towers are known and can therefore be used as close proxies for the actual location of the tracked individual. These raw CDR/Bluetooth data records as such are insufficient to derive meaningful insights. Therefore, a data analytical approach is applied to create value from these data sources.

The first main chapter in this dissertation investigates the value of Bluetooth tracking for modelling the spatio-temporal behaviour of people visiting an indoor shopping mall. The application requires a tracking method that is suited for indoor use as well as a high, store level, precision. The real-life experiment with 56 Bluetooth scanners demonstrated the applicability of Bluetooth tracking for this purpose. 9.81% of the visitors could actually be tracked. This rather high detection ratio ensures that Bluetooth tracking can quickly generate a large sample. Although the experiment revealed that Bluetooth tracking is a good approach, some manual adaptations might be necessary in order to avoid false positive registrations; i.e. registering visitors of a neighbouring store. Therefore, the method has a certain set-up cost. However, the overall cost is very low, certainly when compared to a survey approach.

The second main chapter deals with a larger, outdoor setting. For this setting, Bluetooth tracking is less suited and therefore CDR data is used. Many spatio-temporal analyses need the home location as a start point for further analyses. Nevertheless, the approaches for detecting home locations with CDR data have not been validated sufficiently in literature, due to lack of fine level validation data. Therefore, the existing heuristic methods were benchmarked. This benchmark study revealed an average error of 4.4 kilometres for the best method. On top

of that, this dissertation introduces both a new heuristic method and a predictive modelling approach. The best predictive model reduces the error to 2.8 kilometres. This demonstrated that a labelled predictive approach should be applied when possible. Furthermore, the inclusion of social network based variables further improved the performance. This chapter offers a more solid base for the often crucial first step of identifying the home location.

The third and last main chapter applies the location data, embedded in CDR data, to a business case. The fraud detection application indicates the wide variety of applications where location has a crucial role. The successful business case demonstrated that CDR data can be used to identify customers that use the new telecom product in a non-authorized way. Furthermore, the case required the development of a new analytical method, as no historical data about fraud had been observed. The methodological problem is therefore situated in the area of novelty detection. The research indicated that the traditional one-class novelty methods were not satisfactory in this setting that deals with human spatio-temporal behaviour. Therefore, expert knowledge was used in order to transform the one-class problem into a two-class problem, so that more traditional, better performing methods could be used. This boosted the fraud detection rate from 8.56% to 48.72%.

1

General Introduction

Will you allow www.randomsite.be to access your location? This notification pops up when visiting numerous websites. Why would any website be interested in a visitors location? Because this information is valuable. For their business, for marketing insights and for applications. The website or parts of it can be tailored to your preferences in order to enhance your comfort as well. Consider how many people daily use one or more location based smartphone apps, how often social media posts include a location tag, how often GPS is used in cars or smartphones. Location matters, it is a crucial aspect of human behaviour.

The high importance and relevance of human spatio-temporal behaviour makes it worth studying, both from an academic and a practitioners approach. The number of studies about human spatio-temporal behaviour is still growing at a high pace (Barbosa et al., 2018). Applications can be found in research about commute behaviour (Kung et al., 2014), the impact of mobility on our carbon footprint (Isaacman et al., 2011), traffic prediction (Lv et al., 2014), consumer research in marketing (Yamin and Ades, 2009) and epidemiological studies (Pfeiffer et al., 2008) amongst many others. Surprisingly, now that human mobility is strongly reduced due to the COVID-19 pandemic, location in general and tracking of spatio-temporal human behaviour has become even more important. The spatio-temporal analysis is crucial for epidemiological studies in order to model the spread of viruses as well as to study the impact of measures taken in order to deal with the COVID-19 crisis for example. The development of an application

such as Coronalert (Sciensano, 2020) is exemplary. This application uses the Bluetooth signal of smartphones, the topic of Chapter 2 in this dissertation, in order to estimate the risk of infection. This example already shows the applicability of Bluetooth for location based analysis, but many more options can be considered.

1.1 Data Sources for Tracking Spatio-Temporal Human Behaviour

Research in the field of human mobility was traditionally based on travel surveys, road side surveys and travel diaries. This survey based approach comes with major shortcomings such as small samples, short survey durations and under-reporting. Respondents may report inaccurate locations and times and report on a typical day rather than the actual day (Zhao et al., 2015). On top of that, these methods are labour intensive and very costly per respondent. These methods are usually not scalable, as their total cost grows linearly with every respondent.

Tracking methods have taken over from the survey based approach, thereby mitigating the main disadvantages of the traditional methods. With tracking methods, actual spatio-temporal behaviour can be captured. This overcomes the problem of under-reporting, which can occur due to oblivion of respondents as well as due to the social desirability bias (Grimm, 2010). At the same time, tracking methods usually have a lower cost and are scalable, which also solves the problem of low sample numbers. The small set of costly respondents in the traditional survey methods becomes a large set of low-cost tracked individuals.

However, tracking humans is technologically challenging. Tracking an individual as such is not straightforward, the use of a *proxy* is required. Mobile phones are considered as the most suitable proxy for this purpose. They are typically linked to a single individual, while this individual also consistently carries the mobile phone very close by, in his/her pocket or purse. This means that the mobile phone very closely follows the actual location of the individual. Moreover, mobile phones reach worldwide penetration rates of up to 96% (Iqbal et al., 2014; Vanhoof et al., 2018c), which immediately solves the low sample problem as well. Therefore, mobile phones offer a consistent proxy for research on human spatio-temporal behaviour throughout the world (Kung et al., 2014).

An overview of the main characteristics of different tracking methods using mobile phones and surveys for spatio-temporal data collection is presented in Table 1.1.

The most obvious choice for location based analyses using mobile phones as proxies, would be using the signal of the Global Positioning System (GPS). For certain analyses, this can indeed be the optimal choice. However, GPS has some downsides that need to be considered. First of all, GPS will not work in an indoor setting (as considered in Chapter 2). Secondly, accessing this GPS data, requires the a priori installation of an application by users. This way, the tracked individuals become (self-)selected respondents again. Two other technologies in a mobile phone do not suffer from this drawback; Bluetooth and the standard mobile phone signal. The first leads to Bluetooth tracking data, as used in Chapter 2. The latter leads to Call Detail Record (CDR) data, as used in Chapter 3 and 4 of this dissertation.

Bluetooth tracking, the main topic of Chapter 2, uses the Bluetooth signal of a mobile phone. The most known example nowadays in Belgium is the Coronalert smartphone application (Sciensano, 2020). Similar to other Bluetooth tracking applications in research, this application is based on the *proximity principle*. If two devices detect each others Bluetooth signal, this interaction is stored. The strength of the signal can be used to more precisely estimate the distance between the two devices. The specific Coronalert application does require the installation of the app. Usually Bluetooth tracking does not require this, as it only uses the Bluetooth signal that is broadcasted by a mobile phone. In the marketing based application in this dissertation, customers are tracked based on their Bluetooth signal in a shopping mall. Again, the proximity principle is used. The proximity registration now happens not between two user devices, but between the user device (mobile phone) and a detecting device; a Bluetooth scanner. For the shopping mall application, Bluetooth scanners were set up in every store in order to study the spatio-temporal behaviour of the visitors. This method is especially suited in an indoor setting, but can also be used in an outdoor setting. The methodology is usually applied on a more limited scale, as there is a set-up cost for every Bluetooth scanner. However, in contrary to survey methods, scalability in terms of time does not require extra investment. Furthermore, Bluetooth tracking can provide precise results, as the range in which one scanner can detect devices is limited. An alternative for Bluetooth is Wi-Fi tracking (e.g. Musa and Eriksson (2012)). The technical set-up and collection of information is very similar to Bluetooth tracking. The choice between both technologies largely depends on the detection ratio; the number of mobile phones that can actually be tracked. The evolution of this detection ratio in the future will to a large extent determine which of both technologies might become the preferred one. Furthermore, hybrid approaches that combine both Bluetooth and Wi-Fi tracking can also

	Surveys	GPS (app-based)	Bluetooth (and Wi-Fi*)	Call Detail Records (CDR)
			Standard CDR: Calls and SMS	Enhanced CDR: 3G/4G/5G and ping data
Sample size	Limited sample size	Limited to moderately high (depends on number of users that install app and agree to tracking)	Quickly large sample size (due to moderate, but growing detection ratio)	Large sample size (entire telecom provider customer base)
Duration of data collection	Short survey duration	Enables longer period analyses	Enables longer period analyses	Enables longer period analyses
Reported/actual behaviour	Reported behaviour (e.g. under-reporting, social desirability bias)	Actual behaviour	Actual behaviour	Actual behaviour
Respondent participation	Participation required	Participation required	Non-participatory	Non-participatory
In-depth qualitative info	In-depth, qualitative questions possible	Can be enriched with in-depth survey based questions for a selected sample, using the app	Can be enriched with in-depth survey based questions for a selected sample	Can be enriched with in-depth survey based questions for a selected sample
Social network	Limited possibility	Limited possibility (user consent needed)	No social network information	No social network information.
Number of (tracking) data points	Low (limited number provided by respondent or by survey maker)	Very high	Moderate to high (depends on number of installed scanners or beacons)	Moderate to very high (minimally as high as standard CDR. Usually a lot higher)
Accuracy	Dependent on information provided by respondent	Very high accuracy	Moderately high accuracy in indoor setting. Very high accuracy with new Bluetooth 5.1 (which supports Radio Direction Finding).	Accuracy related to the distribution of cell phone towers in the area
Indoor tracking	No, but questions in survey about indoor behaviour are possible.	No indoor tracking (unless with infrequent and expensive GPS-repeaters)	Ideal for indoor tracking (and positioning)	Indoor visibility, but no accurate location tracking
Cost	High cost per respondent, linear cost increase per respondent.	Low cost (development of app, no hardware installation necessary)	Higher set-up cost (hardware: Bluetooth scanners), but low cost per respondent: scalable.	Low cost (data is collected anyway, no set-up needed)

Table 1.1: Main characteristics of different spatio-temporal data collection methods. The table is non-exhaustive, but aims to give an overview of the methods that are most relevant in light of this dissertation. *Wi-Fi tracking shares many similarities with Bluetooth tracking, as discussed in Section 2.2.2.4.

be implemented (Kao et al., 2017). Due to the strong resemblance between Bluetooth and Wi-Fi tracking, many of the insights derived from the Bluetooth application in Chapter 2 can be transferred to Wi-Fi tracking as well.

The characteristics of Bluetooth (and Wi-Fi) tracking are more or less opposite to the second data source investigated in this dissertation, call detail record (CDR) data. This approach is more suited for outdoor applications, can be operated on a very large scale (without extra investments), but provides a lower level of precision. Standard CDR data is the information that telecom providers capture, every time that a customer makes or receives a call / SMS. Enhanced CDR data adds observations every time a customer (or his or her phone) makes use of his or her mobile data connection (3G/4G/5G). Enhanced CDR data can also include so called ping data. Pings include for example automatic location updates requested by the telecom provider and location updates when crossing location boundaries. Every record in standard CDR data contains interactional aspects (a caller and receiver id) as well as temporal (timestamp and duration) and location aspects. Enhanced CDR data offers more observations, but the added observations do not provide additional information regarding the social network as no second party (e.g. the receiver of a call) is involved. The registration of a location in both cases is again based on the proximity principle. Whereas Bluetooth tracking requires the set-up of Bluetooth scanners, the detecting devices in this approach are the cell phone towers. The geographical coordinates of these towers are used as an approximation of the actual location of the individual. The fact that these towers have a much larger detection range and that they are more sparse explains the lower precision of a CDR approach, when compared to Bluetooth tracking with pre-installed scanners.

Both approaches share the advantage of being *non-participatory* in nature. Both in the case of Bluetooth tracking and recording CDR data, people are to a certain extent unaware of being tracked. This means that it can be expected that their behaviour will not be affected due to the data collection method, which results into an unbiased measurement of their spatio-temporal behaviour. Nevertheless, people might become more and more aware of being tracked either with Bluetooth or by means of CDR data. It remains to be investigated how this might affect their behaviour. Enabling Bluetooth and using all functions of a mobile phone leads to the highest comfort, hence the question becomes how much of this comfort people are willing to give up in order to safeguard their privacy and prevent being tracked. In this respect, apps such as Signal or communication services such as Sky ECC are gaining traction. Recent events indicated the high awareness of being tracked with regular mobile phones and networks

in criminal environments. The awareness of being tracked is probably less high for the general public and their willingness to adapt their behaviour can be expected to be even far less.

Moreover, remark that this non-participatory aspect does not imply that people could not be tracked individually. In the case of Bluetooth tracking, people can be uniquely identified by the registration of the unique Media Access Control (MAC) address of the device (Delafontaine et al., 2012). The same holds for CDR data, where every individual is uniquely identified by his or her mobile phone number. Clearly, this immediately raises important privacy concerns. For Bluetooth tracking, anonymity is preserved because a MAC address can not be linked to the actual identity of an individual person. In the case of CDR data, the actual phone number should not be used for research purposes and is therefore replaced with a hashed identifier (Knuth, 1973). Throughout the main chapters of this dissertation (Chapter 2, 3 and 4), the aspect of privacy will always shortly be touched upon, but never becomes the main focus of the different applications. Nevertheless, when dealing with this type of data, one needs to be aware of the aspect of privacy. Section 5.1.2 provides a deeper discussion regarding this aspect.

When using either CDR data or Bluetooth tracking data for location based applications, one needs to keep in mind that these data sources were originally not designed with this purpose in mind. Furthermore, the difference between primary and secondary data becomes relevant. Primary data is data collected with a specific (research) purpose in mind. It therefore provides a higher level of control over the data. Although Bluetooth technology was not designed with the purpose of tracking in mind, Bluetooth tracking does lead to primary data, as it is generated by an experimental set-up with a specific purpose in mind. Secondary data on the other hand refers to data that has been collected for other purposes. For example, CDR data is collected with the initial purpose of creating the bill for customers. This secondary data, such as CDR data, usually leads to larger sample sizes, but provides less control over the data when used for other purposes. It is therefore interesting to investigate how well CDR data is suited for spatio-temporal analyses after all. The data analyst, who aims to create value from these data sources needs to be aware of this aspect.

1.2 From Tracking Data to Modelling Behaviour

CDR data has been called the most important, game-changing data of the last decade for the analysis of human mobility (Barbosa et al., 2018). Gonzalez et al. (2008) and Song et al. (2010) reported in the leading journals

Nature and *Science* that human mobility is strongly predictable, by using CDR data. However, both CDR data and Bluetooth tracking data as such are raw data in essence. In order to unleash the full potential of both CDR and Bluetooth data and achieve the results Gonzalez et al. (2008) and Song et al. (2010) published, a value adding step is required. A human being is needed in order to interpret, annotate and model the data in such a way that the actual value can be extracted. The data analyst, data scientist, data engineer or data specialist, uses his or her knowledge, possibly in cooperation with a field expert, in order to use the data in a meaningful way that leads to the desired result.

The aspect of the human analyst, the human expert for the analysis of human behaviour is central in this thesis. All chapters deal with the analysis of human behaviour and require an analyst with the right tool set of methods to translate the data into meaningful insights. However, throughout this dissertation the focus will shift from the value of the data as such, to the more methodological aspect of the analysis. In Chapter 2, the main focus lies on the value of Bluetooth tracking data (as a means for marketing analysis). Chapter 3 is situated on the balance between the value of the raw CDR data and the investigation of a new heuristic and predictive method for the problem at hand. Chapter 4 uses the CDR data as an input, but the focus is strongly on the added value of the human expert and the development of a new expert methodology for the non-trivial one-class classification problem.

The data analyst can tackle the majority of the traditional applications with binary classifiers. Examples of this paradigm that will also be used in this dissertation are logistic regression, decision trees, random forest, AdaBoosting, neural networks and Support Vector Machines (SVM). In this work, these different binary classifiers will be used not so much to explore the difference in performance between themselves. Their goal will be to investigate both the robustness of the more general method presented in each chapter and to assess the predictive power of the underlying CDR and Bluetooth data. For example, in Chapter 4, the new two-class expert based method will be benchmarked against the one-class classification benchmark classifiers. The two-class expert based method does not specify a certain binary classifier. The different binary classifiers will be investigated as a robustness check for the general two-class expert based method. However, other research does focus on the specific differences between these binary classifiers. The interested reader can be referred to more focused literature about this aspect (e.g. Baesens et al. (2003)). Note that the difference in performance between these methods might be context dependent, which again stresses the importance of the human expert in order to correctly se-

lect the optimal approach.

Next to the binary classifiers, other methods that will be applied in this dissertation are one-class classifiers (e.g. one-class support vector machines), visualisation techniques, clustering methods, consumer behaviour metrics and decision rule based methods. The modern data analyst has a large set of options at his or her disposal in terms of software and programming languages. Both proprietary and free, open-source alternatives present advanced, well-equipped options. In this dissertation, the choice for open-source was made. In each of the three main chapters, R (R Core Team, 2020) and PostgreSQL will be the main tools of analysis. R is a programming language extensively used by statisticians and data analysts. PostgreSQL is a relational database system, with a strong reputation for reliability and performance. With respect to the cross-industry standard process for data mining (CRISP-DM) framework (Chapman et al., 2000), PostgreSQL will be used primarily for the first steps (business understanding, data understanding and mainly data preparation). R will primarily be used in the subsequent steps (mainly modelling, evaluation and deployment).

1.3 Tracking and Modelling Applications

The value of Bluetooth tracking and CDR data and the methods to turn the raw data into value will be examined in three different applications. These constitute the three main chapters of this dissertation. An overview of the structure is presented in Table 1.2. Note that this is a PhD dissertation by publication, meaning that this manuscript is a collection of individual research papers. Therefore, each chapter can also be read independently. Chapter 2 has been published in *Applied Geography* (Oosterlinck et al., 2017), Chapter 3 in *Expert Systems With Applications* (Oosterlinck et al., 2021) and Chapter 4 in the *European Journal of Operational Research* (Oosterlinck et al., 2020). In every chapter, a specific application will be selected. In Chapter 2 and 4 the methodology itself is not limited to the selected domain and many insights from this specific case can be extended to other applications. Chapter 3 is most focussed on one specific application. However, the results provide an important base as a primary step for subsequent analysis in the field of human mobility with CDR data.

Chapter 2 will investigate the value of Bluetooth tracking for modelling the spatio-temporal behaviour of visitors in a shopping mall. The shopping mall used to deploy a yearly survey in order to gather information about for example what stores customers visit, for how long and in which order they do this. As discussed before, it is clear that a survey approach will fall short

<i>Data</i>	<i>Application</i>	<i>Focus</i>
Chapter 2: <i>Bluetooth Tracking of Humans in an Indoor Environment: An application to Shopping Mall Visits</i> <i>Applied Geography (Oosterlinck et al., 2017)</i>	<ul style="list-style-type: none"> - Bluetooth tracking data - 19 days 	<ul style="list-style-type: none"> - Marketing application - Modelling visitor behaviour in a shopping mall - Bluetooth as tracking method - Positioned as marketing tool - Indoor - Limited range
Chapter 3: <i>Home Location Prediction with Telecom Data: Benchmarking Heuristics with a Predictive Modelling Approach</i> <i>Expert Systems with Applications (Oosterlinck et al., 2021)</i>	<ul style="list-style-type: none"> - Call Detail Record (CDR) data - 5 weeks 	<ul style="list-style-type: none"> - Home location detection - Data / Analytical Methodology - Outdoor (Indoor) - Validation of home detection methods - Large geographical range - Development of new methods
Chapter 4: <i>From One-Class to Two-Class Classification by Incorporating Expert Knowledge: Novelty Detection in Human Behaviour</i> <i>European Journal of Operational Research (Oosterlinck et al., 2020)</i>	<ul style="list-style-type: none"> - Call Detail Record (CDR) data - 2 x 5 weeks 	<ul style="list-style-type: none"> - Fraud application - Analytical Methodology - Home location detection - Household relation detection - Development of new predictive methodology for novelty detection by including human expert knowledge - Outdoor (Indoor) - Large geographical range

Table 1.2: Structure of this dissertation. Modelling spatio-temporal human behaviour is approached from different angles. The mobile phone will be used as the proxy for tracking humans in all applications. In a limited indoor setting, the Bluetooth signal is proposed, whereas the advised data source becomes CDR data in cases with a much larger geographical range. The focus in this PhD dissertation shifts from more data oriented towards a higher focus on the analytical methodology over the chapters.

when it comes to measuring actual movement of the visitors. In this indoor setting, Bluetooth is proposed as the most suited technology for actually tracking customers with individual store level precision. A real-life test experiment with 56 Bluetooth scanners is established. The main research goal is to examine the applicability of Bluetooth technology for tracking purposes in an indoor setting. As the experiment takes place in a shopping mall, there is also focus on the value of Bluetooth tracking in a marketing context. Prior research has already pointed at the relation between shopping path lengths and sales volume for example (Kholod et al., 2010). The analysis is relevant for shopping malls in order to improve their store lay-out. It can be explored which stores attract many visitors and how they should be positioned in the shopping mall. For example, it might be beneficial to position an anchor store (a store that attracts the majority of the customers) further from the entrance, as this forces the customers to also pass along other stores, thereby increasing the chance of them visiting these otherwise less visited stores. Linking this tracking data with point-of-sales data results into clear business results. These insights can drive the revenue for the stores and the shopping mall. The chapter positions Bluetooth tracking as an alternative marketing tool. However, the experiment was set up with the more general idea of tracking humans in an indoor setting, which makes that the results reach further than the specific marketing based application and attain the main goal of investigating the value of Bluetooth tracking in an indoor setting.

In Chapter 3 the shift is made from a limited indoor setting to a much larger outdoor setting. This requires the use of another approach, which can be found in the use of CDR data. The much larger geographical range is the entire area covered by the selected telecom provider. Literature about spatio-temporal human behaviour revealed the importance of the home location as an important first step for analyses on a large scale. Therefore, accurate home detection is crucial for human mobility analysis with CDR data. The main goal of this study is to provide a solid basis for home detection with CDR data. Firstly, this study improves strongly on the validation aspect, whereas literature failed to properly validate the existing heuristic methods for home location prediction with CDR data by using either too coarse-grained aggregated validation data or very small validation sets. By using a unique data set that contains five weeks of anonymised CDR data that enables ground truth validation, the aim is to strongly improve on prior validation efforts. Secondly, the existing rule based heuristics for home location prediction with CDR data are benchmarked by means of this strongly improved validation data. Thirdly, based on the insights from the benchmark study, a new rule based heuristic is proposed. Fourthly, the

chapter also introduces a labelled predictive modelling approach. Here, the dissertation shifts towards a more methodological focus in terms of analysis. The labelled model also enables to provide an estimate of the maximal performance that can be expected. Lastly, as mentioned in Table 1.1, standard CDR data also has the richness of the embedded social network data. Therefore, the labelled predictive model incorporates social network variables in order to further improve the prediction, along with stand-alone models that use merely the information of others in the network to estimate the home location.

Chapter 4 deals with an application of fraud detection in a new telecom offer (Baesens et al., 2015). The goal of the application is to identify subscription fraud for a new *family plan* (Gosset and Hyland, 1999; Farvareh and Sepehri, 2011), by using CDR location information. It will be examined whether CDR data and the embedded location information can be used to identify fraud. However, the main problem is that no fraud has been observed when starting the analysis, only non-fraudulent behaviour has been captured. This makes that the traditional labelled two-class classification approaches can not be applied as such. The focus in this last chapter will therefore shift even more strongly towards the methodological aspect of the analysis. Hilaris and Mastorocostas (2008) define fraud detection as a field that uses techniques to monitor behaviour that deviates from the norm. Therefore, techniques from the field of *novelty detection* are advised. The fact that no labelled examples of the novel fraud are yet observed in this case, requires the use of one-class novelty detection methods. These novelty detection methods are advised in this situation with no labelled data, but are also relevant in cases with partly labelled data. A broad range of methods have been proposed for handling this type of data, and are typically also known under the name of anomaly or outlier detection. Nevertheless, this research will reveal that these methods do not function well in a case that deals with human behaviour. More traditional applications of novelty detection typically have a set of normal data that is rather stable, combined with pronounced outliers. However, human behaviour leads to tracking data that is much less stable, more volatile and varied in nature. Hence, the call for new methods. It will be examined how the inclusion of human expert knowledge can be used to enhance the data and lead to the creation of a new approach for novelty detection, where the boundary between the normal and the novel data becomes supported from two sides. The newly developed expert based two-class model will be applied on a real-life test set. Two separate periods of 5 weeks of anonymised CDR data will be used as data source. Furthermore, the specific type of fraud to be detected in this case shows a strong link with the home detection applica-

tion in Chapter 3. Whereas Chapter 3 focusses on the detection of the home location of an individual, Chapter 4 will predict whether two individuals do share their home location. The approach can therefore also be considered a re-visit of Chapter 3 from a different angle, where this chapter puts stronger emphasis on the methodology and social network analysis. In summary, the main goals of this chapter are to investigate how expert knowledge can be used to convert an one-class problem into a two-class model and examining the use of (location) CDR data in a fraud case.

2

Bluetooth Tracking of Humans in an Indoor Environment: An Application to Shopping Mall Visits¹

Abstract

Intelligence about the spatio-temporal behaviour of individuals is valuable in many settings. Generating tracking data is a necessity for this analysis and requires an appropriate methodology. In this study, the applicability of Bluetooth tracking in an indoor setting is investigated. A wide variety of applications can benefit from indoor Bluetooth tracking. This paper examines the value of the method in a marketing application. A Belgian shopping mall served as a real-life test setting for the methodology. A total of 56 Bluetooth scanners registered 18.943 unique MAC addresses during a 19-day period. The results indicate that Bluetooth tracking is a sound approach for capturing tracking data, which can be used to map and analyse the spatio-temporal behaviour of individuals. The methodology also provides a more efficient and more accurate way of obtaining a variety of relevant metrics in the field of consumer behaviour research. Bluetooth track-

¹Based on: Oosterlinck, D., Benoit, D., Baecke, P., & Van de Weghe, N. (2017). Bluetooth tracking of humans in an indoor environment: an application to shopping mall visits. *Applied Geography*, 78, 55 - 65.

ing can be implemented as a new and cost effective practice for marketing research, that provides fast and accurate results and insights. We conclude that Bluetooth tracking is a viable approach, but that certain technological and practical aspects need to be considered when applying Bluetooth tracking in new cases.

2.1 Introduction

Collecting data to develop insights into the spatio-temporal behaviour of individuals can be of interest for many domains. Crowd management, safety management, operational research and consumer research in marketing provide examples of such domains (e.g. Yamin and Ades (2009)). The specific application discussed in this paper is situated in the domain of marketing research. However, the methodology itself is not limited to this domain and many insights from this specific case can be extended to other applications. We present the case of a Belgian shopping mall that wants to track customers in order to obtain insight into their behaviour. The traditional approach is periodically (e.g. every six months) hiring market research firms that survey random customers of the shopping mall about their shopping trip. Information collected this way inherently suffers from various drawbacks, such as inaccurately reported information and sample bias (Andres, 2012). Tracking methods overcome many of the disadvantages that characterize the more traditional methods. Actual paths, exact time measurement and other high quality statistics can be obtained. Still, surveys have some advantages that cannot be ignored, such as the possibility to gather more in-depth information about certain consumer preferences and characteristics (e.g. age, gender). Depending on the desired kind of information and the available budget, an adequate combination of both methods can be used.

Tracking individuals requires some form of identification. Previous research mainly used Radio Frequency Identification (RFID) technology for this purpose (Liao and Lin, 2007; Kanda et al., 2007; Hurjui et al., 2008; Takai and Yada, 2010; Fujino et al., 2014). However, unlike Bluetooth tracking, this methodology suffers from various specific drawbacks (see Section 2.2.2.3). This research is to our knowledge one of the first applications of Bluetooth tracking in a real-life indoor retail setting. The real-life use-case examines and demonstrates the value and possibilities of the method. The findings have implications that go beyond the specific setting of a shopping mall.

In the remainder we discuss the application of Bluetooth technology with the purpose of tracking humans in an indoor environment. In Section 2.2 we review literature concerning tracking in general. We also discuss the

use of Bluetooth for positioning purposes and touch upon developments in indoor positioning in general. Section 2.3 discusses the design of the study, along with some important preceding test cases that were used to optimize the methodology. The results are reported in Section 2.4. Section 2.5 summarizes the main findings, while we discuss some limitations and suggestions for further research in Section 2.6.

2.2 Literature Review

2.2.1 Customer Tracking

Researchers have studied the movement of people in different contexts and on various scales. Specifically for marketeers, tracking data provides valuable information. Larson et al. (2005) state that a better understanding of the shopping process could lead to important discoveries for retailing. Knowledge about in-store customer behaviour by means of path tracking is beneficial for improving Customer Relationship Management (CRM) and streamlining store operations (Celikkan et al., 2011; Vukovic et al., 2012). Already in 1966, it was clear that store layout impacts shopping behaviour (Farley and Ring, 1966). Sorensen (2003) provides an illustrative example by demonstrating that a store entrance on the right side generates an average boost in sales of two dollars per customer. This placement of the entrance favours counter-clockwise movement in the store, which apparently leads to higher sales. Knowledge about customer trajectories is a treasured input for the process of optimizing store configuration. Observing customers by means of tracking their path makes it possible to position products and stores in a way that increases sales and for example draws customer attention to previously less attracting parts of the store (Vukovic et al., 2012). Customer tracking unveils information about what sections people like, dead spots and favourite recurring spots. It can also be used to better estimate and improve many key performance indicators in the field of marketing. The fact that there is a market for these ideas is illustrated by the establishment of companies that deliver exactly these insights to their customers. Examples are Senion (Senion, 2016), Crossscan (Crossscan, 2016) and Bluenion (Bluenion, 2016). These companies mainly use Bluetooth and Wi-Fi signals. More about these technologies in the context of tracking can be found in the next section (2.2.2). Apple's Bluetooth based iBeacon is another player in this market (Apple, 2016a). iBeacon is frequently used for push marketing applications, where customers get notifications when they are close to such beacon. Nevertheless, this technology could also be used for tracking and positioning purposes. However, this requires the in-

stallation of an application by users. An important advantage of the method discussed in this research is that it does not require this.

Tracking humans is technologically challenging. Research therefore commonly makes use of proxies. These are devices or objects that can be tracked more easily and are therefore used as an approximation of the location of the user. Shopping carts are a commonly used proxy for the movement of customers (Sorensen, 2003; Larson et al., 2005; Celikkan et al., 2011). In this case, the shopping cart is equipped with an RFID-tag in order to enable localisation. The dependence on a shopping cart however limits the generalisability of the method. We therefore suggest mobile phones as a more suitable proxy, since almost every individual carries it along at all times. Also, a mobile phone is typically linked to one individual only and people typically have it in their pocket, meaning that the location of the phone very closely follows the location of the individual.

2.2.2 Methods for Tracking

Traditionally, researchers and practitioners had to rely on methods such as shadowing (Quinlan, 2008; Millonig and Gartner, 2008), collecting travel diaries (Axhausen et al., 2002) and surveys. These methods can provide rich information, but are very labour intensive. Liebig and Wagoum (2012) state that surveys are the most common method for gathering customer information, but that the high cost and the low representativeness due to the non-random sampling process strongly deteriorates its practical value. Another method that is used to track customers is the use of light curtains. A light curtain consists of two poles and is usually found at the entrance of a store. When a customer passes through the light curtain, the invisible light beam is interrupted and the visit of a customer is registered. This method is very limited, since it cannot uniquely identify customers and is therefore basically limited to visitor counts. It is impossible to construct paths, nor is it possible to calculate duration of stay or to obtain information about return behaviour. Video tracking is another approach for gathering information about customer behaviour. Saxena et al. (2008) argue that the use of video data to track individual movements forms a challenging task. The need for manually filtering out the information renders this method too labour intensive (Versichele, 2014), whereas computer vision-based alternatives need to make use of advanced image processing algorithms that may not always be a straightforward option (Celikkan et al., 2011). The most valuable methods for tracking are based on wireless signals. In the remainder of this section we discuss GPS, RFID, Bluetooth and Wi-Fi

2.2.2.1 GPS

Being explicitly designed for positioning purposes, GPS possesses interesting characteristics such as high accuracy. Unfortunately, there is no GPS coverage in an indoor environment. To overcome this issue, GPS repeaters have been developed. An outdoor GPS antenna, generally installed on the roof of a building, is connected to an indoor antenna which re-transmits the signal. Nonetheless, this solution is not sufficient in the case of tracking customers. The re-transmitted signal is limited to the location of the outdoor antenna and this would imply installing multiple, expensive repeaters per location of interest. Furthermore, this method is not applicable in a multi-level building, such as a shopping mall. Locata is another proposed solution to the indoor limitations of GPS. This local positioning system emits signals that are equivalent to GPS signals for the receiving device (Rizos et al., 2010). The system is however quite expensive. It is also more aimed at indoor positioning, not so much at tracking (see also 2.2.3 Indoor Positioning).

2.2.2.2 RFID

Several researchers used Radio Frequency Identification (RFID) as a technology for tracking humans (Liao and Lin, 2007; Kanda et al., 2007; Hurjui et al., 2008; Takai and Yada, 2010; Kholod et al., 2010; Fujino et al., 2014; Kaneko and Yada, 2016). Due to its high positional accuracy and its applicability in indoor settings, RFID promises to be a valuable option for tracking customers in a marketing context. The methodology is however limited due to the fact that people do not simply carry an RFID-tag along. Therefore, researchers have equipped shopping carts with RFID-tags (Sorensen, 2003; Larson et al., 2005; Nakahara et al., 2010; Jung and Kwon, 2011; Vukovic et al., 2012). The advantage is that a high percentage of shoppers will be tracked. This approach however strongly limits the generalisability of the method to settings where shopping carts (or similar proxies, cf. 2.2.1) are available. In addition, the sample of tracked customers might be biased, since the methodology excludes shoppers without shopping cart. Customers may also leave their shopping carts behind at times, for example to quickly grasp something from a previous aisle without having to move the cart. Hence, the location of the shopping cart does not resemble the location of the customer at all times. Celikkan et al. (2011) also mention the high cost of using RFID, since both scanning and scanned devices (tags) need to be installed.

2.2.2.3 Bluetooth

Bluetooth has been used for tracking spatio-temporal behaviour in various domains and settings: to measure throughput time in airport security (Bullock et al., 2010), on a large scale open air festival (Versichele et al., 2012a), on a trade fair (Delafontaine et al., 2012), in an office setting (Abedi et al., 2014) and to study the movements of tourists in a city (Versichele et al., 2014a) amongst others. However, this study is to our knowledge the first application of Bluetooth tracking in an indoor retail setting.

The methodology of Bluetooth tracking uses the wireless Bluetooth signal that is a feature of most mobile phones. Bluetooth scanners are being installed at locations of interest. These scanners continuously scan for Bluetooth signals emitted by the mobile phones and register every detection. The method is based on the *proximity principle*. This means that the location of the device is approximated by the location of the scanner. Practically this means that if a device is detected by a certain scanner, this is registered in the log of that scanner. The path can then be reconstructed by combining the logs of all scanners, where the location of each scanner is added as the location of the device.

The Bluetooth methodology facilitates non-participatory tracking of customers (Versichele et al., 2012a). The non-participatory aspect means that ‘participants’ are unaware of being tracked. As such, they do not need to invest any kind of effort, which would be the case with a survey or with methods that require the installation of a smartphone application (e.g. iBeacon). It also means that the method will have no influence on their behaviour. Bluetooth tracking thus enables unbiased experiments and uninfluenced observations (Delafontaine et al., 2012). Celikkan et al. (2011) argue that in certain contexts it is acceptable that people carry a tag. A method that can avoid this dependency is however more valuable for marketing purposes. The non-participatory aspect does not mean that people cannot be individually tracked. Individual identification is made possible by registration of the unique Media Access Control (MAC) address of the device (Delafontaine et al., 2012).

Bluetooth tracking is especially useful for studying movements in ‘uncontrollable’ settings. We define uncontrollable settings as settings in which individuals of interest can freely move in space, without any obvious means of identification. In a regular supermarket for example, a shopping cart can be used as a proxy. In the setting of a festival, tags can be included in wristbands. In a shopping mall however, there is no suited proxy and the setting can be defined as uncontrollable. Many public places, such as train stations, libraries and museums can be defined as uncontrollable. Tracking

individuals becomes hard in these settings and Bluetooth tracking provides an excellent methodology to keep track of the spatio-temporal behaviour of these actors. Versichele et al. (2012a) state that the methodology enables studying visitor flows at mass events. Bluetooth tracking is thus able to handle large amounts of observations. When compared to other tracking methodologies, Bluetooth can be characterized as cost effective (Abedi et al., 2014). The implementation is low cost and easy (Delafontaine et al., 2012), we nonetheless argue that the implementation should be executed with great care.

An aspect to keep in mind is that interference can affect the detection range of the Bluetooth scanners. Physical objects, radio-frequency, electronic devices and environmental factors cause different levels of interference (Harwood, 2009). Metal constructions have the strongest impact on the signal, as Agostaro et al. (2004) demonstrate. Also, Bluetooth technology was not developed with tracking or localisation purposes in mind; it is therefore not surprising that it does not have the same positional accuracy as other technologies such as outdoor GPS (Delafontaine et al., 2012). There are however ample situations in which room-level accuracy is satisfactory. Combined with the low cost, this technology will often be an interesting alternative to the genuine location-aware technologies.

Bluetooth has three operational states; off, on-invisible, on-visible (Abedi et al., 2015). It is the user of the mobile phone who chooses one of these states. The default setting varies amongst brands and devices. Only the latter of the three states is detectable for the Bluetooth scanners. This makes that only a certain ratio of the population will actually be tracked. Researchers found different detection ratios, 7% (O'Neill et al., 2006; Hagemann and Weinzerl, 2008), 11% (Versichele et al., 2012a), 13% (Versichele et al., 2012b) and 8% (Versichele et al., 2014b). The fact that users are able to switch to an invisible state is negative for the sample size in a non-participatory approach. However, this also means that in settings where participation is allowed or even required, people can easily be asked to switch to the on-visible state.

2.2.2.4 Wi-Fi

Mazuelas et al. (2009), Bonne et al. (2013), Danalet et al. (2014) and Abedi et al. (2015) demonstrated the use of Wi-Fi for tracking purposes in various settings. Tracking by means of Wi-Fi signals is very similar to the Bluetooth approach. Both methods use the unique MAC address to identify individuals and most (dis)advantages are also applicable in the case of Wi-Fi. In terms of data, the same information can be captured. When

compared to Bluetooth, Wi-Fi tracking might have a lower accuracy, since Bluetooth is more specifically designed for short range wireless connections. Furthermore, people might be less likely to use Wi-Fi when they are away from home, since they can use mobile broadband more conveniently, whereas there are no immediate alternatives for Bluetooth, definitely not in the same device. This could lead to the conclusion that Bluetooth is the more mobile technology of both and therefore lends itself better for tracking purposes. However, Abedi et al. (2015) report a higher detection ratio for Wi-Fi, compared to Bluetooth. The conclusion is that both technologies are very analogous with regard to the application of tracking. Our assessment of the methodology with Bluetooth scanners therefore also provides insights for the use of Wi-Fi.

2.2.3 Indoor Positioning

Researchers have suggested Bluetooth not only as a tracking, but also as a positioning technology (Anastasi et al., 2003; Feldmann et al., 2003; Hallberg et al., 2003; Madhavapeddy and Tse, 2005; Rodriguez et al., 2005). For both purposes, the same Bluetooth signal and similar devices can be used. Tracking and positioning however differ to the extent that the initiator is different. Whereas for tracking the initiator is an external party (e.g. shopping mall management) that wants to gather information about the location of devices (e.g. as a proxy for a customer), for positioning, the initiator is the smartphone owner. The smartphone owner wants to know his or her own position, usually in order to reach a certain destination. In positioning cases people need to install an application, which then calculates their position with respect to static base stations.

The location of the Bluetooth scanner can be used as an approximation of the location. This approach is known as the *proximity principle* (Bensky, 2007) and also often referred to as the node-based approach (Nolte and Lynch, 2007). A more complex approach uses Received Signal Strength Indication (RSSI) values. These values are negatively correlated with distance (see also Figure 2.6). Therefore, it is theoretically possible to calculate a reasonably accurate location in the presence of multiple base stations. Multiple stations are required in order to enable some form of multilateration, triangulation or fingerprinting (Agostaro et al., 2004). The basic principle of these techniques is that if one knows the distance (approximated by the RSSI value in our case) to multiple points, it is possible to calculate a more refined estimate for the actual position. Several authors argue that RSSI is not a straightforward means to obtain accurate positioning (Soh et al., 2007; Versichele et al., 2010). The main problem is that the 2.4 GHz

frequency that Bluetooth uses, behaves unstable (Helen et al., 2001). Helen et al. (2001) therefore argue that the use of RSSI values as such is insufficient. Applying an Extended Kalman Filter (EKF) is suggested to improve accuracy. EKF is the nonlinear interpretation of an ordinary Kalman Filter (KF). Kalman Filtering uses Bayesian inference to estimate a result from multiple measurements, in order to be more precise than a single measurement. It is a widely used concept in navigation software (Chen et al., 2014b). The search for more than room-level accuracy might however be irrelevant in many applications (Cheung et al., 2006; Hay and Harle, 2009; Fernandes, 2011). In this research for example, room-level accuracy was sufficient and therefore the proximity principle could be used. The low cost characteristic of Bluetooth and the simplicity of the proximity principle make it a compelling option in situations where the trade-off between cost and accuracy balances in favour of cost.

Gu et al. (2009) argue that few commercial solutions for indoor positioning are available and that they are costly and complex to install. The same companies that are involved in indoor positioning often also provide tracking applications. As mentioned before, the technological part of both applications is usually quite similar. A distinction can be made between initiatives that focus mainly on the software side versus the ones that focus mainly on the hardware side. The first category comprises initiatives that participate in indoor mapping. Examples are Google Maps Indoor (Google, 2016), Apple Indoor Maps (Apple, 2016b) and Open Street Maps (OSM, 2016). The more hardware focussed initiatives either use existing technologies, develop new ones or make a combination. Whereas GPS is the standard technology for outdoor positioning, there is no such standard at this point in time for indoor positioning. The question remains if and when there will be such technology. The answer might lie in a combination of different technologies, where Bluetooth can be one of these. Given such a standard, another question is whether the approach will be infrastructure-free or infrastructure-based. Leveraging existing Wi-Fi networks or using geo-magnetic and sound signals by means of fingerprinting is infrastructure-free. Google Maps Floor Plan Marker is a good example here. Fingerprinting maps the space in terms of signal strength. Based on these registered values, it is possible to make a reasonable estimation of the location of a certain device. Installing beacons that use RFID, Bluetooth or other technologies are infrastructure-based approaches. The Microsoft Indoor Localization Competition provides a yearly test field for both infrastructure-free and infrastructure-based solutions to indoor positioning (Lymberopoulos et al., 2015). In the 2014 edition, a Bluetooth based solution ranked 10th out of 22 in a ranking that is mainly determined by accuracy and to a smaller

Class	Radius (m)	Max. permitted power (mW)	Use
Class-1	100	100	Mainly industrial use
Class-2	10	2.5	Mobile phones, car kits,...
Class-3	1	1	Very short range devices (e.g. keyboard, mouse,...)

Table 2.1: Bluetooth Classes

extent by cost and simplicity.

2.3 Experimental Design

2.3.1 Equipment

The scanners in this research make use of Bluetooth signals. Bluetooth is situated in the ISM-band (Industrial, Scientific and Medical) at a frequency of 2.4 - 2.485 GHz (Bhaskar et al., 2013). The ISM-band is shared by other technologies, such as Wi-Fi and NFC. Frequency hopping is used in order to bypass interference in this ISM-band (Bensky, 2007). Bluetooth devices are categorised into three classes, depending on their energy level (see Table 2.1). Most portable devices are Class-2, with a detection range of approximately 10 metres and an energy level of 2.5 mW.

The Bluetooth scanners consist of three parts: a calculating unit, a storage unit and a Bluetooth sensor. The first part is detailed as follows: an Alix 2D2 embedded board equipped with a 500 MHz AMD Geode LX800 processor and 256 MB DDR RAM. The storage unit comprises a 1 GB Industrial CompactFlash (CF). Three different Bluetooth sensors are used in this research: D-Link DBT-122 (Class-2), Sena UD100 (Class-1) and Trust 15542 (Class-2). A USB cable connects the sensor to the motherboard. The entire device is enclosed within a protective plastic cover. A 18V power cord adapter provides the scanner with electricity. (See Figure 2.1).

Our scanning software is installed onto the CF card. The scanners are programmed in a way that they constantly scan for ‘visible’ devices within their range. The scanner subsequently writes the retrieved MAC address to a log file on the CF card. Along with the MAC address, a time stamp, state and device class are registered. The MAC address is a 48-bit long number that uniquely defines the device. The first six hexadecimal digits identify the manufacturer of the tracked device (also called the Organizationally Unique Identifier). The last six hexadecimal digits represent the



Figure 2.1: Bluetooth Scanner

series number (Bhaskar et al., 2013). In order to avoid privacy issues, the ‘friendly name’ of the device is discarded (cf. 2.5. Discussion). In our experiment, the registered state can take on three distinct values: in, out or pass. ‘In’ represents a first detection of a certain MAC address by scanner, ‘out’ represents the last detection. The ‘pass’ state is used for solitary detections. This way of encoding the data strongly reduces the amount of data, with zero loss of valuable information. The RSSI values are stored in a separate log file.

The log files can easily be transferred through connecting a USB stick to the scanner. The data is manually transferred every two days in order to quickly identify problems. In a more continuous setting, data could also be transferred by integrating the scanners in a network. This approach is used in Versichele et al. (2014a).

2.3.2 Description of Study Area

The actual study took place in a Belgian city centre shopping mall, during a three-week period. The shopping mall houses 39 commercial premises, divided over a below surface, ground level and first floor storey. Higher floors within the building are devoted to offices and apartments. Level -3 and -2 accommodate parking facility, connected to the shopping mall by elevators and staircases. There are three main entrances and one side entrance to the shopping mall.

The 56 scanners were set up in strategic places. Scanners were placed in the commercial premises, at the entrances of the mall, in the parking facility and around the offices. The latter for example, makes it possible to distinguish non-customers, who only pass through the shopping mall to

reach their offices. To investigate whether enough devices could be detected to make meaningful conclusions, a preliminary test was performed. During two full days, a scanner placed at the main entrance scanned 3,491 unique MAC addresses. An extensive part of this number represents passers-by. Nevertheless, this result indicates that a sufficient sample size can be obtained with this approach.

2.3.3 Test Cases

Previous research used similar equipment (Versichele et al., 2012a, 2014b). These experiments were executed in an outdoor environment and with a far less dense spreading of the scanners. It is therefore crucial to pre-test certain aspects of the methodology, before initiating the actual shopping mall experiment. Three test cases were developed. The main goals were to identify issues and investigate the applicability of the *proximity principle*. In other words, we investigate whether using a detection by scanner x (located in store x) as an indication of an actual visit to store x is a valid approach. The first two test cases took place in stores of the shopping mall, whereas the third case took place at the research facility.

The first test case was executed in one store of the shopping mall (Store III, see Figure 2.2). The D-link DBT-122 Class-2 sensor should theoretically detect devices within a range of 10 metres, as displayed in the figure by the sphere. The used test device (Nokia E71) was nonetheless detected far beyond this theoretical range. The most distant captured point (bottom right in Figure 2.2), is almost 30 metres separated from the scanner. The Nokia E71 was not detected in neighbouring Store II, due to the fact that a load bearing wall separates both stores. The scanner did however register the test device in neighbouring Store IV. In this case the wall is not load bearing. These preliminary findings stress the importance of tuning the methodology when applied in an indoor environment.

Since Test Case 1 also revealed the possibility of tracking devices outside in the adjoining street, the same ground level store was selected for the second test case, along with its two neighbouring stores (see Figure 2.3). The choice for ground level also enables the assessment of possible detections above and below the selected stores. The Bluetooth scanners are positioned deep in the store, in order to limit the number of detections of people who just pass by the entrance. Few devices were tracked in Store II due to the fact that the scanner was placed above metal shelves. The test case revealed that the RSSI values of the registered MAC addresses fluctuate to an extent that it is not straightforward to determine whether a captured device is actually situated inside the store. This issue is input for

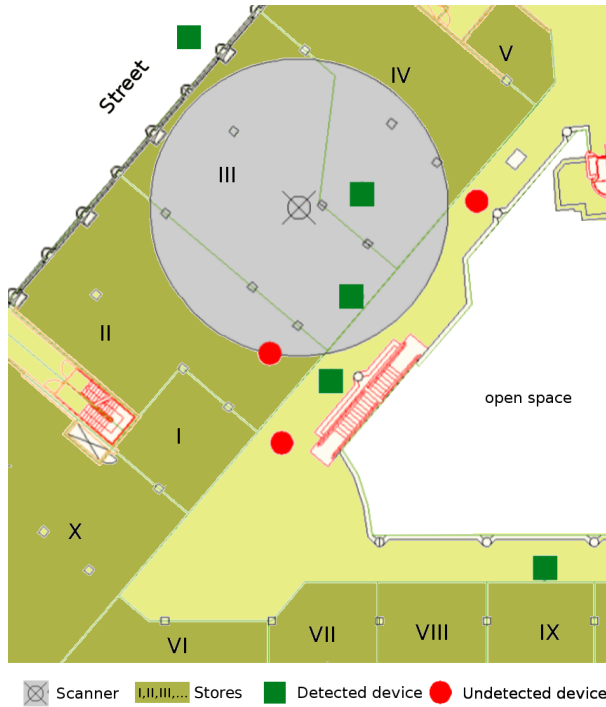


Figure 2.2: Test Case 1: Detection of Bluetooth Signal. The test device was detected outside the theoretical range. Squares indicate detections, whereas the circles indicate places where the device could not be detected.

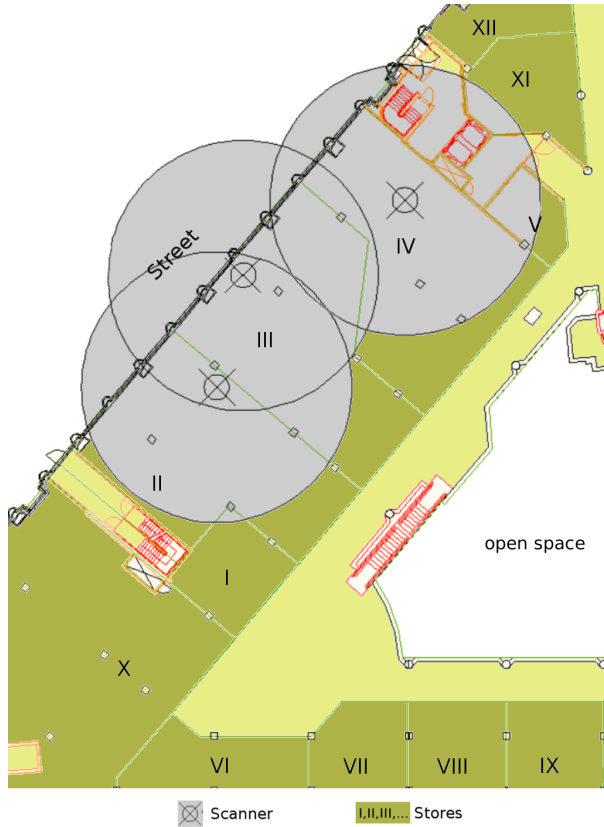


Figure 2.3: Test Case 2. Positioning of scanners in three neighbouring stores.

Test Case 3, in which a solution is proposed. Nevertheless, one has to be aware of the fact that even after applying that solution, noise in the data is inevitable. Type I and Type II errors can occur. Type I, the more likely case, comprises devices that are registered as having visited a certain store, while in reality the person visited the neighbouring store or was only window shopping. A Type II error occurs when devices that did enter the store are not registered. This error type is however unlikely to occur.

The third test case investigated the relation between distance and RSSI. An indoor office space served as the first indoor test setting. The space was measured by means of a Pentax R-325. Forty-five markers on the ground were measured (see Figure 2.4). A scanner with a D-link DBT-122 sensor was used to track the Nokia E71 test device. Figure 2.4 displays the measured values in terms of dBm. dBm is an alternative to RSSI that also

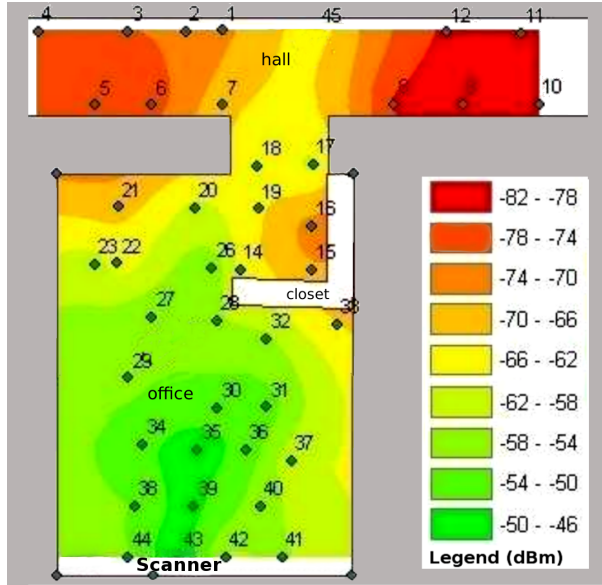


Figure 2.4: Test Case 3a: Interpolation of RSSI values. The figure shows a decay that is not perfectly circular, but a clear directionality is present. The figure maps an office space and a corridor, separated by a wall. The forty-five numbered points indicate the discrete points where signal strength was measured.

expresses signal strength. The only difference between both units of measurement is that RSSI is a relative value, whereas dBm represents absolute power levels, expressed in milliwatts. The closer dBm is to zero, the stronger the signal. Figure 2.4 is generated through interpolation of the forty-five measured values. An important finding is that the decay is not perfectly circular, but a clear directionality is present. The mirrored L represents a metal closet and shows the influence of obstacles on the signal. This aspect is crucial to keep in mind when deploying the methodology in an indoor setting, which is usually characterised by all kinds of obstacles (Feldmann et al., 2003; Zhou and Pollard, 2006). This aspect can also be used to the advantage of the researcher, by creating or making use of obstacles to define the area. A second measurement in the same room resulted in a very similar map.

A simple, but effective practice can significantly enhance the workability of the proximity principle. Wrapping the Bluetooth sensor with aluminium foil creates the effect of a Faraday cage (Cheung et al., 2006), which results in a much faster decay of the signal. This strongly reduces the

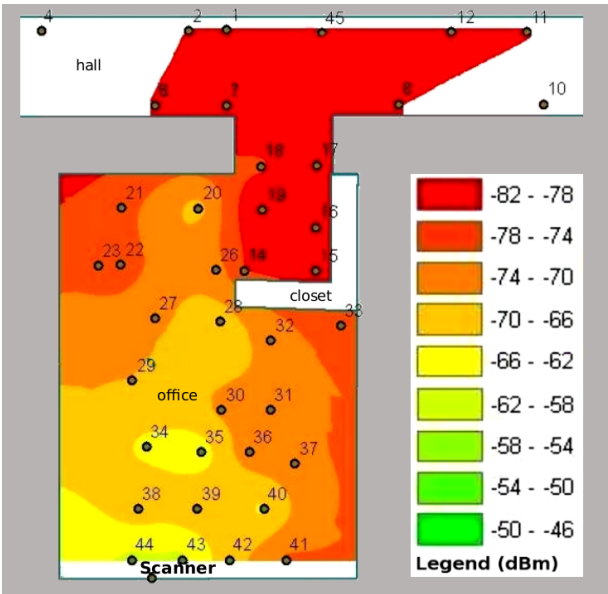


Figure 2.5: Test Case 3b: Interpolation of RSSI values, scanners wrapped with aluminium foil.

chance of tracking devices outside the room of interest. With this simple trick, it is much more straightforward to define a threshold for the RSSI. The resulting values can be found in Figure 2.5.

In order to test the relation between RSSI and distance without the interference of objects and materials that are typically found indoor, an outdoor evaluation was executed as a final part of the third test case. Three Bluetooth sensors were tested. The results are displayed in Figure 2.6. Two conclusions can be drawn from this figure. The first is that there indeed is a negative relation between distance and RSSI. The strength of this relation is emphasized by the high R^2 values. The second conclusion is that there is a difference between the different brands of Bluetooth sensors. It is not surprising that the Sena UD100 is on top, since this is a Class-1 device. However, the difference between D-link and Trust is more surprising, because those are two Class-2 devices. It is therefore important to maintain consistency in the use of a certain type of sensor throughout an experiment.

The test cases show that one needs to consider many aspects when deploying Bluetooth tracking in an indoor environment. In the actual shopping mall experiment, three different types of scanners were used, optimised in function of the particular space. Aluminium foil covered multiple

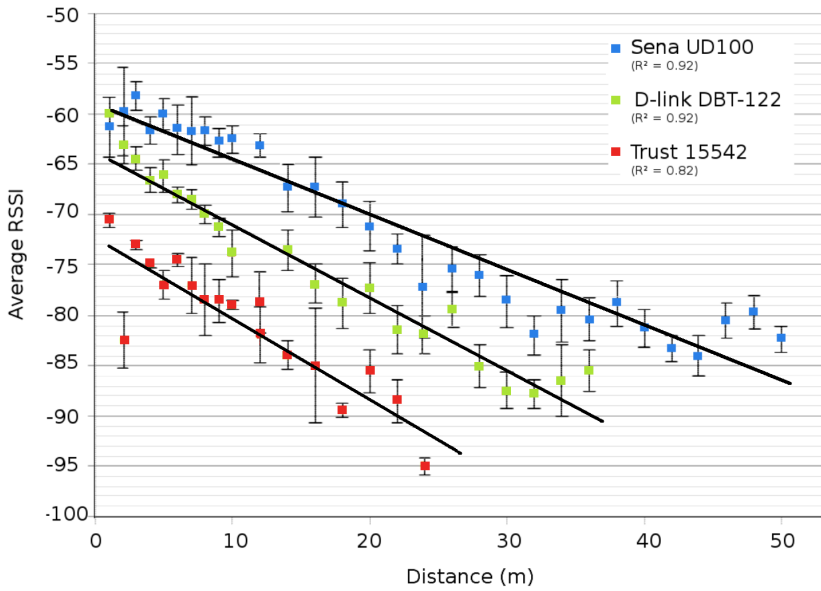


Figure 2.6: Test Case 3c: The Relation between RSSI and Distance. An inverse relation is clear. The test also revealed differences between brands.

scanners in order to limit their reach.

2.4 Results and Analysis

2.4.1 Data

A well thought-out data filtering method can significantly improve data quality. Therefore, after installing the 56 Bluetooth scanners on site, the RSSI values at the entrance of every store were determined. These provide cut-off values for the observations such that if the measured signal strength falls below the threshold, the observation is removed. Test Case 3 provided evidence for the use of this methodology. Observations that were only tracked for a very short period can be categorised as passers-by, rather than as an actual visitor of the store. Based on that information, these observations are eliminated. Dependent on the type of store, the optimal threshold for this time interval may be different. It is therefore advisable to consider every store individually. This process reduces the number of Type I errors, but inevitably also leads to a higher number of Type II errors. However, the reduction in Type I error is much more outspoken than the increase in Type II error. In the few cases where a MAC-address was detected at the same time by two scanners, the one with the highest signal strength was selected.

As could be expected, the data comprises multiple recurring visitors. Since we are mainly interested in the paths per visit, the observations were divided into trajectories of maximal 12 hours, aligned with the opening hours of the mall. Depending on the setting and the desired analysis, the optimal threshold might be different and it requires no more effort than simply changing one parameter to enable other analyses.

We used two methods to determine the detection ratio. Method 1 concerns a manual count, based on video footage of a security camera. The one and a half hour footage resulted in a head count of 1,121 visitors. 106 distinct MAC addresses were scanned at the same location. This results in a ratio of 9.46%. The second method used the counts of light curtains in three stores, during 17 full days. These counting results in a ratio of 1,721/17,486, or 9.84%. Calculating a weighted average, the final reported ratio is 9.81%. This value lies within the expected range of values found in literature (cf. 2.2.2.3).

Nineteen days of scanning customers resulted in 18.943 unique MAC addresses. When excluding the MAC addresses that were only registered in the parking or at an entrance, we get 10.719 unique devices that actually visited a store in the shopping mall. Different device types were registered.

Mobile phones and smartphones accounted for 89% of the registered devices. These devices therefore possess all necessary characteristics of a good proxy. They are linked to the individual and one can therefore be confident about a one-to-one relationship between the location of a scanned MAC address and the location of an individual. They are also by far the most popular category of tracked devices. The remaining 11% goes to audio/video devices (mainly headsets and car kits, accounting for 9%), wireless phones (1%) and computers (1%).

2.4.2 Applications

Below, we present some applications and results. First of all, to determine which stores attract the most visitors, one could simply calculate the ratio of unique MAC addresses that were scanned in the store to the total number of unique detected MAC addresses. In our case, the analysis reveals that 75% of detected visitors visits the most popular store. This result confirms our prior expectations, since it is an *anchor store*. Anchor stores are used to generate traffic in a shopping mall (Pashigian and Gould, 1998). They attract large numbers of visitors, which subsequently are likely to visit the other, smaller stores as well. The second most popular store follows at 27.5%, while the least popular store only attracts 0.5% of the visitors.

The uniqueness of a MAC address allows us to analyse multiple customer metrics. Customer loyalty can be approximated by how often customers return. This analysis can be run both on the individual store-level as on the level of the entire shopping mall. The analysis revealed that 50.3% of the public visited the shopping mall more than once in the observed period. Analysis of the data also reveals insights into the number of stores a customer visits. Figure 2.7 plots these findings. The distribution is quite similar for weekdays versus weekends, but people are slightly more likely to visit a higher number of stores during the weekends. These analyses can also be used to identify types of customers. How many people are drawn to the shopping centre because of one specific store, how many want to enjoy the full shopping mall experience and visit multiple stores, what is the role of the anchor store, et cetera. An analysis of the duration of store visits resulted in a median visit duration of 3.5 minutes, whereas the median duration of the total shopping mall visit was 24 minutes (see Figure 2.8). The result of 24 minutes is in line with our expectations, since we are dealing with a rather small shopping mall of 39 stores. Very long visits can usually be attributed to employees. This provides an additional filtering method to exclude employees from the analysis, besides a more manual filtering process of explicitly scanning their MAC addresses.

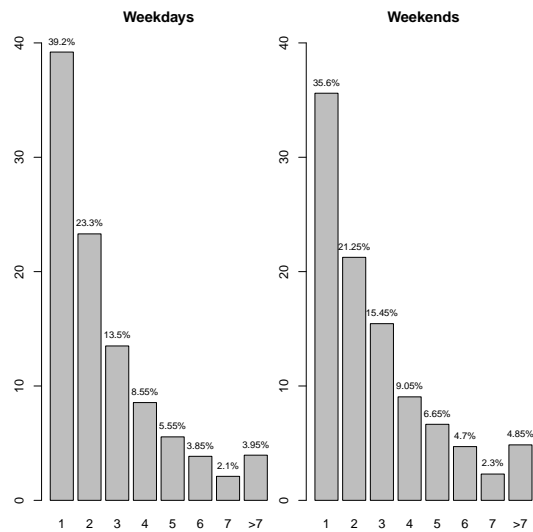


Figure 2.7: Number of Stores Visited (Relative Frequencies)

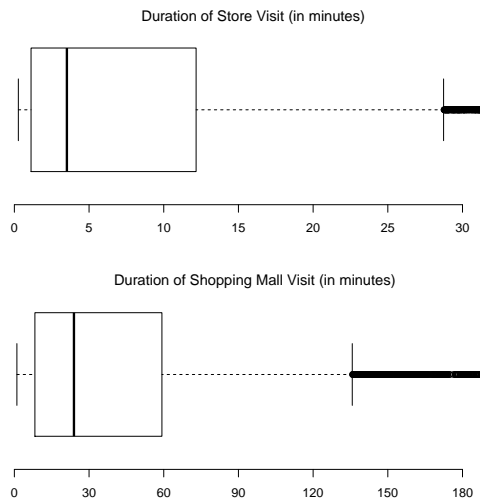


Figure 2.8: Time Measurement of Visits

Before installing the Bluetooth scanners, one has to reflect on the desired kind of information. If one of the objectives is to acquire insight in the means of transport of customers, it is crucial to place scanners in the parking facility or nearby public transport hubs. In our case, the shopping mall was mainly interested in the use of the parking facility. We calculated that during weekdays 18.6% of the customers comes by car, while this percentage grows to 31.2% during weekends. An explanation can be that during the weekend most of the customers intentionally visit the shopping centre as a planned event. These people might come from outside the city and therefore come by car. During the week visitors are probably mostly people who live or work close-by the shopping mall and therefore do not need a car to reach the shopping mall, which is located in the city centre. To confirm this hypothesis, Bluetooth scanners could be placed at multiple locations in the city and on access roads to the city.

For sequential analysis of the data an important condition arises. Since the data comes from scanners on multiple close-by locations, the timestamps have to be very precise. Therefore all internal clocks were synchronised at the beginning of the experiment.

Combining the tracking data with the map of the shopping mall makes it possible to calculate path lengths. The path length does not include the distance travelled within stores, since the spatio-temporal behaviour is recorded at store-level. The average path length amounted to 279 metres. There are several options to visualise the path data. We present two suggestions to visualise the number of customers that visit store x after having visited store y . The first visualization is a plot of the flows on the map of the shopping mall (see Figure 2.9). To maintain readability of this map, the map is restricted to the most frequent flows. We also restricted this figure to one floor of the shopping mall, since mapping the flows between all shops over the three floors would result in a cluttered 3D-plot. An alternative visualization that overcomes this latter problem is a migration plot (see Figure 2.10). This plot displays the major flows in the entire shopping mall. It is clear that store B, the anchor store (consisting out of three departments) accounts for a large part of the flows. The first plot shows more clearly that people are likely to visit a neighbouring store as the next store. This kind of observations are not immediately clear from a migration plot. Hence, different visualisations can provide rich and complimentary insights.

A last application is a cluster analysis on the path data. The goal of cluster analysis is to segment the vast amount of paths and visitors in order to identify groups. We clustered customers on the stores they visited, regardless of the specific order in which they visited the stores. This enables us to categorise paths such as A-B-C in the same cluster as path C-B-A.

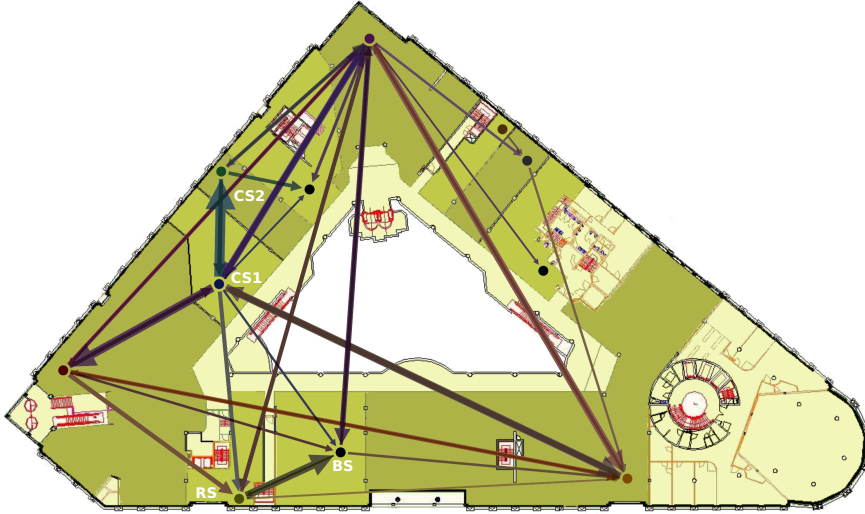


Figure 2.9: Major Flows on First Floor of Shopping Mall. *The linewidth is relative to the size of the flow. From this figure we can observe that many people that visit the record store (RS) subsequently visit the neighbouring bookstore (BS). The same holds for the two similar neighbouring clothing stores (CS1 and CS2).*

For our purpose, this latter path is more similar to A-B-C than A-B-D is for example. Depending on the desired output and insights, one needs to decide on whether or not taking the actual sequence into account. When ignoring the actual sequence, more traditional clustering methods can be used. A matrix was created in which a value of 1 was assigned to the store variable if the store was visited and a 0 in the opposite case. Based on that new dataset, we applied hierarchical clustering (Murtagh, 1985) on the entire sample of 10.719 visitors. Figure 2.11 shows a dendrogram for a random sample of 50 customers. The figure was restricted to this low sample number in order to maintain readability. The full analysis does nevertheless include the entire sample of 10.719 visitors. The clustering method allows the user to define the optimal number of clusters, depending on the desired level of detail. Graphically, that means drawing a horizontal line on the desired height in Figure 2.11. The higher the line, the less clusters will be defined. In the figure, moving the line upwards would for example imply that the two most left clusters become one larger cluster. When opting for a manageable number of five clusters, the omni-presence of the anchor store became clear again. Cluster 2, 3 and 4 each represent customers that visit almost exclusively this store. Cluster 1 contains customers that come to

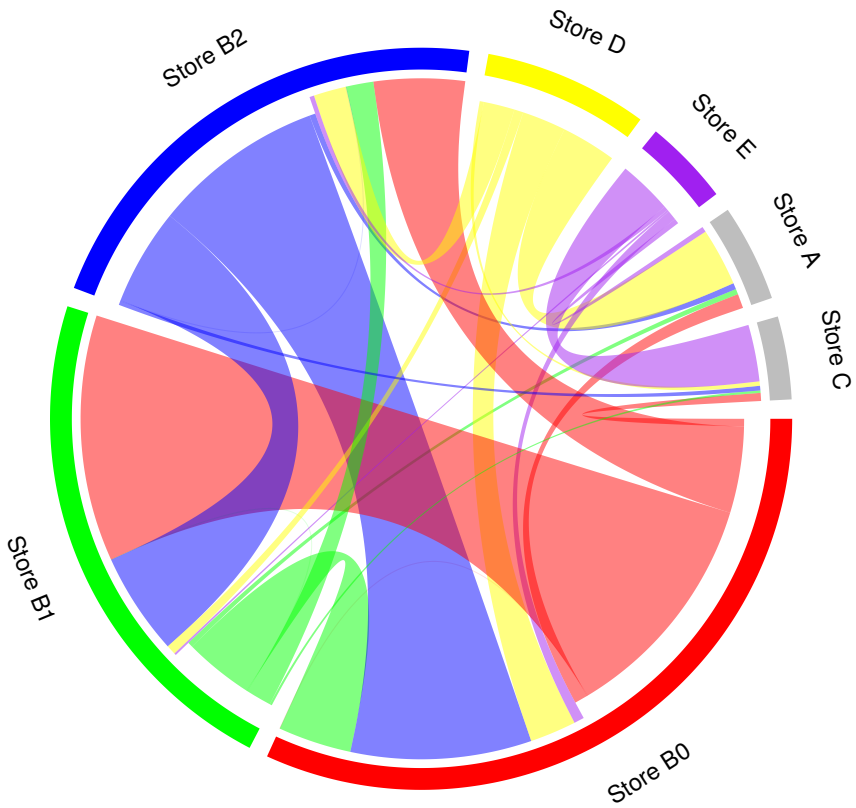


Figure 2.10: Migration Plot of Major Flows. The circle segments represent the different stores as origins and destinations of the flows. The width of the links represent the size of the flows. The colours/greyscale indicate the direction of the flows, links have the same colour as the origin. There is also a gap between the start of the link and the segment to indicate the origin.

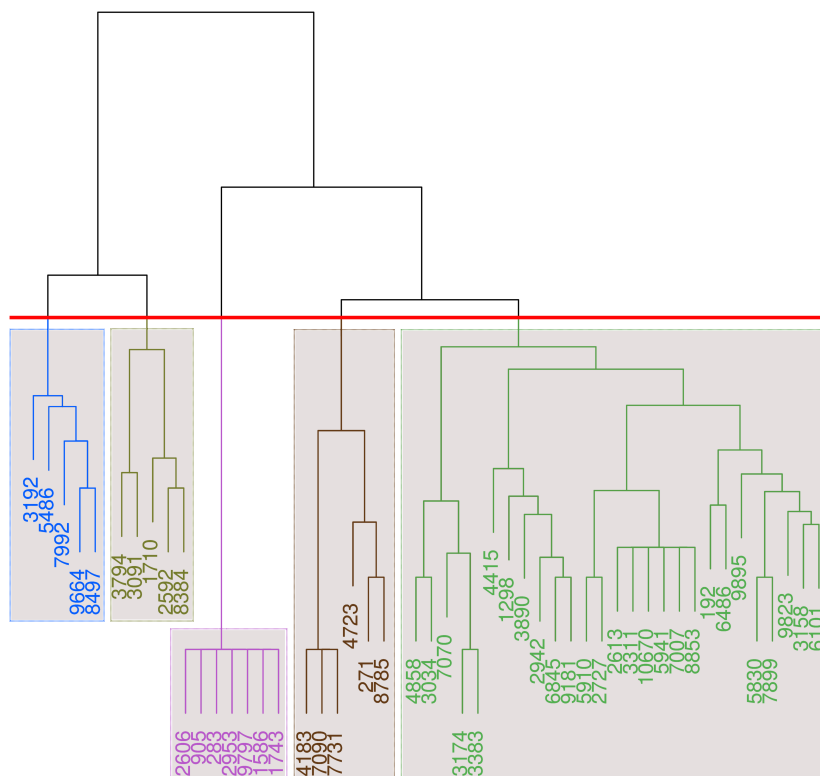


Figure 2.11: Cluster Dendrogram of Hierarchical Clustering on 50 Random Customers. The example shows five extracted clusters.

the shopping mall mainly for eating and multi-media stores, but not for the clothing stores. Customers who visit a variety of stores are to be found in cluster 5.

2.5 Discussion and Conclusion

The main goal in this research was to investigate the applicability of Bluetooth tracking in an indoor environment. Many factors make it far from evident to obtain sequential path data in indoor settings. This research provides a real-life example of the application of Bluetooth tracking in an indoor shopping mall. We have shown that the method is able to analyse the spatio-temporal behaviour of individuals. The method is effective and comes at a rather low cost when compared to the competing technologies such as RFID.

Three test cases were executed prior to the actual experiment. These cases revealed that it is crucial to invest a fair amount of time and effort in the tuning of the scanning devices and their position. Different types of scanners and a wrapping of aluminium foil were used to optimize the accuracy. Additional filtering of the data forms a further step to improve data quality.

We demonstrated the relevance of knowledge and advances in the field of geography for other domains such as marketing. A shopping mall was used as research setting for this marketing application. In a broader sense, Bluetooth tracking is relevant in many settings that we defined as ‘uncontrollable’ (see Section 2.2.2.3). Also, Bluetooth is used in indoor positioning and navigation.

This research is to our knowledge not only the first to deploy Bluetooth tracking in a real-life indoor retail setting, but also the first that uses Bluetooth tracking as an alternative for gathering marketing insights. The collected data not only reveals paths as such, but can also be used to discover many other relevant statistics. Furthermore, this information is of higher quality than most of the traditional surveys. A simple example is the duration of a visit. Bluetooth tracking data provides the actual values, whereas this is not feasible for people. In addition, the non-participatory approach (cf. 2.2.2.3) results in natural behaviour of the tracked individuals. Of course, certain more in-depth information cannot be obtained with this method and if this is desired, a traditional survey can provide this. Furthermore, Bluetooth tracking definitely lowers the cost of information acquisition. The detection ratio of 9.81% is high enough to quickly generate a sufficiently large sample. In contrast to traditional survey methods, where the cost is proportional to the sample size, the cost stays constant in case of

Bluetooth tracking. The method therefore is a sound long-term investment.

Considering the legal importance of privacy in marketing research, the non-participatory aspect should be approached with great care, since tracking human beings without their consent likely generates legal problems. A considerable advantage of Bluetooth tracking therefore lies in the use of MAC addresses. These cannot be linked to the true identity of the person (Giannotti and Pedreschi, 2008; Abedi et al., 2014). A problem with GPS tracking for example is the fact that it can easily link the person's identity through the used smartphone application (Giannotti and Pedreschi, 2008; Abedi et al., 2014).

We can conclude that Bluetooth tracking is a sound methodology to gather information about the unbiased spatio-temporal behaviour of individuals. Furthermore, shopping mall management did consider the experiment interesting and successful. The current line of thinking is to use Bluetooth tracking next to the already existing bi-yearly customer survey.

2.6 Limitations and Future Research

The detection ratio is a significant limitation of the methodology. Devices must be in on-visible mode in order to be scanned. In our experiment, the detection ratio amounted to 9.81%, which is similar to ratios found in previous research (O'Neill et al., 2006; Hagemann and Weinzerl, 2008; Versichele et al., 2012a,b, 2014b). For many purposes this will lead to a sufficient sample size. Furthermore, Bluetooth tracking might introduce a sample bias. Further research is needed to investigate whether certain segments of the population are more likely to carry discoverable Bluetooth devices. The evolution of the detection ratio over time is another aspect that deserves attention. At the moment, it is not clear whether the technology usage will increase or decrease.

Another drawback of the methodology is its currently rather unautomated set-up. Covering scanners with aluminium wrapping is an inconvenient approach to limit the reach of the scanners, in order to increase accuracy. Further research should investigate whether other, more automated, procedures could attain the same effect. It might become possible to automatically program the scanners so that they only scan a specific region. Until now, a solution lies in putting a strict threshold on the RSSI values that the scanner should allow. This software approach does however not necessarily lead to better results than the hardware approach. The non-centralised data collection is another aspect that could be improved upon. As suggested by Versichele et al. (2014a), connecting the scanners via a (wireless) network might significantly improve the practical usability.

Clearly, there are downsides characterizing the Bluetooth tracking method. However, no currently existing tracking method is flawless, nor are the more traditional methods used in marketing and other research. This research shows that, depending on the context, the desired information and level of detail, Bluetooth tracking proves to be a viable approach.

2.7 Acknowledgments

We would like to thank Edward De Mûelenaere and the team of prof. dr. Nico Van de Weghe for the technical set-up and the data collection. We would also like to thank the people from the shopping mall for allowing us to conduct this research and their cooperation in this project.

3

Home Location Prediction with Telecom Data: Benchmarking Heuristics with a Predictive Modelling Approach¹

Abstract

Correctly identifying the home location is crucial for human mobility analysis with telecom data, more specifically call detail record (CDR) data. To that end, multiple heuristics have been developed in literature. Nevertheless, due to the lack of ground truth home location data, no study has thoroughly validated these widely used methods so far. We present a detailed performance analysis of existing home detection heuristics, using a unique dataset that enables this important validation on the lowest level, being the level of the cell tower. Our research indicates that simple heuristics surprisingly outperform their more complex counterparts. The benchmark study revealed that the best heuristic is able to identify the home location with an average error of approximately 4.5 kilometres and selects the correct home

¹Based on: Oosterlinck, D., Baecke, P., & Benoit, D.F., (2021). Home location prediction with telecom data: Benchmarking heuristics with a predictive modelling approach. *Expert Systems with Applications* 170, 114507.

tower in 60.69% of the cases. Based on the insights provided by our study, we propose a new heuristic that increases the accuracy to 61% and lowers the average distance error to 4.365 kilometres. Secondly, if the home location is known for possibly only a fraction of the instances, we propose a labelled predictive modelling approach. Adding social network based variables in this predictive model further enhances the predictive performance. Our best model reduces the average distance error to 2.848 kilometres and selects the correct home location in 72.08% of the cases. Furthermore, this result provides an indication of the upper bound for home detection with CDR data. Finally, models that only make use of social network based data are developed as well. Results show that even without using data of the focal individual, these models are able to select the correct home tower in 37.65% of the cases and achieve an average distance error of 8.1 kilometres.

3.1 Introduction

The number of studies about human mobility displayed a steep increase around 2008 and is still growing at a high pace (Barbosa et al., 2018). The interesting field of human mobility includes a large variety of applications and is therefore able to have a large impact on everyday life. Human mobility analysis sparks the development of smart cities, enables socio-economic studies, facilitates the understanding of mobility patterns and boosts better data-driven decision making amongst others. Vanhoof et al. (2018c) states that human mobility analysis will render important insights in the wider structures governing our society. The study of human mobility data is remarkably important for epidemiological studies in order to model the spread of viruses and even assess the impact of measures taken during the Covid-19 crisis. More applications can be found in the research about commute behaviour (Kung et al., 2014), commute distances, the impact of mobility on our carbon footprint (Isaacman et al., 2011) and even traffic prediction (Lv et al., 2014). Insights from human mobility analyses can further optimize telecom and transportation infrastructure.

Research has shown that telecom data obtained from mobile phone networks, call detail record (CDR) data, has great value for these analyses. Furthermore, Gonzalez et al. (2008) and Song et al. (2010) prove that human mobility is strongly predictable when using CDR data, as people spent most of their time in a limited number of locations. However, due to matters of confidentiality, CDR data typically lacks contextual information (e.g. the content of the messages or calls), which makes it not obvious to interpret the location traces in the raw data (Liu et al., 2013). It is therefore crucial to investigate methods that annotate the raw data into meaningful locations.

The home location is one of the most important meaningful locations as the analysis of human mobility typically requires identifying the home location as a first step. This makes that home location prediction is very often a (first) part of more complex studies (Bojic et al., 2015). An accurate identification of the home location is therefore essential for this area of research. However, despite its large impact on the mobility analysis, literature failed to devote significant attention to this critical aspect. The main reason for this absence of attention has been the lack of proper validation, due to unavailability of ground truth data on an individual level. We fill this gap in literature by thoroughly evaluating the existing home detection methods using a unique data set that contains the closest cell phone tower to the actual home location. Throughout this study, the term *home tower* will be used to refer to this location. We extract the different categories of methods in literature and execute a benchmark study using 5 times 2-fold cross-validation in order to provide robust results.

The advantage of these heuristics is that they can be used even when no single home location is known. However, if the home tower is known for a part of the data set, we propose to use a labelled predictive modelling approach. Our results show that a labelled approach is able to significantly enhance the results. The performance of such a model also indicates to a large extent what the maximal attainable performance of home detection algorithms with CDR data is.

Previous research has already shown the value of social network data in multiple settings (e.g. Verbeke et al. (2014); Benoit and Van den Poel (2012)). In the context of home detection, research has been done in an online social network (OSN) (Backstrom et al., 2010). We will investigate the added value of using social network data in a CDR context. Furthermore, we build stand-alone social models that provide interesting insights for academics and telecom providers. The results show that using only the data of individuals in the social network of the focal individual, has predictive power for the home location of the latter, which also opens up possibilities for further research as it can be expected that this is valid for other than home locations as well. Telecom providers might gain intelligence about non-customers by using this knowledge as well. The data they have about their existing customers is very valuable to them, however, gathering data about non-customers is much harder, but possibly even more valuable as this helps them to gain insight in competing telecom operators and their customers. This information might strongly help them to attract these customers and therefore with enlarging their market share.

3.2 Literature Review

3.2.1 CDR Data for Human Mobility and the Need for Home Detection

Research in the field of human mobility, urban planning, transportation engineering and mobility patterns was traditionally based on travel surveys, road side surveys and travel diaries (Calabrese et al., 2011; von Mörner, 2017; Wang et al., 2018). These survey methods have major shortcomings such as small sample rates, short survey durations, under-reporting and a high cost (Calabrese et al., 2011). Meanwhile, a variety of other data sources has been used, such as circulating bank notes (Brockmann et al., 2006), Foursquare check-in data (Noulas et al., 2012), tweets (Hawelka et al., 2014; Mahmud et al., 2014; Hironaka et al., 2016) and GPS data (Vazquez-Prokopec et al., 2013; Tang et al., 2015).

However, Barbosa et al. (2018) report in their review paper that call detail record (CDR) data is the most important, game-changing data of the last decade for analysing human mobility. CDR data is the information that telecom providers capture, every time that a customer makes or receives a call / SMS. Every record in a CDR data set contains interactional (a caller and receiver id), temporal (timestamp and duration) and, importantly, location aspects. The location refers to the geographical coordinates of the cell tower that is used and is therefore always an approximation of the actual location of the user. It was shown already by Gonzalez et al. (2008) and by Song et al. (2010) that human mobility is highly predictable when studied using CDR data. People will show different usage patterns on different key locations. From these it becomes possible to derive meaningful insights (Karikoski and Soikkeli, 2013; Blondel et al., 2015). An extensive review of the research on CDR data is published by Blondel et al. (2015).

CDR data is ideally suited for both large scale location analyses, as well as for research on individual level mobility. Mobile phones have a worldwide penetration rate of 96% (Iqbal et al., 2014). This makes that CDR data overcomes the low sample problem of the survey approach by capturing almost the entire population and offers a consistent approach for research on mobility patterns throughout the world (Kung et al., 2014). Using CDR data also implies tracking all mobile phones in a provider's network, not only the users that installed a certain application on their device (Scherrer et al., 2018), which would be the case if GPS smartphone data is used. CDR data does also not require additional sensor data and map information, which substantially lowers the cost of data collection as well as being rather easily transferable to other regions (Liu et al., 2013). The

problems of self-reported behaviour in surveys (Eagle et al., 2009a) and their traditional high cost are also mitigated (Isaacman et al., 2011).

Of course, when using CDR data for location based applications, one needs to keep in mind that this data source was not originally designed for this objective. The original function of the data was to count the usage per customer for billing purposes. This makes that the observations are recorded only when somebody makes use of the network by calling or texting. The data generation is thus non-continuous. The precision level of the location also depends on the distribution of the cell towers. These aspects might lead to a low temporal and spatial granularity. One also needs to take into account the different market share of providers and the difference in calling plans between customers. However, most of these effects are largely mitigated due to the large geographic coverage and the high penetration of mobile phones (Wang et al., 2018).

Research has established the value of CDR data for human mobility analysis. Multiple authors (Kung et al., 2014; Vanhoof et al., 2018b; Bojic et al., 2015; Dash et al., 2014) indicate that the correct identification of the home location is an important prerequisite for the large majority of applications in human mobility, such as home-work commuting, commuting patterns, mobility profiles, mobility and epidemiological models (Tizzoni et al., 2014). Isaacman et al. (2011) point out that people spend most of their time on a limited number of locations. It is crucial to identify these key locations, such as the home location, in order to understand human mobility, social patterns and implement technology and policy decisions like the deployment of telecommunications and transportation infrastructure. Nevertheless, the methods of home detection are often obscured in literature (Vanhoof et al., 2018b). Despite that this initial step largely determines the quality of the subsequent analyses, the performance is hardly ever validated. The absence of a ground truth is in many cases the main reason why this validation can not be done. Vanhoof et al. (2018b) underlines that research with individual level ground truth needs to be done in order to assess the quality of the different home detection methods.

3.2.2 Categories of Home Detection Algorithms

The home location in CDR data usually refers to the coordinates of the *home tower*, the cell tower that is closest to the actual home location. As the level of precision is restricted to the level of the cell towers, this is also the case for the home locations and thus *home towers*.

Literature commonly distinguishes two broad classes of home detection methods, as we report in Figure 3.1. *Single-step* methods apply home

detection rules directly on the individual towers. The second category, the *two-step* methods, add an extra initial step by first clustering towers. This is done in order to counter the fact that cell phones might switch towers despite remaining at the same location. In the second step, these methods apply home detection rules similar to the single-step methods, to these clusters. Hence, both approaches need decision rules in order to identify important places and label the correct location as home (Vanhoof et al., 2018b). Based on the literature, we further divide the single-step methods into two classes, the activity and inactivity heuristics. The former takes the standard approach by using data when somebody is actively using his or her phone, whereas the latter models the periods in between usage. This results into three classes, presented in Figure 3.1.

3.2.2.1 Decision Rules and Heuristics Based on Activity (Activity Heuristics)

This first category of home detection algorithms contains single-step methods. Decision rules are investigated to identify the home tower. Vanhoof et al. (2018b) examine three decision rules that fit in this category. A first heuristic, also used by many other authors (e.g. Tizzoni et al. (2014); Song et al. (2010)), states that the home tower is the cell tower where the majority of both outgoing and incoming calls and texts are observed, also called the amount of activities criterion, we will refer to this method as Act_1.

A second decision rule aims to improve upon the former heuristic by taking into account the regularity of a certain location. An extraordinary event might lead to a lot of exceptional calls made or received by someone on a certain location. Applying the Act_1 method to this case, would result into selecting this exceptional location as most used and the location will therefore be labelled as the home location. The aim of the second criterion is to counter this undesirable effect. Regularity will be modelled with the amount of distinct days criterion. If somebody is regularly on a certain place, the chances are higher that this is the home location. The Act_2 method counts the amount of distinct days on which a location was used. The tower with the maximum number of distinct days with phone activities is then selected as home location.

Another attempt at further improving the amounts of activities criterion (Act_1) is done by including time constraints. These time constraints heuristics add the restriction that only activities between specific hours should be taken into account. The aim is to select a time frame during which people are most likely to be at home. Vanhoof et al. (2018b) select the time frame from 7PM till 9AM (Act_3). By selecting such time

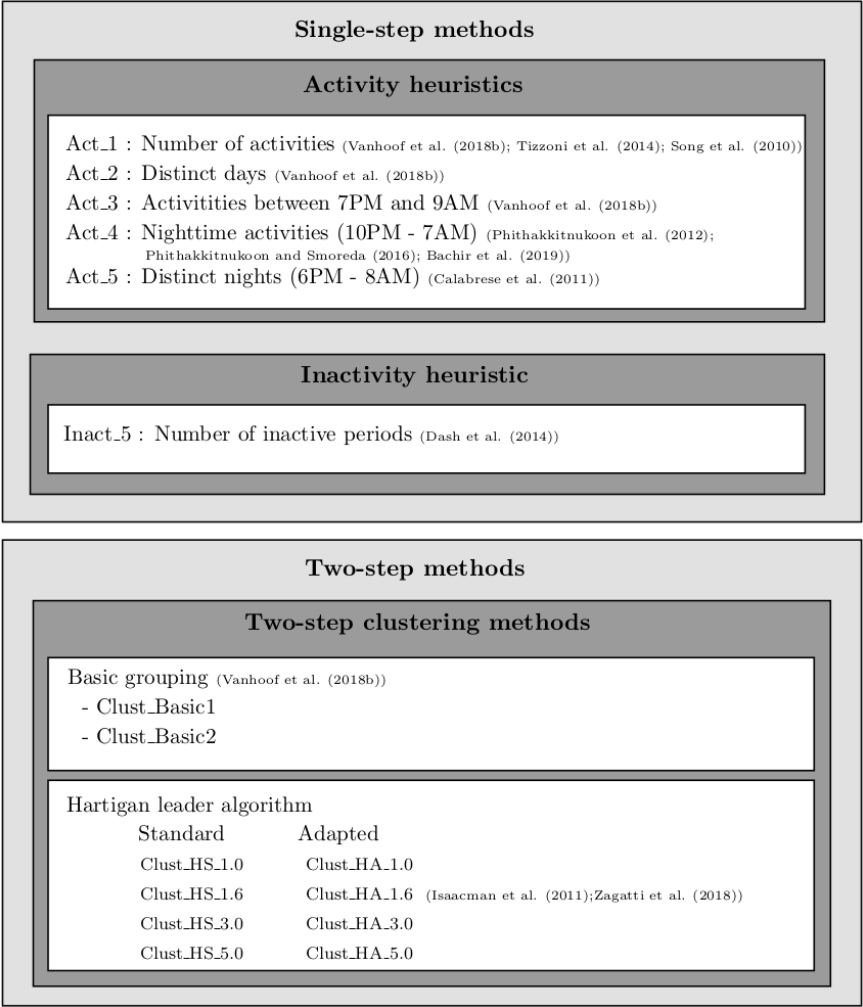


Figure 3.1: Overview home detection methods. Three main categories of home detection methods were identified in literature; activity heuristics, inactivity heuristics and two-step clustering methods. The different methods that will be evaluated in our benchmark study can be found in this table.

frame, the observations during working hours are excluded. The underlying assumption is that people are less likely to be at home during working hours and more likely to be at home during non-working hours. Other authors (Phithakkitnukoon et al., 2012; Phithakkitnukoon and Smoreda, 2016; Bachir et al., 2019) use this same criterion, but with a different parameter choice. They use a more stringent range and try to model night time, for which the time frame of 10PM until 7AM is selected (Act_4).

A final activity based heuristic combines the ideas in the above methods. Calabrese et al. (2011) apply both the concept of regularity and the refinement in terms of time frame. This results into the number of distinct nights criterion. The authors select the time frame from 6PM until 8AM as night time. The number of nights with activity on each location is counted and the location with the maximum number of distinct nights is selected as home location (Act_5).

It is clear that for the time restricted methods, the parameter choice will influence the performance. However, since proper validation with individual level ground truth lacks in literature, it has been difficult to select the optimal parameters. We will therefore benchmark these methods on our unique dataset.

3.2.2.2 Decision Rules and Heuristics Based on Inactivity (Inactivity Heuristic)

Instead of taking into account the activity of a user on a certain location, this second category counts the number of times that somebody is inactive on that location (Dash et al., 2014). Given that one usually sleeps at his/her home location, the idea behind the inactivity heuristic is to set a threshold for a period without activities in order to model the sleeping hours. In practice, this means that a location is counted if the time between an activity on that location and the next activity (on any location) is longer than the selected threshold. In other words, the periods with no data are used and the last location preceding this empty period is registered. The advantage is that this also works for shift workers, as inactivity does not need to be observed during specific hours during night time. The location with the highest count of inactivity is selected as the home location. The threshold was set to five hours by Dash et al. (2014).

3.2.2.3 Two-Step Clustering Approaches

The two-step, clustering, approaches deviate from the single-step approaches by first clustering certain towers together. Because of physical boundaries

and other properties of the cell towers and the environment, it is possible that the cell phone switches connection between different towers, although the user stays at the same location. In order to mitigate this effect, towers are clustered together based on their location. This first step is executed for every individual, thereby resulting into different clusters of towers for every individual. The second step is to label the clusters by scoring them according to criteria, similar to the single-step methods.

The idea of clustering certain towers is closely related to the fundamental concept of stemming in the text clustering literature (Porter et al., 1980; Bharti and Singh, 2015; Abualigah et al., 2018a,b; Abualigah, 2019). The application of clustering as a first, pre-processing, step in the two-step clustering approaches, aims to avoid biases by clustering similar elements together. In text clustering the elements are words, in our application the elements refer to cell towers. Stemming transforms inflectional forms of certain words to the same root or *stem* by removing the prefixes and suffixes of each word (Abualigah et al., 2018b). For example, the words consult, consultant, consulting and consultative will result in the same stem 'consult'. It is evident that for many applications the difference between the individual words is irrelevant and that the performance of the resulting model will significantly improve due to stemming. The clustering of cell towers aims to achieve a similar improvement, by clustering certain towers based on geographical proximity, thereby reducing detrimental differences in location.

Different options for the first, clustering, step are examined in literature. Vanhoof et al. (2018b) present a first example of clustering cell towers. This basic approach clusters all activities for a selected individual that are recorded within a spatial perimeter of 1 kilometre around the cell tower. Thus, if we look at one cell tower, all activities of the user on other towers within the 1 kilometre perimeter are added to the first tower. This approach differs from the other clustering approaches by not scoring the clusters as a whole, but scoring every tower by including the records of the towers within their cluster. This results into scores for every tower, as opposed to scores only for every cluster. The scoring for the basic clustering by Vanhoof et al. (2018b) is selected from their activity heuristics. Vanhoof et al. (2018b) apply their first activity heuristic (Act_1) resulting into the first option; Clust_Basic1, where the home tower will be selected as the tower with the highest number of activities on this tower, including the activity within the perimeter. They also use their third activity heuristic (Act_3) resulting into Clust_Basic2, which is identical to Clust_Basic1, except that only activities between 7PM and 9AM are taken into account.

Other clustering approaches are presented by Isaacman et al. (2011) and

Zagatti et al. (2018). In the first step of this two-step procedure, cell towers are clustered based on the Hartigan leader algorithm (Hartigan, 1975). One of the advantages of this clustering algorithm is that it does not require an a priori chosen number of clusters. We will discuss both the standard algorithm and an adapted version which was used by Isaacman et al. (2011) and Zagatti et al. (2018).

The standard Hartigan leader algorithm starts by ranking the observations (in this case, the cell towers) based on a selected feature (in this case the number of distinct days). The algorithm assigns the first tower in the list as the cluster centre for a first cluster. The algorithm then descends through the sorted list of towers and checks whether the next tower falls within a chosen threshold distance from an existing cluster centre. If this is the case, the tower is added to the existing cluster, if not, a new cluster is formed with this tower as centre.

Isaacman et al. (2011) and Zagatti et al. (2018) implemented the algorithm with the adaptation that the cluster centres can move. If a new tower is added to an existing cluster, the cluster centre is recalculated, weighted by the distinct days that a tower is used. This process can be observed in Figure 3.2.

In the second step of this two-step procedure, the clusters need to be scored. Zagatti et al. (2018) propose to score the clusters by summing +1 for every activity in the evening (7PM - 7 AM) and during the weekend and -1 for daytime hours (8AM-5PM) and weekdays (Scoring a in Figure 3.1). Furthermore, the algorithms can be executed with different thresholds for the spatial perimeter. Isaacman et al. (2011) and Zagatti et al. (2018) use 1.6 kilometres, but discuss that other options could be suitable as well. This study will investigate threshold values ranging from 1 to 5 kilometres. The naming convention for the different implementations of this category in Figure 3.1 is as follows. HS refers to the standard Hartigan leader algorithm, while HA refers to the adapted version. The number refers to the threshold (in kilometres) that is set in the algorithm.

3.2.3 Validation

As mentioned earlier, one of the main shortcomings in literature is the lack of proper validation. Validation is either not done (e.g. Zagatti et al. (2018)) or limited to validation on an aggregated level, usually through census data (e.g. Phithakkitnukoon et al. (2012); Kung et al. (2014); Ahas et al. (2010b); Calabrese et al. (2011)). This means that the sum of estimated home locations is compared to the population count at that location. Common validation metrics on the aggregated level measure the correlation

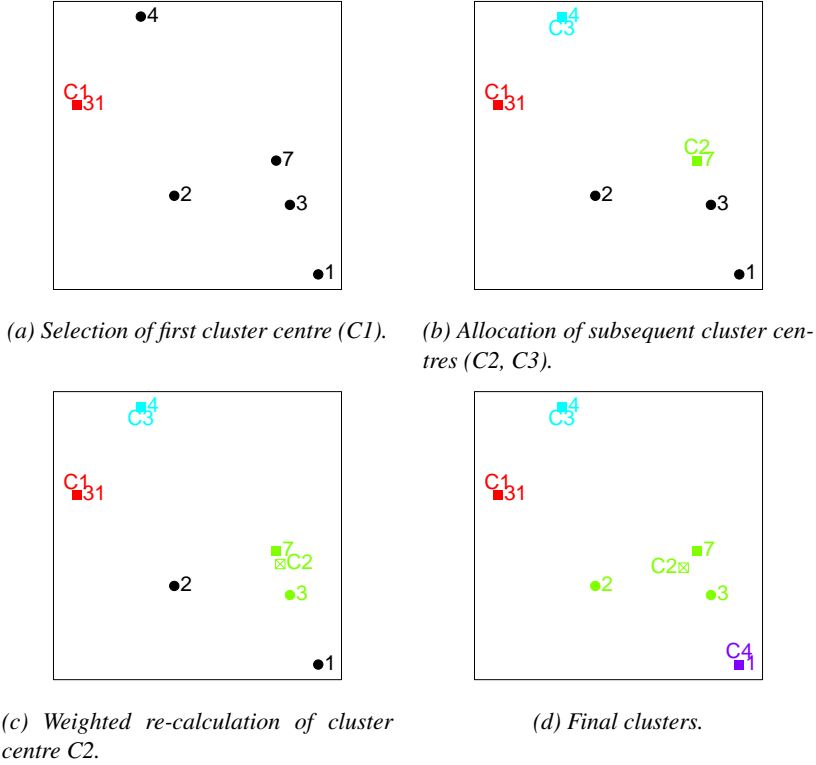


Figure 3.2: Example of adapted Hartigan leader algorithm. The algorithm starts with the calculation of the number of distinct days (cf Act_2) for every tower, represented by the numbers in the figure. The tower with the highest number is selected as the first cluster centre, indicated with a square (a). The algorithm then selects the tower with the second highest number of distinct days and so on. If none of the distances to an existing cluster centre satisfies the threshold value, the tower is selected as a new cluster centre, in this example resulting into the clusters with centre C2 and C3 (b). The distance from the fourth tower in the ranked list (with 3 distinct days) to cluster centre C2 satisfies the threshold. This is the first tower in this procedure that can be added to an existing cluster. The cluster centre needs to be re-calculated due to the addition of the new tower. The new centre is weighted by the number of distinct days, explaining why C2 now lies closer to the original cluster centre; the tower with 7 distinct days (c). The square with an x inside indicates that the cluster centre is the result of the weighted re-calculation. The algorithm continues by adding the observation with 2 distinct days to the second cluster, which leads to a new cluster centre again. The last tower does not satisfy the threshold to any existing cluster centre and results into a final single-observation cluster C4 (d).

between the count of estimated home locations and the census data. Examples are the cosine similarity metric (Vanhoof et al., 2018b), Pearson's r (Vanhoof et al., 2018a) or a simple matching coefficient (Bojic et al., 2015). Two studies contain a small sample ground truth and report distance between the estimated and actual location (Isaacman et al., 2011; Dash et al., 2014).

The problem is that the level of granularity of the census data (e.g. city or village level) is very unlikely to align with the granularity level of the CDR data (the distribution of the cell towers). An additional issue is that this type of validation gives no guarantee whether the algorithm actually detects the correct individual home location as illustrated by the following simplified example. Census data reports that both city A and city B count 1000 inhabitants. Now, consider an algorithm that would actually locate the entire population of city A in city B and vice versa. It is clear that the accuracy should be zero, however as validation is only done on the aggregated, city, level, the reported accuracy will be 100%. Third, using CDR data is generally limited to one provider. The spatial distribution of the market share of this one provider is typically unknown, which also hinders this type of validation. The fact that there is no consensus in literature on which home detection methods are best, is to a large extent due to these validation problems.

It is clear that there is a high need for a thorough validation on the level of the individual, as already suggested by Vanhoof et al. (2018b). The anonymised CDR data in this study contains the tower closest to the home location of the individual users and therefore enables the crucial validation.

3.2.4 Social Network

In their review paper about human mobility research, Barbosa et al. (2018) discuss that individuals in a social network, such as friends, family or colleagues are likely to share locations and mobility patterns (Axhausen, 2005; Carrasco and Miller, 2006; Dugundji and Walker, 2005). In the context of online social networks, Liben-Nowell et al. (2005) were among the first to show a relation between distance and online friendships. Backstrom et al. (2010) find that the location of close contacts may predict the location of the focal individual in the online social network (OSN). Bojic et al. (2015) confirm that it is possible to identify meaningful places such as home location, for users that do not reveal any location information themselves, by using merely the location data of their friends on the OSN. Backstrom et al. (2010) report that the home location can be estimated within 40 kilometres of their actual home location for almost 70% of the US-based Facebook

users that have more than 15 friends.

Multiple researchers confirm these findings in the context of mobile phone users (Wang et al., 2011; Lambiotte et al., 2008; Krings et al., 2009; Phithakkitnukoon et al., 2012). Furthermore, Phithakkitnukoon et al. (2012) discuss that a mobile phone social network derived from CDR data is a better representation of actual everyday personal networks than online social networks. Both the insight that the location of other individuals in the social network has predictive power for the home location of the focal individual and the insight that CDR based social networks outperform online social networks, are a clear indication that augmenting the home estimation models with social network data might add to the performance of the models.

The results of Backstrom et al. (2010) in an online social media context, can even be enhanced by adding information about how strongly individuals are connected (Chen et al., 2014a). Phithakkitnukoon et al. (2012) showed that around 80% of visited locations are within 20 kilometres of peoples nearest social ties locations. Increasing this geo-social radius to 45 kilometres makes the figure rise to 90%. Hence, it is valuable to take the concept of *tie strength* into account to further enhance the predictive performance in the home location prediction models as well. We will define tie strength based on previous research (Onnela et al., 2007; Nitzan and Libai, 2011; Roelens et al., 2016; Meyners et al., 2017). The tie strength between individual i and individual j is defined as the ratio of the communication volume between individual i and individual j and the total communication volume of the individual i . As suggested by Nitzan and Libai (2011) and later used by Roelens et al. (2016), the weight of a text message is set equivalent to a one minute call.

$$Comm_{ij} = sms_in_{ij} + sms_out_{ij} + minutes_calls_in_{ij} + minutes_calls_out_{ij} \quad (3.1)$$

$$Comm_i = \sum_j (sms_in_{ij} + sms_out_{ij} + minutes_calls_in_{ij} + minutes_calls_out_{ij}) \quad (3.2)$$

$$Tiestrength_{ij} = \frac{Comm_{ij}}{Comm_i} \quad (3.3)$$

3.3 Methodology

The methodology section introduces the four important methodological aspects of this research. First, we introduce the dataset that will be used for the main parts of the analysis. Secondly, we explain our validation procedure, crucial to this research. Thirdly, we shortly discuss how we benchmark the heuristic methods encountered in literature to our dataset, including some value adding adaptations. In a fourth methodological section, we introduce our predictive modelling approach for the problem of home detection.

3.3.1 Data

The CDR data used in this study contains five weeks of anonymised CDR data including voice calls and SMS. For an effective evaluation of home detection methods, it is advised to avoid holiday periods as too many people display a temporary change in behaviour and might even be away from home for several weeks (Blondel et al., 2015). Data of the provider in the period from Monday 2 May 2016 until Sunday 5 June 2016 was selected for this purpose.

The ground truth home location in our data set is based on the billing address of the customer. As discussed in Section 3.2.2, this billing address is substituted by the cell tower that is closest to this location and will be referred to as *home tower*. One needs to be aware that people may divide their time between more than one home location. People might have a secondary home in which they spend time during weekends for example. However, in line with previous research, we aim at modelling the main home location. Although our ground truth can not provide conclusive evidence that we are in fact dealing with the main home location for every single individual, we can assume that this is the case for the vast majority of the data set. As people generally receive their bills on their main home address, using the billing address for this purpose can be considered as a suitable choice for defining the ground truth home location.

Several researchers remove less active customers, by requiring a minimum threshold number of used towers (Barbosa et al., 2018), a minimum number of calls (Kung et al., 2014; Dash et al., 2014; Bojic et al., 2015) or a minimum number of days with observations (Ahas et al., 2010b). This filtering however limits the scope and artificially boosts the performance as only cases with a lot of data points are retained. It is of course much easier for the algorithm to detect the correct home location for these more informative cases. In other words, a too strict filtering artificially improves the

performance. We opted for a very light restriction of minimum 5 observations in the five week period, so that we at least exclude idle numbers. Our intention is to compare the different methods on an equal, fair basis and to make sure that the results have a broad scope, including less active users. We retain 93.57% of the users by imposing this mild threshold.

A random sample of 100,000 customers satisfying the threshold is selected. This results into 54,567,294 records or 15.59 observations per day for the average customer. 100,000 customers results into 100,000 home towers in the data set. Customers also use non-home towers, this leads to 2,159,444 tower - customer id combinations. The average customer thus used 22.59 towers, in other words, on average, 4.43% of the used towers are home towers.

3.3.2 Validation Metrics

Whereas literature assessed the home detection methods merely on an aggregated level, we are able to assess the performance on individual level. We will calculate two important measures at this level: accuracy and distance error.

Accuracy is calculated as the percentage of predicted home locations (home towers) that are actual home towers. It is important to keep in mind that a random model would achieve an accuracy of 4.43% as this is the percentage of home towers in the data.

As the accuracy alone does not capture the entire picture, the average distance between the predicted and the actual home tower will be used as a second validation metric. Consider an algorithm that predicts a location as home, that is 100 kilometres separated from the actual home tower. It is logical that this case will be counted as incorrect in the accuracy measure. However, if the same algorithm predicts a location only 1 kilometre separated from the home tower as home, it will also be counted as incorrect, although the algorithm performs very well in this case. The accuracy measure might therefore be a serious underestimation of the actual performance, as several towers might be an acceptable solution. The distance measure takes this into account.

Furthermore, for robustness of the results, we use five times two-fold cross-validation (5x2cv) (Alpaydin, 1999). In a first step, the 5x2cv method randomly splits the data into two folds. Each fold is used once as a training set and once as test set. This procedure is repeated five times. As opposed to predictive modelling method that we present in Section 3.3.4, the benchmark heuristics in the following section do not require a training phase. Performance of these methods can therefore immediately be assessed on

the test sets resulting from the 5x2cv. For both approaches, this leads to ten test values for every metric. The 5x2cv method also encompasses the 5x2cv F-test to assess the significance of the difference in performance between the models (Alpaydin, 1999).

3.3.3 Benchmarks

We deploy the methods described in literature (Section 3.2.2) to our CDR data set described in Section 3.3.1. For the two-step clustering algorithm, next to the proposed 1.6 kilometres of Isaacman et al. (2011) and Zagatti et al. (2018), we also investigate a threshold of 1, 3 and 5 kilometres. Furthermore, whereas single-step methods identify one tower as home location, the two-step methods can result into a new cluster centre that lies in between towers. For validation purposes, the identified cluster centre needs to be brought back to the closest tower to this centre. This makes that the results are evaluated on the correct level of granularity and that they are compared on a fair basis between the different categories of home detection algorithms.

3.3.4 Predictive Modelling Approach

The presence of a home location in our unique dataset not only allows for a thorough validation of existing measures, but also allows for a labelled predictive modelling approach. We can use the labelled data to train models that learn how to identify the home location. To our knowledge, the only research that adopted a similar approach was done by Liu et al. (2013). Their dataset was however restricted to only 80 users with labelled data. Furthermore, their study was not aimed specifically at home detection.

We will first explain how the models are constructed, with respect to the decisions about the dependent variable, the independent variables and the classification algorithms. It is to be expected that using state-of-the-art classification algorithms on a labelled dataset will increase the home detection performance. Of course, the scope of a labelled approach is more limited than the unlabelled heuristic methods, as this method is only applicable if at least some home locations are known. This method also provides a substantiated indication of the maximal attainable performance of home detection algorithms based on CDR data.

3.3.4.1 Binary Dependent Variable: Home Tower

In order to build a model that will identify the most likely home tower, we need to feed our algorithm with observations to learn from. We construct a

base table that contains observations for home towers as well as non-home towers. Both categories are needed for the model in order to learn how to distinguish between both. The structure of the base table is represented in Table 3.1.

Selecting observations for the home towers (class 1) is straightforward, as we know for each user in our training set what the home tower is. Selecting non-home towers is more involved as this choice is less obvious and will affect the results. In theory, every tower that is not the home tower for a certain user can be selected as part of this class (class 0). However, this would mean that there would be zero activity for the selected user on the majority of the towers in this class. The resulting model would only learn to distinguish between used and non-used towers and will therefore be useless for the identification of the home location. We therefore select only towers that have been used at least once. By doing this, it becomes much harder for the model to distinguish between both classes. However, the model will be much more informative and relevant.

Id	Tower	Dependent variable	Independent variables		
			calls_in_nbr	calls_in_dur_total	...
1	home tower	1	5	210	...
1	non-home tower 1	0	2	103	...
1	non-home tower 2	0	1	504	...
2	home tower	1	10	1243	...
2	non-home tower 1	0	3	239	...
2	non-home tower 2	0	12	2087	...
2	non-home tower 3	0	2	96	...
...	...	1/0

Table 3.1: Structure of the base table for the predictive modelling approach. The dependent variable indicates whether the tower is the home tower for the id. The independent variables are calculated based on the CDR data observed on that tower for the id. This structure results into a base table with home and non-home towers and can therefore be used to build a model that distinguishes between both types of towers. The tower with the highest predicted home tower probability will be selected as the home location.

3.3.4.2 Independent Variables

We constructed 30 independent, or explanatory, variables. Twenty-two of these are constructed based on the three categories identified in Section 3.3.3. Eight social network based variables are included as well. We present the variables in Table 3.2. Note that it follows from the structure of the base table that every variable is calculated per user, per tower.

In order to use the logic in the home detection heuristics, the variables needed to be translated in order to fit in the structure of the base table. The Act_2 heuristic for example selects the tower that had the most distinct use days as the home tower for the selected user. Translating this into an independent variable, this becomes the number of distinct days on this tower, for this user. It is straightforward to see how the other heuristics were translated into variables as well, following the same logic.

The first activity heuristic, Act_1, selects the tower with the highest number of activities. We decided to split this into multiple variables, separated into incoming versus outgoing and voice calls versus text messages. We also enriched this by calculating other measures such as average and standard duration of calls and the percentage of the activity of the user on this tower.

Literature showed the predictive power of social networks. We therefore augmented the base table with variables that take into account this social network. We included three variables based on how frequently the contacts of an individual also use a certain cell tower. Five variables were included based on tie strength as defined in Section 3.2.4.

Furthermore, we will explore the added value of social network data by building three categories of models: the *full* models are trained on all variables, the *withoutsocial* models do not use any social network based variables, whereas the *socialonly* models do only use the social network based variables.

3.3.4.3 Binary Classification Algorithms

We will examine four frequently used binary classification algorithms; logistic regression, random forest, adaboosting and neural network models. R statistical software was used to implement these models (R Core Team, 2020). The random forest models were run with 1,000 trees, as recommended by Breiman (2001). The adaboosting models are implemented following the method of Friedman et al. (2000) and allowed 150 boosting iterations. The neural network models are implemented with 40 units in the hidden layer and we restricted the algorithm to perform a maximum of 2,000 iterations. Evaluating multiple classifiers assesses the robustness of the labelled predictive modelling approach.

The output of these models is the probability of belonging to a certain class. For every user, the tower with the highest predicted home tower probability is selected as the home tower. We will evaluate the predictive modelling approach based on the same measures as the benchmark methods; accuracy (percentage correctly predicted home towers) and the average

Activity based	
calls_in_nbr	Number of incoming calls.
calls_in_dur_total	Total duration of incoming calls.
calls_in_dur_avg	Average duration of incoming calls.
calls_in_dur_sd	Standard deviation of duration of incoming calls.
calls_in_perc_on_tower	Percentage of incoming calls.
calls_out_nbr	Number of outgoing calls.
calls_out_dur_total	Total duration of outgoing calls.
calls_out_dur_avg	Average duration of outgoing calls.
calls_out_dur_sd	Standard duration of outgoing calls.
calls_out_perc_on_tower	Percentage of outgoing calls.
sms_in_nbr	Number of incoming text messages.
sms_in_perc_on_tower	Percentage of incoming text messages.
sms_out_nbr	Number of outgoing text messages.
sms_out_perc_on_tower	Percentage of outgoing text messages.
act_2_distinct_days	Number of distinct days.
act_3_7pm_9am	Number of activities between 7PM and 9AM.
act_4_nighttime_10pm_7am	Number of activities between 10PM and 7AM.
act_5_distinct_nights_6pm_8am	Number of distinct nights (6PM - 8AM).
Inactivity based	
inact_5	Number of inactive periods (>5h).
inact_7	Number of inactive periods (>7h).
Clustering based	
clust_Basic1	Number of activities within a 1 kilometre perimeter.
clust_Basic2_7pm_9am	Number of activities, 1 km perimeter (7PM - 9AM).
Social network	
soc_sum_cdr	Sum of number of activities of contacts.
soc_nbr_contacts_use_loc	Number of contacts that use this location.
soc_perc_contacts	Percentage of contacts that use this location.
soc_tiestrength_avg	Average tie strength with contacts on this location.
soc_tiestrength_median	Median tie strength with contacts on this location.
soc_tiestrength_min	Minimum tie strength with contacts on this location.
soc_tiestrength_max	Maximum tie strength with contacts on this location.
soc_tiestrength_sum	Sum of tie strength with contacts on this location.

Table 3.2: Variables in predictive model. Every variable is calculated based on the structure of the base table, therefore ‘on this tower for this id’ can be added to the description of every variable. All variables are built on the 5-week period of CDR data.

error distance to the actual home tower, thereby following the same 5x2cv procedure.

3.4 Results

3.4.1 Benchmarks

We present the results of our benchmark study of the heuristic home detection methods introduced in Section 3.2.2 in Table 3.3. These methods do not need a training phase, as opposed to a predictive modelling approach and are therefore immediately applied to the test folds in the 5x2cv approach. The reported numbers are the averages of the metrics over the 10 test folds.

Model		Correct home tower (%)	Distance to home tower (km)
Activity	Act_1	57.56	5.671
	Act_2	60.16	4.591
	Act_3	56.54	5.800
	Act_4	41.69	8.425
	Act_5	59.11	4.816
Inactivity	Inact_5	60.69	4.499
Two-step clustering	Clust_Basic1	53.69	6.885
	Clust_Basic2	55.01	6.276
	Clust_ha_1	6.41	25.133
	Clust_ha_1_6	6.03	26.155
	Clust_ha_3	5.34	28.850
	Clust_ha_5	4.87	32.587
	Clust_hs_1	5.88	25.517
	Clust_hs_1_6	5.15	26.875
	Clust_hs_3	3.76	30.456
	Clust_hs_5	3.22	34.898

Table 3.3: Benchmark results. The results indicate that the inactivity category performs best, with the highest accuracy (% correct home tower) and the lowest average error distance, followed by the activity and two-step clustering based methods. The maximal attainable performance with the existing home detection algorithms lies in the range of 60% accuracy and an average error distance of 4.5km.

The best performing category is the inactivity category. The average accuracy is 60.69% and the average distance error is only 4,499 metres. The Inact_5 method scores better than the best activity method (Act_2). The

difference is highly significant with an F-value of 72.44 and an associated p-value of $8.95e-05$ for the 5x2cv F-test for difference in accuracy. The idea of modelling sleeping hours, that is the basis of this heuristic, does not require the choice of a specific time of day and therefore allows for a correct home identification for a much larger range of people with different behaviour. This makes this heuristic much more broadly applicable, across cultures and countries as well. Despite its simplicity, the inactivity method also scores a lot better than the more complex two-step clustering methods.

The performance within the second best category, the activity category, is rather consistent. The best method is Act_2, which states that the home tower is the tower that is used on the maximum number of distinct days. The underlying assumption of regularity in this method seems to improve the performance a lot, when compared to Act_1 for example. This method selects the correct home tower in 60.16% of the cases, while the average error in terms of distance is limited to 4,591 metres. Act_2 is significantly better than the second best activity method (Act_5) (F-value 52.89, p-value $1.9e-3$).

A surprising result is that the, more complex, clustering methods do not improve the performance. Comparing the two basic clustering methods with their single-step counterparts (Act_1 for Clust_Basic1 and Act_3 for Clust_Basic2), confirm this. The Hartigan leader based approach tremendously reduces performance, the best model using this approach achieves an accuracy of only 6.41%, whereas even the worst non Hartigan leader based algorithm still achieves 41.69%. In terms of distance, a similar performance drop is observed. We also observe that the adapted version (HA) performs slightly better than the standard version (HS). The higher the threshold value, the further the method deviates conceptually from the single-step methods. The results indicate that a higher threshold value further reduces performance. In the next section of the paper, more settings for the Hartigan based methods will be explored in order to identify what causes their unacceptable performance.

Furthermore, it is remarkable that the heuristics that take into account specific hours for night time (Act_4, Act_5 and Clust_Basic2) do not achieve a higher performance than their counterparts that do not take time into account. The lower performance can partly be explained due to the fact that part of the users do not have any data points during the specified time frame. This obviously limits the maximal performance that can be achieved. Table 3.4 reports the percentage of users for which no prediction can be made because of this issue.

Model	No Prediction
Act_3	1,67%
Act_4	16,12%
Act_5	0,15%
Inact_5	0,13%
Inact_7	0,14%
Clust_Basic2	1,67%

Table 3.4: Percentage of individuals for which no prediction can be made. Due to imposing a time frame in model Act_3, Act_4, Act_5 and Clust_Basic2, it is impossible to make a home location prediction for individuals that have no observations during the selected time frame. Act_4 uses the most restricted time frame, which results into the highest number of these cases. The inactivity methods will produce no prediction if no inactivity period is observed.

3.4.2 Optimisation of Benchmarks

Given that the inactivity method performed best, we decided to further investigate this method. The question is whether optimising its main parameter, being the threshold value for the number of idle hours, further enhances the results. We present the results in Table 3.5 and graphically in Figure 3.3, where Inact_X means the inactivity method with a parameter value of X. Based on the accuracy metric, Inact_6 can be selected as optimal. However, based on the distance metric Inact_7 is optimal. The negligible difference in accuracy between Inact_7 and Inact_6 however is not significant (F-value 0.55, p-value 0.803). We therefore suggest to use Inact_7. Furthermore, the difference between this optimal setting and the arbitrarily chosen value of 5 by Dash et al. (2014) is significant (F-value 6.78, p-value 0.024). We also observe that the performance parameters show an inverted U shape. The performance increases gradually when starting with a threshold of 3 hours, to reach its maximum at the 6/7 hour threshold, after which the performance steadily declines again. This optimisation further demonstrates the added value of having high quality validation data.

Model	Correct home tower (%)	Distance to home tower (km)
Inact_3	60.05	4.748
Inact_4	60.48	4.612
Inact_5 (Dash et al., 2014)	60.69	4.499
Inact_6	60.81	4.455
Inact_7	60.80	4.413
Inact_8	60.73	4.421
Inact_9	60.61	4.448
Inact_10	60.33	4.556
Inact_11	59.90	4.669

Table 3.5: Inactivity method. Optimising the parameter of the inactivity function reveals that the arbitrary 5-hour value of Dash et al. (2014) does not lead to optimal performance. The Inact_7 method is selected as optimal, as it reaches the lowest distance error and an accuracy that is not significantly different from the optimal accuracy of the Inact_6 method.

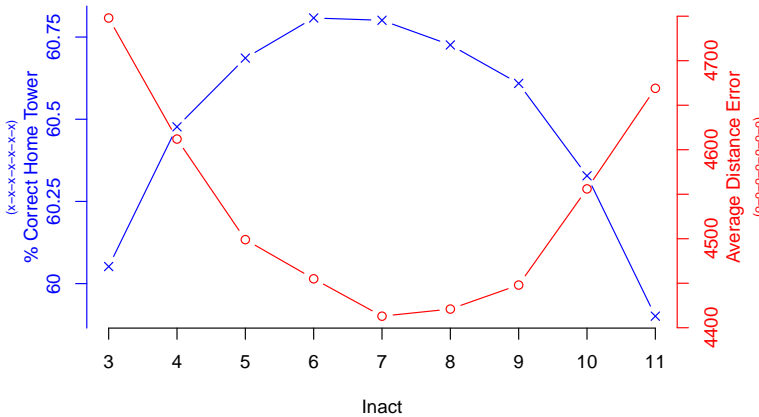


Figure 3.3: Inactivity method. The optimisation exercise shows a continuous shape for the performance functions. The value of 5 as suggested by Dash et al. (2014) does not lead to optimal performance. In terms of average distance error, 7 is the optimal value. In terms of percent correct home tower, 6 is optimal, however the difference with 7 is not significant. We can therefore conclude that 7 is the best value.

The initial analysis revealed that the Hartigan leader based two-step methods have a much lower performance than the other methods. In order to further investigate this and better benchmark this category with the single-step methods, we decided to use different scoring rules in the second

step. We score the clusters with the Act_2 and the Inact_5 method. The results are reported in Table 3.6.

First of all, we again observe that the adapted Hartigan leader algorithm (HA) performs better than the standard version (HS). Also, except for the combination of the adapted Hartigan leader algorithm with Inact_5, a higher threshold distance leads again to lower performance, indicating again that the further the method deviates from the single-step methods, the worse.

Clust_HA_1_Act_2 is now the best two-step counterpart for the single-step Act_2 method. The accuracy however drops from 60.16% to 55.70% and the distance error rises from 4.591 to 5.167 metres. This difference is highly significant (F-value 1248.96, p-value 7.51e-08). The same holds for the Inact_5 versus the Clust_HA_1_Inact_5 method; from 60.69% to 58.28% and from 4.499 to 4.525.

Model	Correct home tower (%)	Distance to home tower (km)
Clust_ha_1_act_2	55.70	5.167
Clust_ha_1_6_act_2	49.40	5.534
Clust_ha_3_act_2	37.82	5.700
Clust_ha_5_act_2	30.04	5.582
Clust_hs_1_act_2	46.75	5.489
Clust_hs_1_6_act_2	36.53	5.883
Clust_hs_3_act_2	20.47	6.150
Clust_hs_5_act_2	14.71	6.259
Clust_ha_1_inact5	58.28	4.525
Clust_ha_1_6_inact5	53.33	4.532
Clust_ha_3_inact5	41.77	4.493
Clust_ha_5_inact5	32.35	4.490
Clust_hs_1_inact5	50.12	4.600
Clust_hs_1_6_inact5	40.59	4.686
Clust_hs_3_inact5	22.62	4.983
Clust_hs_5_inact5	15.39	5.373

Table 3.6: Scoring the Hartigan leader based methods with known activity/inactivity heuristics. The scoring method proposed by Zagatti et al. (2018) resulted in bad performance. We therefore executed the clustering method with scoring methods based on the best inactivity/activity heuristics. This enhances the results for the clustering method strongly. However, when compared to the single-step methods, the performance is reduced despite the higher complexity of the two-step method.

Our results indicate that the most important part of the two-step approach is the second step, as the performance of scoring with better measures (Act_2 and Inact_5) clearly outperforms the standard option. This

second step corresponds to the single-step methods and underlines the importance of a well chosen heuristic. Nevertheless, this study shows that adding a clustering step to the single-step methods does not add any value.

The performance depends heavily on the assumptions and parameters in the heuristic. An optimal approach would be able to automatically select the best parametrisation and assumptions in every case. This idea is more embedded in a pure predictive modelling or classification approach.

3.4.3 Predictive Modelling Approach

The results of this approach can be found in Table 3.7. We will first compare these results with the benchmark heuristic approach, followed by a discussion of the different binary classifiers and a discussion of the added value of the social network data.

Model	Correct home tower (%)	Distance to home tower (km)
full_logreg	61.36	4.453
full_rf	71.86	2.952
full_adaboost	71.86	2.925
full_neuralnet	72.08	2.848
withoutsocial_logreg	60.78	4.682
withoutsocial_rf	71.42	3.094
withoutsocial_adaboost	71.40	3.070
withoutsocial_neuralnet	71.66	2.941
socialonly_logreg	37.65	8.098
socialonly_rf	32.74	9.712
socialonly_adaboost	37.64	8.480
socialonly_neuralnet	38.15	8.365

Table 3.7: Predictive method results. *The improvement of the logistic model, when compared to the best unlabeled heuristic methods is minor. However, the other classifiers strongly enhance the results. The best method (full_neuralnet) reduces the average distance error with 1,565 metres and improves the correct home percentage with 11.28 percentage points.*

The best predictive model is the neural network model that uses all created variables (*full_neuralnet*). This model predicts the correct home tower in 72.08% of the cases and has an average distance error of only 2.848 metres. Compared to the results of the optimal heuristic method (Inact_7, 60.80% and 4.413 metres), the advantage of using a predictive modelling approach is evident.

In terms of classifiers, the neural network model is generally best, closely

followed by the adaboost models and random forest. The frequently used logistic regression model scores much lower and actually performs in the same range as the (best) heuristics. The assumptions underlying the logistic model do not seem to fully accommodate the case of home prediction. However, for the models that only use social variables, logistic regression is the second best model in terms of accuracy and even the best in terms of distance error. Taking into account these results and the high interpretability of a logistic regression model, when compared to the other classifiers, this model is advised in a context where only social network data is used.

It is clear that taking all data into account (*full* models) leads to the highest accuracy and the lowest average distance error. The best full model (*full_neuralnet*) performs significantly better than the best model without social network variables (*withoutsocial_neuralnet*) (F-value 33.76, p-value 5.79e-4). The same holds for every classifier in the full model compared with its *withoutsocial* model counterpart. The increase in home detection performance is a bit more outspoken for the weaker logistic regression classifier. In a case where a clearly interpretable model, such as a logistic regression model, is required it therefore becomes even more important to add social variables in order to compensate the loss in performance due to a weaker classifier.

The *socialonly* models obviously can not reach the same performance as all other models that do use the data of the individual itself. Nevertheless, our results do confirm the value of social network data in the context of home detection. In the literature review, evidence was found for home detection in online social networks based on the location of contacts in the social network (Backstrom et al., 2010) and it was anticipated that the value could be much larger in a CDR data set as this is an even better representation of the actual social network of people. Recall that Backstrom et al. (2010) found that the home location was predicted within 40 kilometres of the actual home for 70% of the users. Our best *socialonly* model in terms of distance, the logistic regression based model, lowers this distance at 70% to only 5.2 kilometres. This model is able to classify 95.89% of the data set within the error range of 40 kilometres. The improvement of using CDR data instead of OSN data is substantial. Of course, in order to achieve a fair comparison, one would need to replicate these findings within the same geographical boundaries. Nevertheless, the size of the improvement is a clear indication that a substantial improvement is to be expected.

3.4.4 Exploratory Performance Analysis

In order to attain more insight into the performance of the most important models and the performance of home location prediction in general, we develop a further, exploratory performance analysis. We propose seven variables that can be expected to have a relation with the performance measures used in this research.

A first variable *total number of CDR observations* is constructed in order to evaluate the often stated premise that more data leads to better predictions (e.g. Junqué de Fortuny et al. (2013)). Although this largely holds in many situations, one needs to be aware of its possible adverse effect in the case of home prediction. There is a strong, highly significant positive correlation (44.46% p-value < 0.001) between the total number of CDR observations and the number of *distinct towers used*, a second explanatory variable. The consequence of observing more towers for an individual is that selecting the correct tower from the larger set of towers becomes more difficult. This would be very outspoken when using a naive model that randomly selects one of the used towers as the home tower. Although more advanced models will less be affected by this issue, we can still expect a negative relation with performance.

Predicting the correct home tower can also be impeded by the number of towers in the area surrounding the home tower. Urban environments will have a higher tower density than rural areas, which can again make it more difficult to predict the correct tower. The same rationale is the motivation behind the two-step clustering approaches, which are aimed at reducing this problem by clustering nearby towers. Although the results of the clustering methods were unsatisfactory, the rationale still makes sense, which leads to the construction of the following two variables that model tower density. The *number of towers within 2 kilometres of the home tower* can be expected to have a negative effect on performance, whereas the opposite holds for the *distance to the closest tower* to the actual home tower.

The analysis in Section 3.4.3 indicated the added value of social network data. The *number of contacts* variable is created to assess how it can be more (or less) difficult to predict the home location for people with more social connections.

Finally, we formulate two variables, based on the two best performing heuristic methods, Act_2 and Inact_7. These variables calculate for every individual the total number of counts for these heuristics. For Act_2, this results into *total Act_2 counts*, which is the sum of the distinct days, over all towers used by the individual. This variable captures how many measurements we have to base the well performing Act_2 heuristic on. For

Inact_7, this results into *total Inact_7 counts*, which counts the total number of observed inactivity periods, corresponding to the level of inactivity.

Table 3.8 reports the correlations between the seven explanatory variables and the performance of the best methods in the different relevant categories: activity heuristic (Act_2), inactivity heuristic (Inact_7) and predictive model (Full_neuralnet). Note that a positive correlation with % correct home indicates that performance increases when the explanatory variable is higher. However, the opposite is true for the distance error measure, as a higher distance error implies a lower performance. This explains why this correlation usually switches sign, when compared to the corresponding correlation of the % correct home metric.

The results indicate that having more observations for an individual does not lead to higher performance in the case of home detection. The effect that more observations results into more towers clearly explains this finding, as the number of distinct towers has an even higher negative correlation with the performance metrics. Tower density around the home tower, as measured by the number of towers within a 2 kilometre radius and the distance to the closest tower, has a strong relation with the percentage correct home locations. A higher tower density around the home tower makes it clearly more difficult to detect the correct home tower. The influence of tower density on the distance error is however inconclusive, it is much more limited or even insignificant. This shows that the algorithms still manage to detect a home tower that is relatively close to the actual home tower. As discussed in Section 3.3.2, this also demonstrates the importance of taking the distance measure into account, as the accuracy alone might underestimate the actual performance in the case of high tower density. A higher number of contacts is related to higher performance for the heuristic methods, however the opposite is true for the predictive neural net model.

A higher number of observations on which the Act_2 (number of distinct days) heuristic can be based, surprisingly leads to lower performance for all models. This result is however perfectly in line with the results of the total number of CDR observations and the number of distinct towers used, all three are measures that to some extent actually model the same concept: the level of activity. In general, the results indicate that it is more difficult to identify the correct home location for more active users. This makes it interesting to have a closer look at the opposite category of inactive methods as well. We observe that the reversed effect holds for the correlation between the *total Inact_7 counts* and the performance of the heuristics. This result is not apparent for the predictive neural net model, where the results are inconclusive. For the heuristic methods, we see that observing more inactive periods, as modelled by *total Inact_7 counts* improves per-

	Act_2			Inact_7			Full_neuralnet		
	%Correct	Distance	Error	%Correct	Distance	Error	%Correct	Distance	Error
Total number of CDR observations	-1.30% ***	0.61%		-1.31% ***	0.73%	*	-5.58% ***	2.20% ***	
Distinct towers used	-5.57% ***	5.76%		-5.28% ***	4.82%		-10.26% ***	5.93% ***	
Number of towers within 2 km of home tower	-11.19% ***	0.33%		-11.43% ***	0.44%		-2.99% ***	-0.79% *	
Distance to closest tower	15.89% ***	-1.39%		16.30% ***	-1.54%		6.77% ***	0.16%	
Number of contacts	1.26% ***	-1.25%		1.03% **	-1.48%		-6.62% ***	0.71% *	
Total Act_2 counts	-7.72% ***	2.18%		-7.51% ***	1.59%		-15.34% ***	3.82% ***	
Total Inact_7 counts	2.91% ***	-3.54%		3.93% ***	-4.45%		-7.53% ***	-0.86% **	

Table 3.8: Correlations between explanatory factors and performance measures of the best performing methods. The level of significance is indicated by *** ($p\text{-value} \leq 0.001$), ** ($p\text{-value} \leq 0.01$), * ($p\text{-value} \leq 0.05$).

formance. It also has the most pronounced negative effect on distance error for the heuristic methods. In contrast to the level of activity, a higher level of inactivity is positively correlated with a higher performance.

3.4.5 Combined Inactivity Activity Heuristic Method

The benchmark study revealed that the best home prediction method in literature is Inact_5 with 60.69% of the home towers correctly predicted and an average distance error of 4.499 kilometres (see Table 3.3). An optimisation of the parameter of this approach revealed that these values could be improved to 60.80% and 4.413 kilometres (see Table 3.5). The main success of the inactivity approach lies in the fact that it effectively models the desired concept of inactivity. This concept was introduced to represent sleeping hours, given that one usually sleeps at his or her home location. Given the data period of five weeks in our research, people would sleep 35 times. In an ideal situation and given perfect data, we would therefore observe 35 inactivity periods for every individual. The total number of inactivity periods is represented by *total Inact_7 count* from the previous section. Figure 3.4 displays the distribution of this variable. The distribution has the highest relative frequency at 35. The inactivity method therefore clearly achieves to model the desired concept, explaining its strong performance.

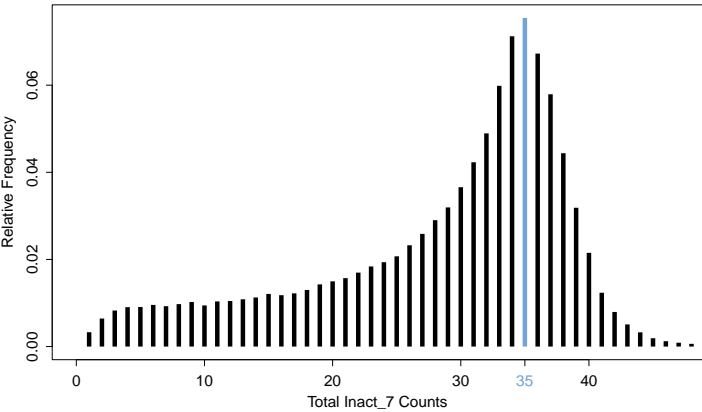


Figure 3.4: Distribution of total Inact_7 counts. This distribution of the total number of inactivity periods is peaked at 35, the number of days in the data set. This indicates that the inactivity heuristic embodies the concept of inactivity, as a proxy for periods of sleep.

Based on the strong performance of Inact_7 and the results in the pre-

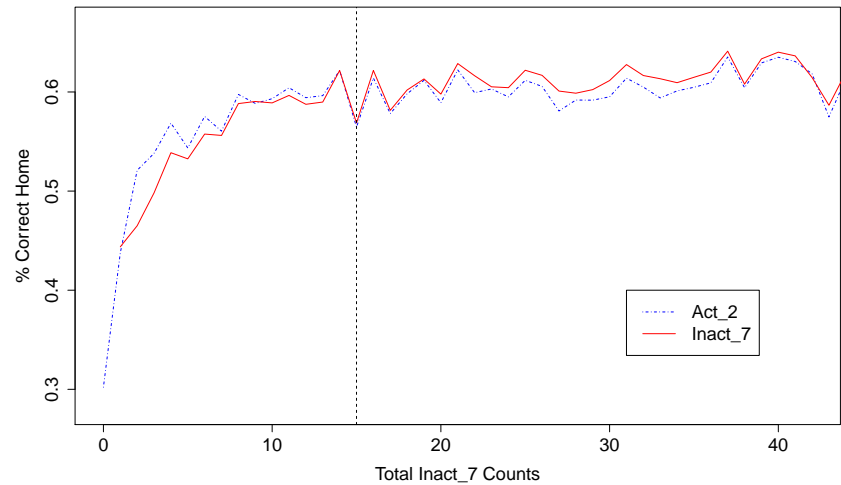
vious section, a new heuristic has been developed in order to further improve the quality of home detection in an unlabelled setting, where the optimal predictive modelling approach (full_neuralnet) can not be used. The strongest correlation between performance and the explanatory variables was observed for the concept of tower density. Nevertheless, this idea can not be used to further develop a heuristic method for home detection, as the calculation of the number of towers around the home location needs the unknown home location as a prerequisite. The inactivity method seemed to perform best, however, we observed that the performance was positively correlated with a higher level of inactivity. This knowledge hints at the possibility that the inactivity method can be improved when replaced by another, activity based, heuristic at low levels of inactivity. The optimal method to do this is the best activity method, Act_2 (distinct days), which has a slightly lower overall performance (60.16% and 4.591km). The performance of Act_2 has a lower positive correlation with total Inact_7 counts. This combined knowledge indicates that the performance of Act_2 could indeed be higher than Inact_7 at lower levels of inactivity.

These premises are empirically validated and represented in Figure 3.5. The figures confirm the positive correlations from Table 3.8 between total Inact_7 and the accuracy (% correct home) for both the activity and inactivity method, as well as the negative correlation with the distance error. Furthermore, the higher correlation for the inactivity method, when compared to the activity method leads to a more rapid increase/decrease, thereby leading to an intersection around the inactivity level of 15. Act_2 does outperform Inact_7 for low levels of inactivity. The figures also indicate that Act_2 always provides a prediction, whereas this is not the case for Inact_7 (0.14% of the individuals have an activity level of zero, leading to no prediction for Inact_7, see also Table 3.4). We therefore propose to use the Act_2 heuristic for individuals with a low level of inactivity (defined here as total Inact_7 count < 15) and use Inact_7 for higher levels of inactivity (total Inact_7 count ≥ 15).

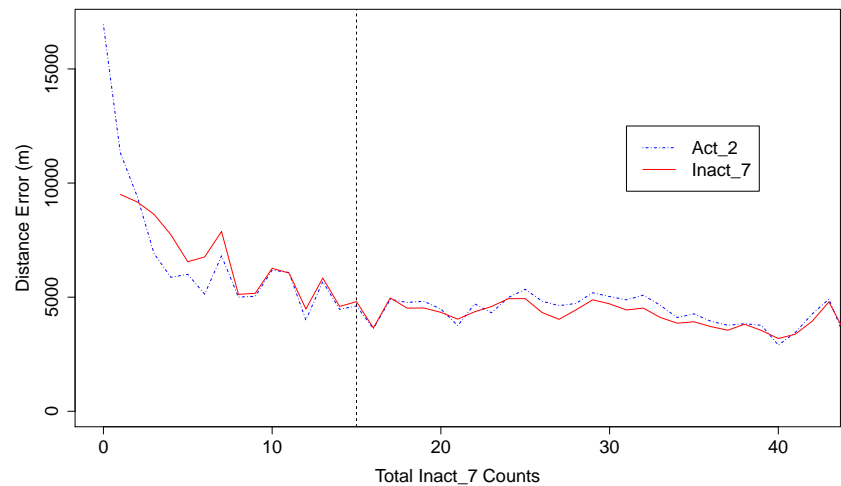
The accuracy rises to 61.00%, while the average distance error decreases to 4.365 kilometers. Applying the 5x2cv F-test informs us that this result is significantly better than both Inact_5, the best heuristic method observed in literature (F-value 157.40, p-value 1.31e-05) and our optimised version Inact_7 (F-value 157.40, p-value 1.31e-05).

3.4.6 Summary of Results

In Figure 3.6, we plot the reversed cumulative distribution of the distance error for the best performing models for the different categories. Three



(a) Accuracy (% Correct Home) in function of the level of inactivity (Total Inact_7 Counts).



(b) Average distance error in function of the level of inactivity (Total Inact_7 Counts).

Figure 3.5: Combined Inactivity Activity Heuristic. At low level of inactivity, the activity heuristic (Act_2) is advised. At a higher level of inactivity, the inactivity heuristic (Inact_7) is advised. This new heuristic method results in an improved accuracy of 61.00%, and a reduced average distance error of 4.365 kilometres.

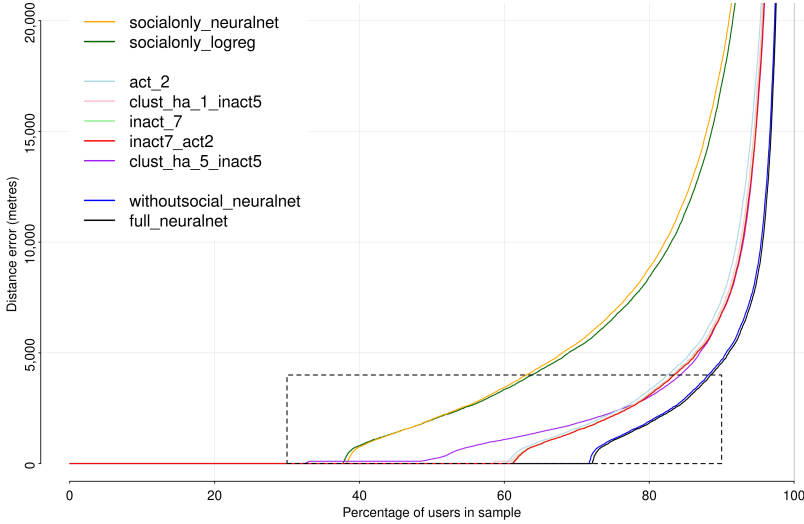


Figure 3.6: Reversed cumulative distance error for best performing models in different categories. The plot displays which error can be achieved for what percentage of the data set. The intersection with the horizontal axis is equal to the percent correct home measure. The more a certain method is situated to the bottom right of the figure, the better the performance, as this means that a higher percentage of the data has a lower distance error. Three major groups can be identified in terms of results; the social only models, the heuristic methods and the predictive models. The legend displays the methods from left to right at the upper horizontal line of the dashed rectangle. Figure 3.7 zooms in on the selected region, bounded by the dashed rectangle, to provide more detailed insights.

groups are identified with similar performance. The first group contains the predictive models that only use purely social network variables. It is clear that using merely these variables weakens the results, however these models are interesting if no location data about the individual itself is available and have great potential in such cases. Figure 3.7 zooms in on the region where the methods start to differ. We observe that *socialonly_neuralnet* has a higher accuracy than *socialonly_logreg*, but nevertheless a higher average distance error, as the neural network model has a larger percentage of users with zero distance error, but is outperformed in terms of distance error quickly as the percentage of users taken into account increases.

A second group consists of the heuristic benchmark methods. Also in that group, we observe that *clust_ha_1_inact5* has a much higher percentage correct home than *clust_ha_5_inact5*, but nevertheless a higher average distance error as well, as the method more quickly increases in the plot.

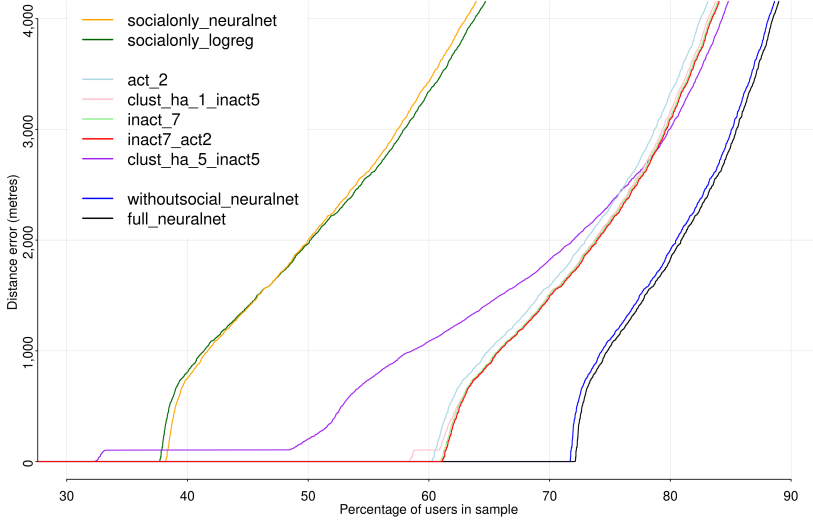


Figure 3.7: Zoomed reversed cumulative distance error for best performing models in different categories. The plot shows a more detailed representation of the performance of the models in the region where they start to differ most. Certain methods combine a lower accuracy (percent correct home) with a lower distance error, as they continue below the other method towards the tail of the data set (e.g. socialonly_neuralnet versus socialonly_logreg).

The third and best performing group consists of the labelled predictive models that use all (*full_neuralnet*) or all, except social variables (*without-social_neuralnet*).

In summarising Table 3.9, we report the performance metrics (*percent correct home* and *average distance error*) for the best method in each category. Remark that for two of three original benchmark categories, the optimal method is the result of an adaptation in Section 3.4.2, namely Inact.5 (Dash et al., 2014) was optimized to Inact.7 and the optimal two-step clustering method was Clust.ha.1.inact.5. All original benchmark heuristics are further outperformed by the newly introduced combined activity inactivity heuristic.

3.5 Conclusion and Future Research

Multiple home detection methods for CDR data have been developed in literature. Nevertheless, there still was a high need for a thorough validation of these methods as ground truth data typically lacks in this type of research. Our unique dataset enabled this benchmark study. The benchmark

Model	Correct home tower (%)	Distance to home tower (km)
Single-step heuristic methods		
Activity heuristics	60.16	4.591
Inactivity heuristics	60.80	4.413
Combined heuristic	61.00	4.365
Two-step heuristic methods		
Two-step clustering	58.28	4.525
Predictive modelling methods		
Social only	37.65	8.098
Without social	71.66	2.941
Full	72.08	2.848

Table 3.9: Results home detection methods. The performance metrics (percent correct home tower and average distance error) are reported for the best method in each category. The Act_2 benchmark is the only benchmark in this figure that was not optimised further; beyond the original method proposed in previous literature. Both Inact_7 and Clust_ha_1_inact5 are adaptations of the original methods. The best heuristic method (act_2&inact_7) is a result of this research. The best overall model (full_neuralnet) improves strongly upon the best heuristic (act_2&inact_7), however is more limited in scope as it requires labelled data.

study revealed that the more complex two-step clustering methods do not lead to higher performance, but on the contrary decrease performance. A second important result is that methods that require the specific choice of a range of hours that define night time perform less good than methods that do not use such parameter settings. The best benchmark heuristics (the distinct days based activity method and the inactivity based method) are both examples of this. The best activity based method relies on the number of distinct days that an individual uses a certain location. This result confirms the research by Vanhoof et al. (2018c). This method does not incorporate the time of the day and is based on the assumption that regularity is an important aspect in identifying the correct home location. The best inactivity method also avoids defining a time of day by modelling periods of inactivity, which aim to model the sleeping hours. These periods can occur throughout the entire day, in order to accommodate for people that work in shifts and people with irregular sleeping hours in general. Our analysis identified that the performance of the inactivity approach can be optimised by changing the required length for the inactive periods. The non-validated arbitrary choice of 5 led to suboptimal results, a longer period of 7 hours improved the performance of this technique. Based on the benchmark results and an analysis of the factors that influence the performance of home detection, we propose a new heuristic that further improves the results in an unlabelled setting. The new combined inactivity activity heuristic uses the activity method for low levels of total inactivity and the inactivity method for sufficiently high levels of inactivity.

Our individual level validation revealed that the best heuristic predict the correct home tower in 61% of the cases and that the average error between the predicted and actual home tower is less than 4.4 kilometres. In an unlabelled setting, where no home locations are known, one needs to rely on these heuristics and the expected performance is as above. In cases where part of the data is labelled, meaning that for a part of the data a home location is known, our research indicates that it is strongly recommended to use a labelled predictive modelling approach. Using a binary classification model enhances the accuracy to more than 72% and reduces the error distance to 2.8 kilometres. The scope of this labelled approach is of course limited to applications with partly labelled data, but this analysis also provides an indication of the maximal attainable performance in home detection when using CDR data.

As a third contribution, we evaluated the value of social network data for home detection in CDR data. We found that adding information about the social network significantly improved the accuracy of the predictive model from 71.66% to 72.08%, and reduced the average distance error from

2,941 metres to 2,848 metres. A predictive model that used merely the data of the social network was able to achieve an accuracy of 37.65% and an average distance error of 8,098 metres, which is a large improvement on previous research using Facebook data (Backstrom et al., 2010).

Identifying the home location is key to many applications, hence our results can be used in a wide variety of research and business applications. Epidemiological models have used CDR data before. CDR data can help modelling the spread of a virus and investigate the impact of measures such as the advice to stay at home during the Covid-19 pandemic. It is obviously crucial to not only have an appropriate data source, but also the accompanying methods to identify what the home location of people in the model is. Previous research (Blondel et al., 2012; Vanhoof et al., 2018b) explained that home detection with CDR data can also be used to replace the outdated or unavailable census data in developing countries. Research about homework commuting and commuting patterns has a great impact on society by affecting people's everyday life and by studying the impact of different commute behaviours on our carbon footprint for example. This type of research requires an accurate home location and thus clearly benefits from a solid home detection method. Telecommunications and transportation infrastructure can be further improved and deployed based on findings derived from CDR data analysis. The developed *socialonly* models are of academic interest, but also spark business applications, as these models enable to get insight in the location of non-customers as well. This might lead to identifying areas where the telecom provider is under-represented and might trigger specific tailored marketing campaigns that boost the customer base of the provider. This way, home detection does not only serve policy makers, human mobility researchers or epidemiologists, but also proves its relevance for marketeers amongst others in business.

We urge researchers to replicate our results on CDR datasets in other countries and cultures in order to assess the robustness in these different settings. Using data of other telecom providers also eliminates a possible selection bias introduced by the fact that a certain provider might attract only a certain segment of the market and therefore produces a biased population sample. This research used five weeks of CDR data, it remains to be investigated how using a longer period of data might affect the results. This aspect is especially important for the proposed new combined inactivity activity heuristic. The concept of using the activity method for a low level of inactivity remains robust. However, the cut-off value where the heuristic switches from the activity to the inactivity method may depend on the length of the observed period. Further research needs to determine whether this value is absolute or relative to the number of observed days. Further-

more, we strongly encourage further research into the promising topic of the models that use social network data, as our preliminary results already reveal great potential. Finally, the ultimate impact of investing in a better home detection method needs to be quantified in the many different applications.

4

From One-Class to Two-Class Classification by Incorporating Expert Knowledge: Novelty Detection in Human Behaviour¹

Abstract

One-class classification is the standard procedure for novelty detection. Novelty detection aims to identify observations that deviate from a determined normal behaviour. Only instances of one class are known, whereas so called novelties are unlabelled. Traditional novelty detection applies methods from the field of outlier detection. These standard one-class classification approaches have limited performance in many real business cases. The traditional techniques are mainly developed for industrial problems such as machine condition monitoring. When applying these to human behaviour, the performance drops significantly. This paper proposes a method that improves existing approaches by creating semi-synthetic novelties in order to have labelled data for the two classes. Expert knowledge is incor-

¹Based on: Oosterlinck, D., Benoit, D. F., & Baecke, P. (2020). *From one-class to two-class classification by incorporating expert knowledge: Novelty detection in human behaviour*. *European Journal of Operational Research*, 282(3), 1011-1024.

porated in the initial phase of this data generation process. The method was deployed on a real-life test case where the goal was to detect fraudulent subscriptions to a telecom family plan. This research demonstrates that the two-class expert model outperforms a one-class model on the semi-synthetic dataset. In a next step the model was validated on a real dataset. A fraud detection team of the company manually checked the top predicted novelties. The results show that incorporating expert knowledge to transform a one-class problem into a two-class problem is a valuable method.

4.1 Introduction

Novelty detection is concerned with detecting data that is different from the known data that characterizes a normal or stable situation. The term novelty detection is frequently used interchangeably with the more narrow term one-class classification. Models in this domain are used when only one class is known, while the other class is absent, poorly sampled or not well defined (Khan and Madden, 2014). One-class models rely heavily on outlier assumptions. These methods are therefore suited for applications with clear outliers, where the novelties do not interfere with the normal data. In machine monitoring for example, where the normal data is stable and outliers are usually pronounced, these methods are applicable (Japkowicz et al., 1995). However, applications that classify human behaviour typically possess a much higher variability in the data, resulting into a less strict boundary between novelties and the normal data. Novelties are not always outliers and outliers are not always novelties (Das et al., 2016). To deal with this increased classification difficulty, we need a method that uses more than solely the data of the one class of normal behaviour. As one-class classification is a harder problem than two-class classification (Tax and Duin, 2001), there is value to be found in the transformation of the one-class problem into two-class. Previous research developed methods to generate artificial data for the unknown class. As will be shown in this study, this generated data is however not informative enough to effectively boost performance in applications with diverse human behaviour. This research therefore proposes a method that incorporates expert knowledge to generate data for the unknown class. Modelling human behaviour with the support of human experts proves to be a good match.

The methodology is evaluated through a case study with a large European telecommunications provider. The company released a new mobile offering where customers can bundle themselves in a so called *family plan* (Desai et al., 2018). Due to unique factors of this product, people could take advantage of the service by using it in a way that is not allowed by general

terms and conditions. This *subscription fraud* leads to losses in revenue. As fraud detection is becoming more and more important in preventing these losses (Barse et al., 2003), the goal was to develop a state-of-the-art fraud detection system (FDS) that distinguishes normal users from fraudsters. Before the launch of the new product, all data contains only non-fraudulent customers by definition. Once the product is launched, fraudulent cases will occur in the dataset, however those are unidentified and unlabelled, which rules out traditional two-class classification. The *post-launch* dataset will be used to validate the proposed model.

In the remainder of the paper, the main novelty detection approaches established in literature and their extensions are reviewed. In the methodology section, our expert method for the transformation into a two-class problem is developed. This method is benchmarked against other methods in the subscription fraud case study and the results are empirically validated by means of manual inspection on the *post-launch* data.

4.2 Literature Review

4.2.1 Novelty Detection

Novelty detection is a major area of research. There are several closely related fields of which many are used as synonyms; one-class classification, anomaly detection, outlier detection, concept learning, data description, single-class classification (El-Yaniv and Nisenson, 2007). Less commonly used are the terms noise detection, deviation detection and exception mining (Hodge and Austin, 2004). The terms novelty and anomaly detection are broader in scope than the often as a synonym used one-class classification (OCC). OCC, as introduced by Moya et al. (1993), is merely one approach to tackle novelty detection. However, it is the most common approach in this respect.

Novelty detection is concerned with classifying data that differ in a certain way from the available data in the training phase (Pimentel et al., 2014). Figure 4.1 displays a simplified representation of the concept. Cases from only one class are known, this class is referred to as target or positive class, while the unknown class is referred to as unstable, negative or outlier class (Hempstalk et al., 2008; Khan and Madden, 2014; Clifton et al., 2014). Throughout this research, the terms positive and negative class will be used.

One-class classification builds a model that describes the positive class, also called a model of stability (Clifton et al., 2014). This model only uses data from the known positive class. At prediction time, the model classifies new examples as a novelty or as being part of the positive class. Different

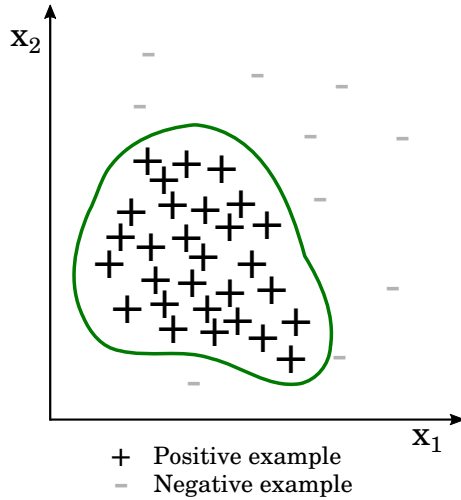


Figure 4.1: Novelty detection. Simplified representation with only two dimensions and perfect class separation. Only instances for the positive class are known. Data for the negative class is not (readily) available or unlabelled. Typically, a one-class model is employed (cf. the line surrounding the positive cases).

types for these models will be discussed in 4.2.2.

Ding et al. (2014) explain that novelty detection is mostly based solely on the positive data since there is enough data about normal events, but none or only scarce data about non-normal events. Furthermore, it is often costly to acquire data about abnormal events. In these situations, it is general practice to fall back to these one-class, unsupervised approaches since developing explicit models for the novelty class is hard (Das et al., 2016).

Novelty detection techniques are traditionally developed and applied in more industrial applications, such as the monitoring of manufacturing processes (e.g. Al-Habaibeh and Parkin (2003)), machine condition monitoring (Carino et al., 2016; Clifton et al., 2014), mobile robotics (Sofman et al., 2011) and medical diagnoses (Tarassenko et al., 1995; Quinn and Williams, 2007; Clifton et al., 2011). These settings are usually determined by a stable positive class. Hence, the negative observations resemble more closely outliers in their most strict definition and the traditional outlier based detection methods are adequate. However, in cases where human behaviour plays an important role, both the positive and negative class are much less stable. Patcha and Park (2007) report that the traditional outlier or anomaly

detection models result into high false alarm rates, when applied to network intrusion detection, a case where human hackers aim to intrude the network. The more volatile and varied nature of this data calls for new methods.

In the following, a concise overview of the prevalent approaches to novelty detection is presented. Figure 4.2 displays the main categories and the positioning of our proposed method within the novelty detection literature. A distinction is made between approaches that use only data from the positive class and those that create data for the negative class. The first category can be called one-class or unsupervised novelty detection, the second category two-class or (semi-)supervised novelty detection.

4.2.2 One-Class Novelty Detection

Within the one-class category, three major approaches are identified; probabilistic, distance-based and domain-based methods.

The probabilistic approaches estimate a probability density function of the positive data. The model will then classify points that lie outside of the high density region as a novelty. Both parametric and non-parametric approaches can be used. The multivariate Gaussian distribution is a frequently used parametric example.

$$P(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right) \quad (4.1)$$

The model will estimate the parameters of the multivariate Gaussian distribution. In the simplified bivariate example in Figure 4.1, a distribution will be constructed so that the positive examples are in the high density region and the (unobserved) negative examples would be classified in the tails of the distribution. In general, probabilistic methods for novelty detection are mathematically sound and effective if there is an accurately estimated probability density function (Pimentel et al., 2014). Higher dimensionality or small training sets are however dreadful for their performance. Pearson (2005) states that it is not convenient to find a suitable distribution. Non-parametric approaches can partly solve that issue, but they suffer from the curse of dimensionality and add computational complexity (Hempstalk et al., 2008). The curse of dimensionality implies that the number of required datapoints exponentially increases with the number of dimensions. Furthermore, a test point will be classified as a novelty if it does not follow the identified distribution, however the assumptions of many distributions might be too simplistic for real-life data. Ding et al. (2014) point at the importance of prior knowledge to circumvent this issue.

Distance-based methods include both clustering and nearest neighbour based methods. Nearest neighbour methods are among the most used approaches for novelty detection (Pimentel et al., 2014). The k-nearest-neighbours (k-NN) method is based on the assumption that normal points have close neighbours in the positive training set. A new data point x_{new} is classified as positive if the distance between x_{new} and its k nearest neighbours $NN_k(x_{new})$ is smaller than the distance between $NN_k(x_{new})$ and its respective k nearest neighbours $NN_k(NN_k(x_{new}))$ in the training set. This leads to the following formula for the kNN score.

$$\text{k-NN Score} = \frac{||x_{new} - NN_k(x_{new})||}{||NN_k(x_{new}) - NN_k(NN_k(x_{new}))||} \quad (4.2)$$

The observations in the new dataset that have the highest k-NN scores will be classified as negative. A common distance formula is Euclidean distance. The Euclidean distance formula between a and b for n dimensions is given by:

$$d(a, b) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2} \quad (4.3)$$

Also Manhattan and Minkowski distance are used. Ding et al. (2014) selected k-NN as the best novelty detection method in an experimental evaluation, using ten benchmark datasets with different scale, dimensionality and problem complexity. Distance based methods have the advantage that they require no assumptions about the probability distribution of the normal data. The curse of dimensionality is again present; as the number of dimensions increases, the distance formula uses so many coordinates that the differences in distance will become relatively small. Clustering has the extra downside that the computational complexity increases quickly and the method is therefore not very scalable.

Domain-based methods construct a boundary using only the positive dataset. The density, which is crucial for probabilistic approaches, becomes irrelevant, because these methods are only concerned with the boundary. The reasoning behind this methodology is that one does not need to solve a more general problem than what is necessary (Schölkopf et al., 2000; Tax and Duin, 2004). The two leading approaches within the domain-based category are both based on support vector machines (SVM) and this category is therefore also referred to as one-class SVM. Schölkopf et al. (2000) developed a method that they describe as a natural extension of SVM to the case of unlabelled data. This algorithm returns a function f that is positive in a region that captures the majority of the datapoints and is negative

elsewhere.

$$f(x) = \text{sgn}\left(\sum_i \alpha_i k(x_i, x) - \rho\right) \quad (4.4)$$

This method requires the user to fix in advance a percentage of the positive data that is allowed to fall outside the boundary that defines the positive class (the ν parameter). This means that outliers in the training data are tolerated more, which helps with the issue that not all outliers are examples of the negative class and not all examples of the negative class are outliers (Das et al., 2016). This flexibility is beneficial for the classification of diverse human behaviour, however the impact is expected to be still limited. Also, this parameter has a strong influence on the overall performance and therefore has to be chosen with great care (Manevitz and Yousef, 2001). The second domain-based approach was developed by Tax and Duin (2004). Their support vector data description (SVDD) method defines the novelty boundary as the hypersphere with minimum volume that includes all (or most) of the positive training data. A result of this definition is that the method is not well suited for high-dimensional spaces because of sparseness issues. In general, one-class SVMs are well known and repeatedly used for novelty detection (Clifton et al., 2014).

4.2.3 Two-Class Novelty Detection

The aforementioned unsupervised, purely one-class algorithms are often criticised for their high false negative rates (Das et al., 2016; Ding et al., 2014). Görnitz et al. (2013) also mention their frequently low predictive performance and point at the need for labelled data. Tax and Duin (2001) explain the inferior performance of one-class methods by the fact that the decision boundary is only supported from one side. On top of that, a vast amount of methods have been developed for (two-class, binary) classification and it would be beneficial if novelty detection could make use of these established methods. Two major approaches have been developed with the purpose of assigning a label to the unlabelled data points; manually labelling existing negative points (e.g. Görnitz et al. (2013)) and generating artificial data (e.g. Surace and Worden (2010)).

Manual or expert based labelling has been done through the inclusion of feedback loops (Abe et al., 2006; Görnitz et al., 2013). Abe et al. (2006) use active learning to interactively query the user. The user needs to manually label selected observations. The model learns from this information. The Active Anomaly Discovery (AAD) method, introduced by Das et al. (2016) is very similar and incorporates expert feedback through an interactive data exploration loop. It is clear that these approaches are very inef-

ficient in most novelty detection applications since the overall presence of novelties is very low and it therefore takes a long time before novelties are discovered.

The artificial data generation approaches do not have this drawback and are therefore expected to be more efficient. Steinwart et al. (2005) mathematically prove that it is worthwhile to generate artificial data in order to apply a binary classification algorithm, given that the artificial samples are well chosen. In certain cases it is not possible to use authentic data for the negative class because for example the target service is under development (Barse et al., 2003). Artificial data provides an interesting solution.

A probabilistic approach to artificial data generation was introduced by Hempstalk et al. (2008). Their technique enhances the standard used one-class probabilistic approach by transforming the problem to two-class. Density estimation is used to form a reference distribution for the artificial class ($P(X|-)$). This distribution should be as close as possible to the positive class. $P(X|-)$ is used to generate data for the negative class. The positive data and the generated negative data are then labelled as such and mixed so that two-class classification can be used. The authors use Bayes' rule to combine the density function of the reference distribution ($P(X|-)$) with the class probability estimates ($P(+|X)$) in order to yield a description of the density function for the positive class ($P(X|+)$). This results in the following relation.

$$P(X|+) = \frac{(1 - P(+))P(+|X)}{P(+)(1 - P(+|X))}P(X|-) \quad (4.5)$$

$P(+)$ can be estimated by the proportion of positive examples in the mixed dataset. Using a balanced dataset ($P(+)=P(-)=0.5$) reduces the formula to the following.

$$P(X|+) = \frac{P(+|X)}{1 - P(+|X)}P(X|-) \quad (4.6)$$

Applying a learning algorithm to this two-class training set (which includes both the positive and the generated negative data) results into a class probability estimator that will take the role of $P(+|X)$. $P(X|-)$ can also be calculated if an appropriate function was selected, e.g. a multivariate normal distribution so that instances can be generated from it. Hempstalk et al. (2008) demonstrated with multiple datasets that this artificial data generation method improved performance.

Surace and Worden (2010) use a largely distance based approach, called negative selection, to generate the artificial data. A data point is pseudo-randomly generated using Gaussian distributions. If it is not similar enough

to the existing positive data, it is labelled as part of the negative class. The similarity is calculated using the cosine similarity, which is an alternative for Euclidean distance, using the following formula (where l refers to the length of the vector, thus the number of variables).

$$\text{sim}(x, y) = \frac{\sum_{i=1}^l x_i y_i}{\sqrt{\sum_{i=1}^l x_i^2 \sum_{i=1}^l y_i^2}} \quad (4.7)$$

Another example of artificial data generation is given by Clifton et al. (2014). They develop a two-class counterpart for the one-class SVMs by generating data with the purpose of calibrating SVM output into probabilities. Their goal is different, but the methodology and their case study is relevant for this research. The monitoring of an industrial combustion engine was tackled by simulating data. The initial training phase of the engine was considered as the positive, stable data. Data for the negative class was generated by simulating unstable combustion through increasing fuel flow rates. This approach is thus based on data simulated by experiments, which would be infeasible when dealing with a human behaviour setting, such as fraud detection.

There is not one established, generally applicable method for novelty detection. This is largely due to the fact that specific settings require specific methods and a well-tailored method usually outperforms the more general method. The use of artificial data is a method that can be well-tailored and therefore supports this idea. There are however two important comments to be made on the use of artificial data for the negative class. Abe et al. (2006) and Hempstalk et al. (2008) notice that it is important that the artificial data is not too different from the positive data, since the risk exists that the classifier would simply learn to distinguish real from artificial examples. A second remark is made by Görnitz et al. (2013), who warn that using artificially created data for the unknown class may in certain cases be inappropriate, since totally new and unseen (negative) classes are not easily picked up with a two-class method that was not trained on such data. One-class methods are expected to outperform the two-class approaches in that respect. However, the benefits of the two-class approach will in many situations outweigh this possible downside. The argument of Görnitz et al. (2013) also suggests that one-class classifiers do not suffer from the drawback that new negative examples are not picked up. It should however be remarked that traditional one-class approaches are based on the implicit assumption that all examples of the positive class are present in the dataset. This assumption is too strong in most cases and the result is that new positive examples will be misclassified, resulting in a higher number

of false negatives (FN). A model that is able to learn from two classes will generally be able to make a better decision in these cases (Tax and Duin, 2001, 2004). Overcoming the risk that the two-class model, created on artificial data, only learns to distinguish between artificial and real data remains the most precarious issue. Our study proposes a method to overcome this remaining issue.

4.3 Method Development

4.3.1 Expert Knowledge

Models based on (partly) artificial data try to enhance the informativeness and therefore improve the predictive performance. The objective of this research is to further enhance the informativeness beyond what has been done in previous research, where artificial data was mainly used to create a decision boundary, without strong assumptions about the negative class. Ding et al. (2014) emphasize that the success of semi-supervised novelty detection is strongly dependent on the quality of the generated negative data. The method that we present takes into account the principle behind the guideline of Hempstalk et al. (2008); namely that it is important to prevent that the classifier only learns to distinguish between real and artificial examples. Our solution however differs because it does not require the artificial data to be close to the positive data. If it is known or expected that the real negative data differs on certain aspects from the real positive data, it would be suboptimal not to include this kind of information. The main advantage of using this information is that the boundary between the positive and the negative class will become supported from two sides rather than from one side as is the case with one-class approaches. From these findings, the most important condition for the artificial data arises; namely that the created instances should be as realistic as possible. They should be a good surrogate for actual negative data and prevent that the model only learns to distinguish between artificial and real data.

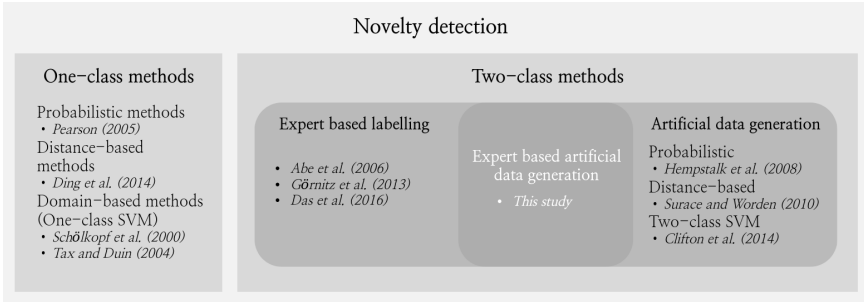


Figure 4.2: Positioning of Expert Data Generation Method in the Novelty Detection Literature.

As illustrated in Figure 4.2, this paper proposes the incorporation of expert knowledge in order to meet the condition of realism. Experts are qualified to come up with representative instances of negative data and to assess the realism of the synthetically created data. It has been suggested that prior (expert) knowledge about the case has the potential of tremendously increasing the performance of a classifier (Li et al., 2000; Larichev et al., 2002; Dayanik et al., 2006; Wang and Zhang, 2008; Lauer and Bloch, 2008; Utkin and Zhuk, 2014). Ashouri (1993) acknowledges that human reasoning enables identifying the structure of a problem and allows a qualitative analysis, but that handling quantitative, objective analysis is less obvious for human beings. The combination of expert knowledge with the predictive model incorporates the best of both worlds.

4.3.2 Expert Scenarios

The remaining question is how to implement the incorporation of expert knowledge in the creation of artificial data. Previous research has incorporated experts through a feedback loop (Abe et al., 2006; Görnitz et al., 2013), while others created synthetic data without incorporation of expert knowledge (e.g. Hempstalk et al. (2008)). Our method includes expert knowledge from the first stage, with the purpose of generating well informed data for the negative class. The expert knowledge defines one or more subspaces where the likelihood of observing negative data is higher. Semi-synthetic data will be constructed in these subspaces. An implementation that incorporates expert knowledge through the construction of scenarios is presented. Workshops with the relevant experts can be organised to come up with scenarios of abnormal or novel behaviour. These scenarios should be realistic and cover as much negative cases as possible.

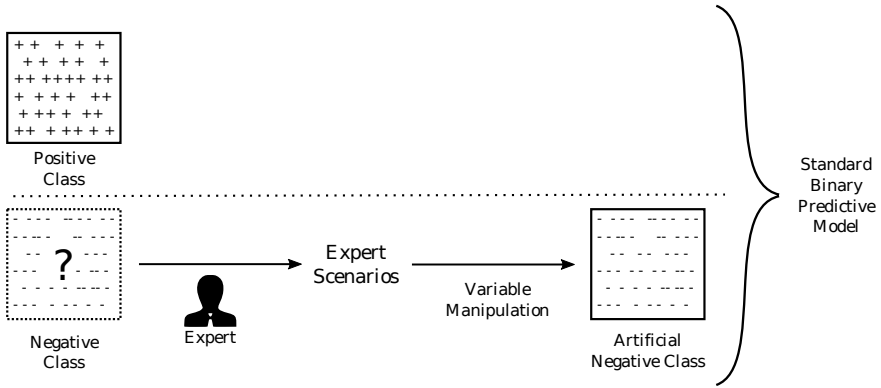


Figure 4.3: Expert Data Generation Method

Based on the selected scenarios, the next step involves generating actual data instances. By taking existing real datapoints as starting point, we aim for realistic new datapoints. The suggested approach can be called *variable manipulation*. Suppose that there are k variables in positive dataset P with p observations. The goal is to generate n instances for the negative class and store these in dataset N . We now proceed with the following steps:

1. Creation of instances of negative class - manipulated variables
 - (a) Based on the expert scenarios, select m variables to manipulate (with $m < k$)
 - (b) Set one or more expert rules for these m variables in order to generate artificial negative class examples. (e.g. variable x_k should have a value higher than 10 in scenario y).
2. Creation of instances of negative class - other variables
 - (a) Calculate the values for all other non-manipulated variables
 - (b) This results into dataset N
3. Stack dataset P (with the existing positive class examples) with set N
4. Build a predictive model, predicting $P(x_i \in P | x_{i,1} \dots x_{i,k})$.

This generic framework is meant to be applied to different cases. The structure and availability of the data will determine how to implement the

calculations in step 2a. The calculations for this step will be network based in our case study.

There is a trade-off between the number of fixed, manipulated variables and the free variables. The higher the number of manipulated variables is, the higher the influence of expert knowledge will be. The goal is to restrict certain variables and see how the other variables react on that. The ratio p/n determines the balancedness of the final dataset. Our method enables a free choice of this ratio. This also eliminates the typical novelty detection problem of extremely unbalanced data; a lot of positive examples, but none or very few negative examples are known. The scenario method therefore offers an alternative to oversampling techniques which are normally used to remedy the unbalancedness (Miguéis et al., 2017). The synthetic minority oversampling technique (SMOTE), as developed by Chawla et al. (2002), is a widely used example. This purely data based technique shares similarities with our method, since it also generates new semi-synthetic instances.

The proposed method also provides freedom of selecting a specific binary classification algorithm, as it only interferes at the data generating process, the modelling part continues as if this was a standard two-class classification problem. In order to clarify the methodology, the next section applies the method to a fraud detection case in the telecom sector.

4.4 Case Study: Telecom Subscription Fraud

4.4.1 Business Problem

A case study is used to demonstrate that incorporating expert knowledge into the data generation phase enhances predictive performance in a real-life setting. The goal of this business case was to detect customers that commit telecom subscription fraud. Hilaris and Mastorocostas (2008) define fraud detection as a field that uses techniques to monitor behaviour that deviates from the norm. This definition comes very close to the definition of novelty detection and it is therefore not surprising that novelty detection methods have been used for fraud detection (Pachia and Park, 2007; Jyothsna et al., 2011; Pimentel et al., 2014).

The problem of fraud is an important and worldwide issue in the telecommunications sector as it leads to an important loss in revenue (Farvaresh and Sepehri, 2011). Fawcett and Provost (1997) estimate that fraud costs the sector hundreds of millions of dollars per year. Hollmén and Tresp (1999) report that telecom companies lose between 2 and 5% of their total revenue to fraud. Fraud involves misuse, but it does not necessarily lead to direct legal consequences (Phua et al., 2010). Different types of telecom fraud have

been identified by Gosset and Hyland (1999), Hollmén and Tresp (1999), Hilas and Mastorocostas (2008) and Farvaresh and Sepehri (2011); such as contractual fraud (subscription and premium fraud), hacking fraud, technical fraud and procedural fraud. This case deals with *subscription fraud* (Gosset and Hyland, 1999; Farvaresh and Sepehri, 2011), a type where advantage can be made of the service by using the mobile offering in a way that is not allowed by general terms and conditions of the subscription.

The company wanted to launch a *family plan* (Desai et al., 2018), which includes up to five SIM cards for a fixed total price. These SIM cards are of the flat-rate use type, which means that they include unlimited SMS and calls. The general terms of this *family plan* allow the SIM cards within one subscription to be used only by people within the same household. However, there is a financial incentive to distribute these anonymous cards to people outside of the household. Since an extra card - up to five - comes at no extra cost, the incentive for fraudsters is very high.

Telecom fraud detection systems are usually based on anomaly detection, where behaviour is compared with past behaviour of subscribers (Yesuf et al., 2017). Van Vlasselaer et al. (2013) state that because of the many domain-specific characteristics of different fraud types, it becomes important to use a domain-specific solution. Our method provides the necessary flexibility due to the incorporation of expert knowledge. Fraud detection also leads to a class imbalance problem, since in most cases there are very few fraud cases compared to the total dataset. As became clear from the previous section, the class imbalance problem is eliminated as the number of generated negative examples can be set as desired.

4.4.2 CDR Data

Identifying fraudulent customers in this case comes down to predicting whether the relationship between two customers that subscribed in the same household truly is a household relationship. To tackle this prediction problem, a vast amount of call detail record (CDR) data is used. Eagle et al. (2009b) and Cho et al. (2011) confirm that CDR data has great potential to reveal relationships between people. Not only does the CDR data contain the calling behaviour between the individuals, it also contains their location. Geo-data has previously been used to infer social ties (Eagle et al., 2009b; Crandall et al., 2010; Cranshaw et al., 2010; Cho et al., 2011).

This research uses the CDR data of two five-week periods. The *pre-launch* data is used for the model building and a first evaluation on a test set. The *post-launch* data enables to perform an additional real-life validation test of the modelling approach.

- Week 1 - 5: Pre-launch data
- Week 9: Product launch
- Week 24 - 28: Post-launch data

All of the analyses are performed on *dyad* level, where a *dyad* consists out of two customers. The main goal is to predict whether the relationship between two customers is a household (positive) or a fraudulent (negative) relationship.

Based on this network data, 66 variables are created in cooperation with the experts of the company. These variables can be categorized as pure network-based (e.g. number of calls within dyad, number of common contacts), spatial / spatio-temporal (e.g. distance between most used locations), network-spatial (e.g. distance between both when calling each other) and variables related to the home address (as approximated by the closest phone tower). An overview of these variables can be found in appendix.

4.4.3 Incorporating Expert Knowledge

Transforming the one-class problem into a two-class problem requires labelled data for both the positive and negative class. Examples of positive dyads are available in the *pre-launch* dataset. Before the launch of the new product, many customers already signed up as a household in order to receive only one bill. Apart from this practical convenience, there was no (financial) incentive to dishonestly sign up as a household. We can therefore use these dyads as examples of positive class data (see also Figure 4.4).

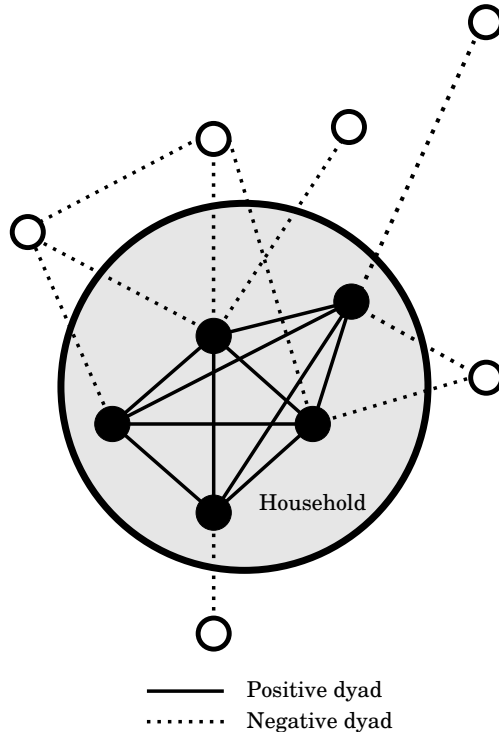


Figure 4.4: Network dyad selection; positive and negative class. Positive links connect individuals in the same household whereas negative links connect people that are not part of the same household.

Since there are no known fraud examples, instances have to be created for the negative class. One approach would be to use a random subset of all links between two individuals that are not within the same household. In other words, negative dyads are constructed by randomly combining two individuals. However, the major part of these instances would be dyads that are not at all related and therefore unrepresentative of the real negative class. This approach would also lead to negatives that are too different from the positives, conflicting with the guideline of Hempstalk et al. (2008). Moreover, complete random selection would also imply that there is no expert knowledge that can guide the selection process. Therefore restrictions are set in our method (see 4.3.2 Expert Scenarios: step 1b). Our method overcomes this issue and meets with the requirements of realistic scenarios.

The experts identified two major fraud scenarios (with each three sub-scenarios): distributing the extra free SIM cards between friends and be-

tween neighbours. Translating these scenarios into usable data is done by using rules for variable manipulation. Based on the information in the existing CDR dataset, thresholds were set for three variables that define the scenarios. The assumption for the *friend* scenario is for example that there is at least one contact between them during the five-week period. For the neighbour scenarios, people that live within a radius of 200 metres are selected. Only taking this distance into account would lead to the previous described problem that most of those neighbours are unrelated. Therefore, the same calling behaviour rules as in the friend scenarios are taken into account.

Scenario	Calls	SMS	Distance
friend	≥ 1	or ≥ 1	
good friend	≥ 10	or ≥ 10	
best friend	≥ 40	or ≥ 100	
neighbour	≥ 1	or ≥ 1	$\leq 200m$
good neighbour	≥ 10	or ≥ 10	$\leq 200m$
best neighbour	≥ 40	or ≥ 100	$\leq 200m$

Table 4.1: Expert Fraud Scenarios. *Or is non-exclusive.*

This data is used for the subsequent analyses (see Table 4.2). An unbalanced dataset with 95% data from the positive class was used. The 5% data for the negative class was sampled from the expert generated negative dataset. This unbalanced set was used since this more closely resembles the true class distribution as expected by our experts. Instances from the different scenarios are merged and labelled as the negative class. Putting all scenarios together into one negative class enables to have a well-sampled representation of real-life data, where all scenarios will also occur together in the data. We use 5-fold cross-validation for all employed models in order to obtain a robust evaluation of the results. The division of the data over the five folds is reported in Table 4.2.

In the following, our expert model will be benchmarked against pure one-class models and models that create artificial data without expert knowledge (cf. Figure 4.2).

4.4.4 Benchmark Model 1: One-Class Classification

We implement the three main categories of one-class classification, to benchmark our method. We follow the most important measures for novelty detection as identified by Ding et al. (2014) to evaluate our models; True Negative Rate (TNR) and AUC.

Scenario	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Total
household data (P)	18,059	18,072	18,090	18,048	18,061	90,330
fraud scenario data (N)	957	944	927	968	956	4,752
friend (N)	168	153	154	169	148	792
good friend (N)	158	154	161	155	164	792
best friend (N)	153	157	161	166	155	792
neighbour (N)	154	166	148	161	163	792
good neighbour (N)	173	151	158	156	154	792
best neighbour (N)	151	163	145	161	172	792

Table 4.2: Pre-launch Data. Number of instances (dyads) in the different folds for 5-fold cross validation.

The main interest lies in the true negative rate (TNR), also called specificity, since the goal is to detect as many of the actual fraudsters as possible. TNR is equivalent to the fraud detection rate (FDR), it calculates the percentages of fraud cases that are detected. The TNR or FDR is calculated as $TN/(TN + FP)$, with TN = True Negative and FP = False Positive. As is common practice in novelty detection, we define the stable (=household) class consistently as positive throughout this research. As opposed to traditional two-class classification, therefore the class that we want to predict as accurately as possible is defined as the negative class, which turns the intuitive interpretation of specificity and sensitivity upside down. AUC measures the *Area Under the receiver operating characteristic Curve*. AUC provides an assessment of the overall performance of the classification model and is not dependent on a chosen cut-off value. The measure can be interpreted as the probability that a randomly chosen positive observation is ranked higher than a randomly chosen negative observation. The value should thus be as high as possible and will be between 0.5 and 1, where 1 indicates a perfect model and 0.5 indicates a random model from which we can not learn anything. All reported measures are the average performance over the folds of the 5-fold cross-validation (5-fold cv) approach. The values per fold can be found in Appendix (Table A2).

4.4.4.1 Benchmark Model 1a: One-Class Probabilistic

We selected a parametric model for the probabilistic one-class benchmark. The multivariate normal distribution was fitted to the positive data in the training set. The resulting density was then used to score the test data. The 5% observations with the lowest density score were classified as the negative class.

The results are displayed in the confusion matrix in Table 4.3, together

with the other one-class benchmarks. All reported confusion matrices are summed over the different folds. The average FDR over the five folds is 7.64%. A random model would on average result into a FDR of 5% since, 5% of the test data is of the negative (fraud) class. This means that the one-class probabilistic model performs slightly better. The average AUC value over the five folds is 0.516. This also confirms that the model does slightly better than a random model.

4.4.4.2 Benchmark Model 1b: One-Class k-Nearest-Neighbours

Literature identified one-class k-NN as the best method in the distance based category. In two-class supervised learning, one can optimize the value for k . In one-class k-NN applications, this is not possible, because there is no ground truth. However, in our case, we can use the test data (see Table 4.2) to select a value for k . Based on the FDR, 1 was selected as k for four folds, $k = 3$ was selected for one fold (see Appendix). Furthermore, all variables were scaled before applying this distance based method.

The results of this benchmark model are again displayed in Table 4.3. The FDR amounts to 6.27%. The AUC for this model is 0.512.

4.4.4.3 Benchmark Model 1c: One-Class SVM

The one-class SVM (OCSVM) method of Schölkopf et al. (2000) is selected as benchmark for the third category of one-class models. Chang and Lin (2011) implemented the approach of Schölkopf et al. (2000) in the popular libSVM package. The interface to libSVM as provided in R was used (Meyer et al., 2017). The OCSVM was defined using the ν parameter. For two-class SVM, this parameter serves as an upper bound for the training error and a lower bound for the number of support vectors, whereas for OCSVM ν is an upper bound for the fraction of negative class data (Hornik et al., 2006). This way, ν can also be interpreted as the *novelty rate*. ν is set at 0.05 and the one-class model is trained on the positive household data using 5-fold cv as presented in Table 4.2.

FDR is 8.56% for the one-class SVM model. The AUC value can not be calculated because one-class SVM outputs only binary decisions without probabilities based on which the observations could be ranked

The performance of these three benchmarks is in the same range. We can observe that OCSVM (Benchmark 1c) scores best, followed by the probabilistic approach and one-class k-NN. Benchmark 1c has the best FDR and will therefore be selected as representative for the one-class approaches.

	Probabilistic (Benchmark 1a)			k-NN (Benchmark 1b)			One-Class SVM (Benchmark 1c)		
	Predicted Positive	Predicted Negative	%Novelty	Predicted Positive	Predicted Negative	%Novelty	Predicted Positive	Predicted Negative	%Novelty
household (P)	85, 938	4, 392	4.86	85, 873	4, 457	4.93	85, 687	4, 643	5.14
fraud (N)	4, 389	363	7.64	4, 454	298	6.27	4, 345	407	8.56
friend (N)	735	57	7.20	706	86	10.86	728	67	8.08
good friend (N)	748	44	5.56	738	54	6.82	740	52	6.57
best friend (N)	670	122	15.40	775	17	2.15	661	131	16.54
neighbour (N)	768	24	3.03	731	61	7.70	764	28	3.54
good neighbour (N)	770	22	2.78	739	53	6.69	770	22	2.78
best neighbour (N)	698	94	11.87	765	27	3.41	682	110	13.89

Table 4.3: Confusion Matrix One-Class Benchmark Models.(5-fold cv) %Novelty displays per class what percentage of the cases was predicted as novelty. For the fraud scenarios, %novelty equals the Fraud Detection Rate and can be interpreted as the percentage of actual fraud cases that are detected. For example, the FDR for the one-class SVM is thus 8.56%. For the positive, household class, %novelty equals 1 - True Positive Rate and can be interpreted as the percentage of cases incorrectly classified as fraud.

4.4.5 Benchmark Model 2: Two-Class Artificial Data Generation Models

As two-class artificial data generation can be seen as an intermediary step towards our expert data generation (see also Figure 4.2), we implement models from this category as a second class of benchmarks.

4.4.5.1 Benchmark Model 2a & 2b: Probabilistic Artificial Data Generation

We follow the approach of Hempstalk et al. (2008) and use equation (4.6) to implement two probabilistic models. Benchmark model 2a uses the multivariate normal distribution as reference distribution for the artificial class ($P(X|-)$), benchmark model 2b uses a uniform distribution (as suggested by Hempstalk et al. (2008)). Hempstalk et al. (2008) stress that $P(X|-)$ should be as close as possible to $P(X|+)$, we therefore use the parameters as estimated for $P(X|+)$ to generate data for the negative class. For the multivariate normal model (2a), this means that $P(X|-)$ is equal to the density estimated in benchmark model 1a. The boundaries of the uniform model are determined by the boundaries of the positive class.

Artificial data is generated from the respective distributions for the negative class. We then use SVM to train the classifier $P(+|X)$. The SVM implementation in R (Meyer et al., 2017) was set up to output class probabilities instead of only class labels, based on Platt et al. (1999). The used type of SVM is C-classification, with a radial basis function (RBF) kernel. All SVM models throughout this research (except for the one-class SVM) use these settings. As an intermediate result, we report the performance of $P(+|X)$ on the generated artificial dataset itself (Table 4.4).

$P(X -)$	AUC	FDR
Multivariate Normal (2a)	0.999	99.73
Uniform (2b)	1	100

Table 4.4: Average Performance (5-fold cv) $P(+|X)$ on Artificially Generated Data.

We observe perfect separation for model 2b and nearly perfect separation for 2a. This intermediate result suggests that the model only learns to distinguish between artificial and non-artificial data.

Combining $P(+|X)$ and $P(X|-)$ using equation (4.6) results in the final prediction (Table 4.5). In line with the novelty rate and the frequency of

artificial fraud dyads in the dataset, the dyads with the 5% lowest densities are classified as fraud.

	Multivariate Normal (Benchmark 2a)			Uniform (Benchmark 2b)		
	Predicted Positive	Predicted Negative	%Novelty	Predicted Positive	Predicted Negative	%Novelty
household (P)	85,944	4,386	4.86	86,017	4,313	4.77
fraud (N)	4,383	369	7.77	4,310	442	9.30
friend (N)	736	56	7.07	686	106	13.38
good friend (N)	747	45	5.68	722	70	8.84
best friend (N)	666	126	15.91	733	59	7.45
neighbour (N)	769	23	2.90	688	104	13.13
good neighbour (N)	771	21	2.65	741	51	6.44
best neighbour (N)	694	98	12.37	740	52	6.57

Table 4.5: Confusion Matrix Two-Class Probabilistic Artificial Benchmark Models (5-fold cv).

We observe that the FDR for the multivariate normal model (2a) is slightly higher than for the corresponding one-class probabilistic model (Benchmark 1a). Selecting a uniform reference distribution (2b) leads to a further improvement in FDR.

4.4.5.2 Benchmark Model 2c: Distance-based Artificial Data Generation

The approach of Surace and Worden (2010) was mimicked to generate artificial data for the negative class. In a first step, data was generated from a multivariate Gaussian distribution (cf. Benchmark models 1a and 2a). In a second step, the average cosine similarity of the generated data points - with respect to the positive set - was calculated. Only cases that had a lower average cosine similarity than the 5% lower boundary of internal cosine similarity in the positive set were retained as generated negative cases.

Now that data has been created for the negative class, a traditional support vector machine (SVM) is used for classification. Therefore, we will refer to this method as two-class artificial SVM as well. Again, the dyads that are in the lowest 5% probability of being household are classified as fraud.

The results are presented in Table 4.6. The extremely high FDR, together with an average AUC of 0.999 over the folds indicates that the model offers almost perfect separation. Similar to the intermediate results of benchmark models 2a and 2b, this raises the alarm that the model might just be learning to distinguish between real data (of the positive class) and artificial data (of the negative class). This important aspect will be dis-

cussed further in the following sections.

	Predicted Positive	Predicted Negative	%Novelty
household (P)	90, 245	85	0.09
fraud (N)	83	4, 672	98.25

Table 4.6: Two-Class Artificial Model (Benchmark 2c): Confusion Matrix on the Artificial Test Set (5-fold cv). Fraud Detection Rate = 98.25%.

Using this model on the expert test set results into a huge drop in performance (see Table 4.7). Nearly all dyads are predicted as household, less than 0.001% of dyads have a lower than 0.975 probability of being household. Using the same absolute cut-off value as in Table 4.6 would lead to nearly all dyads in the test set being classified as positive. However, in accordance with other benchmarks, we classify the dyads with the 5% lowest probabilities as fraud in the confusion matrix (Table 4.7).

Two-Class Artificial (Benchmark 2c)			
	Predicted Positive	Predicted Negative	%Novelty
household (P)	86, 068	4, 262	4.72
fraud (N)	4, 259	493	10.37
friend (N)	728	64	8.08
good friend (N)	736	56	7.07
best friend (N)	646	146	18.43
neighbour (N)	724	68	8.59
good neighbour (N)	741	51	6.44
best neighbour (N)	684	108	13.64

Table 4.7: Two-Class Artificial Model (Benchmark 2c): Confusion Matrix on the Expert Test Set (5-fold cv).

4.4.6 Two-Class Expert Model

The implementation of our two-class expert method uses the same expert based dataset as the other benchmarks. The crucial difference is that the two-class expert model does use the generated negative expert data for the

training of the model. To achieve maximal comparability with the benchmark models where a classifier was needed (1c, 2a, 2b and 2c), we again use SVM as binary classifier. This means that the difference in performance can be attributed solely to the two-class expert method and not to the difference in the background binary classifier. Furthermore, SVM has a strong theoretical foundation and excellent predictive performance (Lessmann and Voß, 2009). It also has a good generalisation performance when applied to noisy data. In this specific case there is a large heterogeneity in the behaviour of customers and a method that generalizes well is desirable.

The predictions on the test set are shown in Table 4.8. The FDR is now 48.72%. The model resulted in an average AUC of 0.824 over the folds. These values indicate decent performance, heavily improving upon the benchmark models. The two-class expert model also has the lowest number of households incorrectly classified as fraud, 2.70% compared to an average of 4.88% for the benchmark models. Keeping in mind that a random model would classify 5% of the households as fraud, this can be considered as a considerable improvement.

Two-Class Expert			
	Predicted Positive	Predicted Negative	%Novelty
household (P)	87,890	2,440	2.70
fraud (N)	2,437	2,315	48.72
friend (N)	225	567	71.59
good friend (N)	339	453	57.20
best friend (N)	440	352	44.44
neighbour (N)	333	459	57.95
good neighbour (N)	454	338	42.68
best neighbour (N)	646	146	18.43

Table 4.8: Two-class Expert SVM Confusion Matrix per Scenario on the Test Set (5-fold cv). Fraud Detection Rate = 48.72%.

When applying the two-class expert method, the user is however not restricted to SVM. We therefore also demonstrate the robustness of the method by applying six different classifiers (Table 4.9 and Figure 4.5). The results are robust over the six different classifiers, with SVM scoring average.

Model	AUC	FDR
SVM	0.824	48.72%
Logistic Regression	0.835	45.56%
AdaBoost	0.884	51.02%
Decision Tree	0.770	41.50%
Random Forest	0.865	51.38%
Neural Network (1-layer)	0.793	40.46%

Table 4.9: Robustness Check. Average performance (5-fold cv) of the Two-Class Expert Method for different binary classifiers on expert test data.

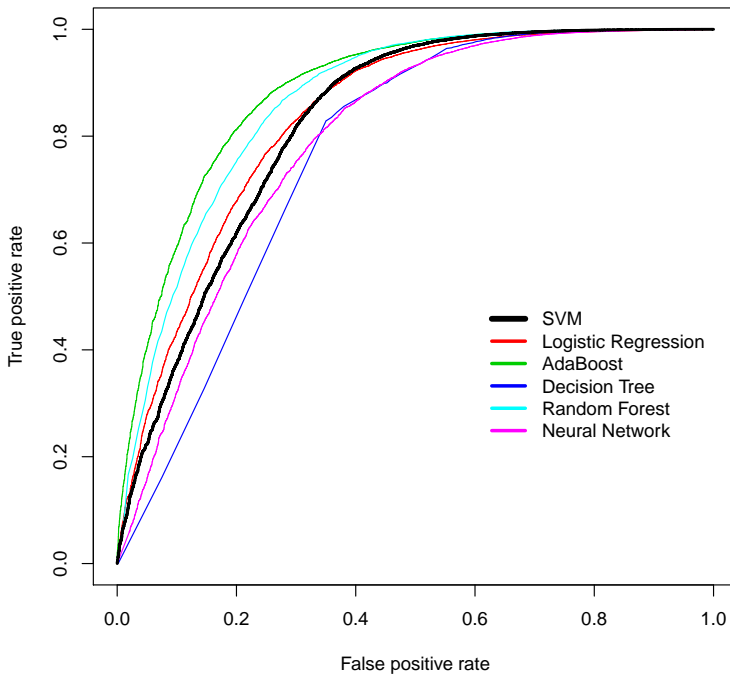


Figure 4.5: Robustness Check. ROC Curves of investigated binary classifiers for the Two-Class Expert Method on expert test data (5-fold cv).

A valuable novelty detection method should be generalisable and avoid overfitting. In other words, it should perform well on a new dataset that

was not involved in the generation of the novelty detection model. Therefore, we also evaluate the performance of the most important models on the dataset generated by the other method. As the pure one-class models do not generate data, only the data generated by the two-class models can be used for this test. For the pure one-class approaches, model 1c was selected as the best performing model. Furthermore, our main objective is to compare a two-class expert method over a one-class method and a two-class artificial method. Hence, keeping the background algorithm (SVM) equal makes it ideal for comparison. For this reason, we also select benchmark 2c as example of the two-class artificial data generation models. The results of this cross comparison are reported in Table 4.10.







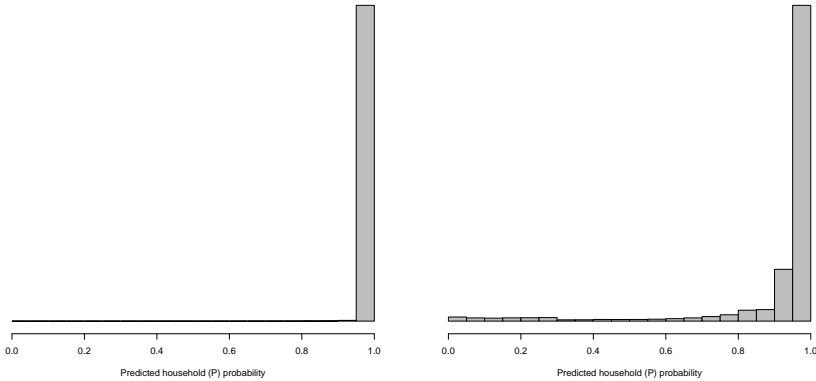
	Artificial Data	Expert Data
One-Class (Benchmark 1c)	 5.89	 8.56
Two-Class Artificial (Benchmark 2c)	 98.25	 10.37
Two-Class Expert	 15.01	 48.72

Table 4.10: *Fraud Detection Rate (FDR) of three novelty detection methods evaluated on two datasets. The artificial dataset contains the negative data as generated by the purely artificial approach (cf. Section 4.4.5). The expert dataset contains the negative data as generated by our expert data generation method (cf. Section 4.4.3). The models were trained on their respective dataset.*

The performance of the two-class artificial model drops tremendously when deployed on the other data set (87.88 percentage point drop). This indicates that the artificial negatives on which the model was trained are indeed too artificial and too different from the positives. It demonstrates the assumption that the artificial model only learns to distinguish between artificial and non-artificial cases. As expected, our two-class expert model also drops in performance when deployed on the other data set, however the drop is much smaller (33.71 p.p.). This illustrates that the expert model detects novelties in a more generalised way and thus not only performs well on the data on which it was trained. The expert model also performs better than the one-class SVM, even on the artificial data set on which it was not developed. These analyses show the value of expert based data generation for novelty detection. That is, the approach finds the required balance between generating novelties that are different enough from the positive data, while not being too distinct, so that classification algorithms do not overfit the artificial data.

4.4.7 Real-Life Post-Launch Implementation and Validation

In this and the next section we implement and validate the model on new real-life data. The two-class models are used to score all (100,000+) dyads of the *post-launch* data, i.e. data that contains potential fraudsters. Figure 4.6b displays the distribution of the predictions of the proposed expert model. As is typical for novelty detection and fraud problems, we observe a strong unbalancedness in the predicted probabilities. Nevertheless, the histogram is fairly dispersed when compared to the predictions of the artificial model (Figure 4.6a). The histogram of the artificial model shows that all cases are classified as positive, non-fraudulent cases. This again indicates that the artificial model actually classifies these cases merely as *real, non-artificial* cases, hence providing no information about fraudulent behaviour. Again, as in the previous section, the artificial model shows to only have learned to identify very specific artificial outliers rather than more general anomalies.



(a) Histogram of predictions of the Two-Class Artificial model (Benchmark 2c) on post-launch data. All cases are classified as positive. The model has no practical value.

(b) Histogram of predictions of the Two-Class Expert model on post-launch data. The predictions are as expected, the major part is classified as non-fraudulent. We clearly observe the characteristic unbalancedness of one-class and fraud problems. Cases in the left side of the histogram can be labelled as fraud suspects.

Figure 4.6: Comparison of Predictions on Post-Launch Dataset.

4.4.8 Manual Checks on Post-Launch Predictions

Now that the expert model predicts some dyads as fraud suspects, the company was involved to verify whether these cases are true fraudulent dyads. The sample of dyads with the lowest household probability scores was transferred to the company. A specialised fraud team that could make use of specific data sources, assessed whether a dyad was fraudulent. Due to the nature of this case and the fact that for many users there was limited additional information, it is impossible to assess all selected dyads. 478 dyads were checked in total in order to obtain 100 validly labelled dyads. Figure 4.7 shows the results for these 100 dyads. Dyads with a decision beyond reasonable doubt were labelled as *likely*. The pure fraud and household categories contain only cases where fraud could be identified by the specialised company fraud team. The results show that a very high proportion (about 90%) of these dyads indeed were considered to be fraudulent by the company fraud team. For the company this result was surprising as most of their fraud models (in different contexts though) suffer from much higher false positive rates.

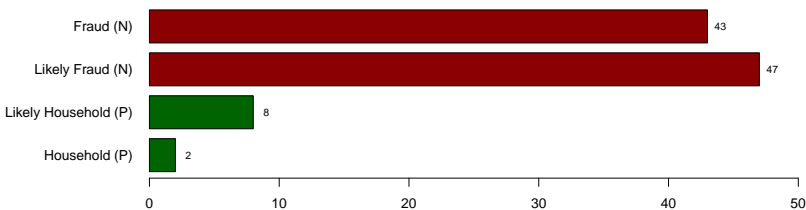


Figure 4.7: Results of manual labelling of predicted fraud suspects by fraud team of the company. The vast majority of these top predicted fraud dyads are indeed labelled as such in the real-life validation.

To illustrate an example, Figure 4.8 and 4.9 provide details on a fraud case that was identified by the expert model. The relation between the individuals depicted in orange and red was part of the top 100. In Figure 4.8 we see that the red individual is situated further in the network structure. In Figure 4.9 we observe the same in terms of location. The manual checks indeed identified this person as a fraudulent part of the household. The predictions for all dyads in this household, together with their actual label are presented in Table 4.11. The table also displays the predictions of representative benchmark models. For this example, it is clear that the latter provide virtually no information about fraud behaviour.

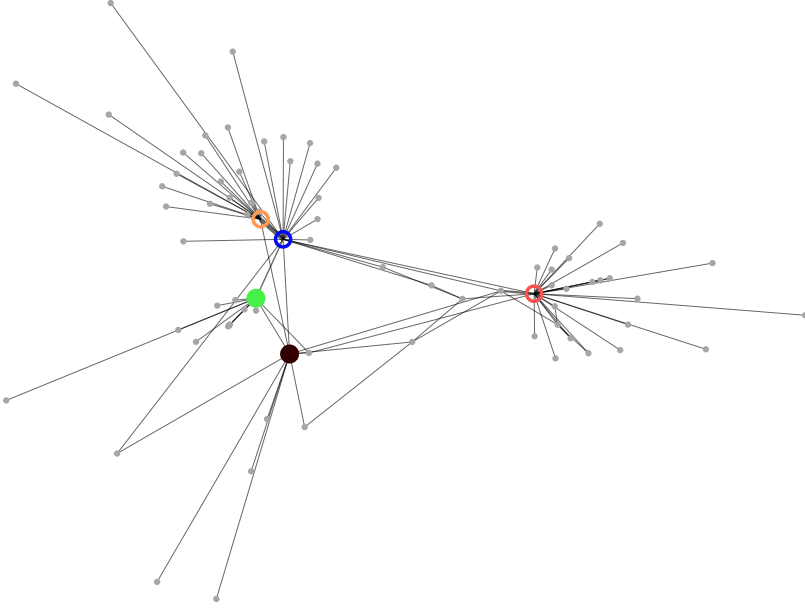


Figure 4.8: Identified Fraud Case: CDR Network Visualisation. Full circles represent individuals that are valid members of the household, hollow circles are fraudulent members. The lineweight of the links reflects the number of calls and SMS between two individuals within the dyad. People with stronger connections are also displayed closer to each other, as calculated by the ForceAtlas2 algorithm in the Gephi software (Bastian et al., 2009). Our model identified the relation between the red and the orange individual as fraudulent. We observe in this network that both individuals have no clear connection. There is very little overlap in their respective social networks as well. In Figure 4.9 we can draw the same conclusion based on location data.

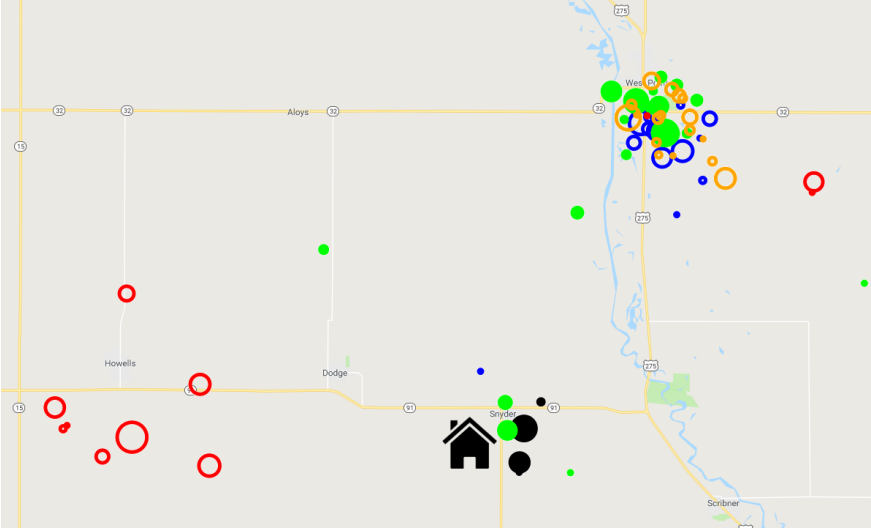


Figure 4.9: Identified Fraud Case: Location Plot. The size of the circles represents the number of calls/SMS on a certain location. The house indicates the home location of the household. The individuals are displayed with the same coding as Figure 4.8. The background map in this figure does not represent the true geographical location in order to preserve anonymity. We can observe that the red and orange individuals - who were identified as fraudulent by our model - have no location tags near the home location. The locations of both individuals also are very distinct from one another.

A	B	Manual Label	Two-Class Expert	Two-Class Artificial (Benchmark 2c)	One-Class (Benchmark 1c)
Orange	Red	Fraud (N)	0.207	1.000	P
Orange	Black	Fraud (N)	0.347	1.000	P
Green	Red	Fraud (N)	0.392	1.000	P
Blue	Red	Fraud (N)	0.407	0.999	P
Orange	Green	Fraud (N)	0.736	1.000	P
Blue	Green	Fraud (N)	0.759	1.000	P
Blue	Black	Fraud (N)	0.919	1.000	P
Black	Red	Fraud (N)	0.923	1.000	P
Green	Black	Likely Household (P)	0.960	0.993	P
Blue	Orange	Likely Household (P)	0.972	1.000	P

Table 4.11: Predictions of the different classes of models for a fraud example. For the expert and artificial model, these are the household (P) probability predictions. The one-class model outputs no predictions, but a binary decision, in this case positive for all dyads, hence all dyads are predicted as household.

4.5 Discussion

The two-class expert method outperformed both the one-class and two-class artificial benchmarks in this study. The latter techniques have nevertheless displayed acceptable results in previous research. An important distinction with the proposed method is that human behaviour is modelled, whereas traditionally applications in for example machine monitoring have been explored. This research demonstrated that the existing methods are not sufficient for the classification of human behaviour.

The one-class methods suffer from the obvious drawback that they can only learn from one class. The artificial two-class methods also failed to significantly boost performance. These artificial data generation approaches take two rather extreme forms, that are both not well suited for a human behaviour application. The first (cf. Benchmark models 2a and 2b) generates artificial data based on the distribution of the original positive class. Intuitively it is clear that we can not learn a lot about the actual negative class in such case. The results indicate that improvement upon the one-class models is indeed minimal. The other approach (cf. Benchmark model 2c) is extreme in the sense that it creates artificial data that is very distinct from the positive class. When dealing with human behaviour, the variability within the data becomes large, both for the positive and negative data, the overlap between both classes is larger than in non-human applications. The boundary between both classes becomes less strict and hence the boundary that will come out of such model will be too strict and too artificial. These artificial, non-expert, data generation methods thus use automated, unrealistic assumptions, whereas we proposed to incorporate well-informed expert based assumptions. This addition of extra information, in the form of expert knowledge, adds strongly to the classification power.

Taking a closer look at the confusion matrix of the expert model (Table 4.8), we observe that the more restricted the scenarios become, the lower the FDR becomes. This can be explained by the fact that in this case, the more strict scenarios for the negative class more closely resemble the positive (household) class, which makes it more difficult for a model to distinguish both classes. Nevertheless, it is important to include these scenarios, because according to the experts these scenarios more closely resemble realistic cases. What the model learns from these cases is likely the most important for the actual detection of fraud cases in the real-life validation.

Even though it remains to be explored how much the expert can add to the more traditional machine monitoring cases, the presented method promises to be well suited to tackle these and other novelty detection prob-

lems as well, due to its flexible nature. The expert scenario method can flexibly introduce scenarios that are not in the original dataset. Hence, creating semi-synthetic data has the benefit of providing data that is well tailored to specific requirements. Furthermore, creating expert data is much cheaper when compared to manually labelling data.

This research is an addition to and not an argument against the traditional one-class approaches for novelty detection. Pure one-class approaches are sometimes considered to be better at identifying complete novel cases. Therefore, the selection of the appropriate methodology will depend on the misclassification cost of these cases. However, when implementing the expert-based two-class methodology, it is important to invest a fair amount of time in the construction of the scenarios, so that all relevant scenarios are represented in the negative data. Furthermore, the two-class expert method is able to detect cases that are not explicitly modelled in one of the scenarios. The classification algorithm detects underlying, shared characteristics between the scenarios that are also shared with novel cases. In the case study, other types of subscription fraud that were not explicitly modelled, were detected. An example is the use of the extra SIM cards by older children of the household that already moved out. Furthermore, as discussed before, most one-class methods assume that the positive class is perfectly represented in the positive dataset. However, this is unlikely to be true in many cases, which leads to a higher number of false negatives.

The major limitation of the presented research is that the expert method is validated in a single case study. To enlarge the validity of this method, future research needs to explore how the method translates to other cases and contexts. Another remaining issue is the trade-off between the number of manipulated variables and the free variables. The higher the amount of expert knowledge, the higher the number of manipulated variables. In general, a higher level of expert knowledge will shrink the space in which data for the unknown class is generated. These better defined regions however come at the cost of possibly losing the generality that characterizes traditional one-class approaches. Further research is needed to examine what the impact of using different levels of expert knowledge could be.

4.6 Conclusion

The transformation of a one-class problem into a two-class problem was examined. This method was assessed in the context of fraud detection for a new telecom service. The absence of labelled fraud examples calls for the use of one-class novelty detection methods. However, traditional one-class methods perform poorly in a case dealing with human behaviour. Hence, a

new method is developed to deal with this issue. Using semi-synthetic data for the negative class has great potential. Previous research used artificial data with the same purpose of better defining a boundary around the positive class, but without clear assumptions about the negative class. We introduced the incorporation of expert knowledge in order to use clear assumptions. This enhances the informativeness of the artificial data and further improves the classification performance. Experts build realistic, representative scenarios that describe the behaviour of the humans belonging to the negative class. Using these scenarios, instances were generated for the negative class with variable manipulation. The method was tested in a real-life telecom subscription fraud case. The two-class expert method clearly outperformed the conventional one-class benchmark models. The method also improved upon the artificial two-class non-expert benchmarks, that were characterized by the problem of creating models that merely learned to distinguish between artificial and non-artificial cases. The performance of the model was also examined in a manual validation phase for a new post product launch dataset. The model performed very well in this real-life setting and was able to detect real fraud cases with a model build on expert fraud scenarios. Including expert knowledge strongly helped to classify the diverse human behaviour data, where less flexible traditional methods failed. The manual checks are costly in terms of manpower and hence a predictive model that prioritises, can generate a lot of value.

4.7 Appendix

4.7.1 Benchmark Model 1b: Optimal Value for k

k	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
1	6.51%	6.40%	4.38%	5.95%	6.68%
2	6.19%	5.88%	5.12%	5.62%	6.39%
5	4.59%	5.46%	5.76%	4.96%	5.80%
10	5.02%	4.82%	5.34%	4.74%	5.70%
20	4.91%	4.72%	4.38%	4.74%	5.31%
50	3.52%	4.83%	4.38%	4.41%	4.62%
100	3.31%	4.41%	4.16%	2.98%	3.64%

Table 4.12: Fraud Detection Rate for different k -values (where k is the number of nearest neighbours). The optimal value $k = 1$ is used for all folds except fold 3, where $k = 5$ will be used.

4.7.2 5-Fold Cross-Validation: Performance Measures of the Different Folds

Model	FDR					AUC				
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Benchmark 1a	7,79	7,87	7,58	7,83	7,17	0,518	0,528	0,509	0,517	0,508
Benchmark 1b	6,51	6,40	5,76	5,95	6,68	0,510	0,505	0,505	0,534	0,506
Benchmark 1c	9,07	9,13	8,22	8,38	8,06	-	-	-	-	-
Benchmark 2a	7,68	7,87	8,11	7,94	7,27	0,526	0,540	0,516	0,526	0,518
Benchmark 2b	10,57	9,97	9,71	8,27	8,06	0,648	0,644	0,644	0,621	0,604
Benchmark 2c	9,61	11,65	10,78	10,36	9,53	0,597	0,610	0,617	0,616	0,603
Two-Class Expert	49,84	48,58	50,16	50,06	45,29	0,825	0,826	0,832	0,816	0,821

Table 4.13: 5-Fold Cross-Validation: Performance Measures of the Different Folds.

4.7.3 Overview of Variables

The following naming convention is used:

- **n_**: ‘Pure network variables’. For these variables only data about calls and sms is used, thus no location data. (For examples see tables below). This category can further be refined into two subcategories:
 - Variables that concern the direct relation between the two ids in the dyad.
 - Variables that concern common contacts.
- **s_**: ‘Pure spatial variables’. These variables only use location.
- **ns_**: ‘Combi variables’. These variables are a combination of calling/sms and spatial information.
- **home_**: ‘Home variables’. All these variables relate to the home address of the customer.

n.calls_nbr	number of calls to each other (sum of both directions)
n.calls_dur_total	total duration of calls to each other (sum of both directions)
n.calls_dur_avg	average call duration of calls to each other (over calls in both directions)
n.calls_dur_sd	standard deviation of duration of calls to each other (both directions taken together)

n.calls_ratio	ratio of number of calls (lowest number divided by highest number) (as a measure of equality/directionality of the relationship)
n.calls_perc_of_calls_max	percentage of calls of id x that are to the other person (this is the highest number of both percentages)
n.calls_perc_of_calls_min	percentage of calls of id x that are to the other person (this is the lowest number of both percentages)
n.sms_nbr_sent	number of sms sent to each other (sum of both directions)
n.sms_ratio_sent	ratio of number of sms sent (lowest number divided by highest number)(as a measure of equality/directionality of the relationship)
n.sms_perc_of_sms_max	percentage of sms of id x that are to the other person (this is the highest number of both percentages)
n.sms_perc_of_sms_min	percentage of sms of id x that are to the other person (this is the lowest number of both percentages)
n.common_contacts_nbr	absolute number of common contacts
n.common_contacts_perc_max	percentage of contacts that are common: #common contacts / #contacts of person with least contacts of both
n.common_contacts_perc_min	percentage of contacts that are common: #common contacts / #contacts of person with most contacts of both
n.com_cont_calls_nbr	total number of calls to/from common contacts
n.com_cont_calls_total_duration	total duration of calls to/from common contacts
n.com_cont_calls_avg_duration	average duration of calls to/from common contacts
n.com_cont_sms_nbr	number of sms to common contacts
s.distance_most_used_loc	distance between the most used locations in general for the dyad
s.distance_most_used_loc_workday	same as s.distance_most_used_loc, but only locations registered during working hours
s.distance_most_used_loc_evening	same as s.distance_most_used_loc, but only locations registered during morning, evening, night; thus when people are expected to be at home

s_distance_most_used_loc_weekend	same as s_distance_most_used_loc, but only locations registered during week-ends
s_overlap_abs_number_of_locations	overlap in locations
s_overlap_loc_perc_max	ratio of absolute overlap to number of distinct locations for one person in the dyad (highest number of both options)
s_overlap_loc_perc_min	ratio of absolute overlap to number of distinct locations for one person in the dyad (lowest number of both options)
s_overlap_abs_number_workday	same as s_overlap_abs_number_of_locations, but on subsample of locations: namely only those during the workday
s_overlap_loc_perc_workday_max	ratio cfr s_overlap_loc_perc_max
s_overlap_loc_perc_workday_min	ratio cfr s_overlap_loc_perc_min
s_overlap_abs_number_evening	same as s_overlap_abs_number_of_locations, but on subsample of locations: namely only those during the evening
s_overlap_loc_perc_evening_max	ratio cfr s_overlap_loc_perc_max
s_overlap_loc_perc_evening_min	ratio cfr s_overlap_loc_perc_min
s_overlap_abs_number_weekend	same as s_overlap_abs_number_of_locations, but on subsample of locations: namely only those during the weekend
s_overlap_loc_perc_weekend_max	ratio cfr s_overlap_loc_perc_max
s_overlap_loc_perc_weekend_min	ratio cfr s_overlap_loc_perc_min
ns_sms_dist_avg	average distance when texting each other
ns_sms_dist_max	maximum
ns_sms_dist_min	minimum
ns_sms_dist_sd	standard deviation
ns_call_dist_avg	average distance when calling each other
ns_call_dist_max	maximum
ns_call_dist_min	minimum
ns_call_dist_sd	standard deviation
ns_calls_same_location	number of calls to each other when on same location
ns_sms_same_location	number of sms to each other when on same location
ns_calls_same_location_workday	calls idem location workday
ns_sms_same_location_workday	sms idem location workday
ns_calls_same_location_evening	calls idem location evening
ns_sms_same_location_evening	sms idem location evening

ns_calls_same_location_weekend	calls idem location weekend
ns_sms_same_location_weekend	sms idem location weekend
home_nbr_calls_max	number of outgoing calls on the home location (largest number of the dyad)
home_nbr_calls_min	idem (but smallest number in the dyad)
home_perc_calls_max	percentage of all outgoing calls that were on the home location (largest number in the dyad)
home_perc_calls_min	idem (but smallest number in the dyad)
home_nbr_sms_max	number of outgoing sms on the home location (largest number in the dyad)
home_nbr_sms_min	idem (but smallest number in the dyad)
home_perc_sms_max	percentage of all outgoing sms that were on the home location (largest number in the dyad)
home_perc_sms_min	idem (but smallest number in the dyad)
home_distance_most_used_loc_max	distance between the overall most used location of ida/b to the home tower (largest number in the dyad)
home_distance_most_used_loc_min	idem (but smallest number in the dyad)
home_distance_most_used_loc_evening_max	idem, but only locations during evening, night and morning
home_distance_most_used_loc_evening_min	idem (but smallest number in the dyad)
home_distance_most_used_loc_weekend_max	idem but only location during weekend
home_distance_most_used_loc_weekend_min	idem (but smallest number in the dyad)
home_distance_most_used_loc_workday_max	idem, but only locations during working hours
home_distance_most_used_loc_workday_min	idem (but smallest number in the dyad)

Table 4.14: Overview of Variables

5

Conclusion

The main goal of this dissertation was to investigate how mobile phone data could be used to model spatio-temporal human behaviour. Raw mobile phone data as such is not sufficient to derive insights. Therefore, a data analytical approach was adopted in order to create value from the raw data. Three different applications were tested in order to assess the power of mobile phone data and the accompanying models. This dissertation offers by no means a full picture of modelling human behaviour with mobile phone data. Nevertheless, several contributions have been made, both in terms of the use of Bluetooth and call detail record (CDR) data as sources for the analysis, as well in terms of methodological contribution. A concise overview of the main contributions, findings and practical applications can be found in Table 5.1. Note that even though in every chapter a specific application was selected, the methodologies itself and the reported value of the data is not limited to the selected application. In this concluding chapter, the structure and the foundations of this dissertation are recapitulated. The chapter starts with a discussion section, that allows to discuss certain aspects that need to be stressed further, or that have not been covered sufficiently in the main chapters. Next, the most important findings are summarised and some theoretical and practical implications are discussed, followed by limitations of the presented studies and approaches. Based on the latter, avenues for future research are provided as well.

5.1 Discussion

5.1.1 The Single Best Method

In 2020, a worldwide user base of 3.5 billion smartphone users and 4.2 billion active smartphones were reported (Newzoo, 2020), where active smartphones are defined as smartphones used at least once a month. The same report estimates that by 2023 there will be 4.1 billion smartphone users globally, taking into account the estimated constant annual growth rate of 6.2%. A large part of this growth will be driven by emerging regions. Although smartphones constitute the main part, they are still only a part of the total mobile phone market. The current number of mobile phones (including smartphones) is estimated at 4.9 billion, accounting for 63.6% of the global population (Turner, 2020). Hence, the already discussed advantages of mobile phones as a proxy for tracking humans are supplemented with a steady growth of the number of people that can be tracked. This means that even larger samples can be attained, that the approach offers a promising future perspective and that the growth in emerging countries enables worldwide application of the approach. Mobile phones offer tremendous potential. Smartphones offer even more potential, as the installation of dedicated applications enables capturing data with more context, that leads to even deeper, more insightful information. However, a significant group of people only uses the primary functions of their mobile phone or smartphone; making phone calls and sending text messages (Gadziński, 2018). Installing an application requires more effort and knowledge that this group may not want to do or does not possess. Although this problem does not affect CDR data (and to a lesser extent Bluetooth), the richness of the captured data is less than with a smartphone application based approach. This dissertation mainly favoured the use of CDR and Bluetooth data, partly due to their advantages in this respect. However, this dissertation should by no means be interpreted as an argument against the use of these more dedicated apps. It is important to recognise that both approaches have their right of existence. Depending on the desired depth of the results, the desired sample size and possible specifics of a certain tracking case, a choice should be made for one, or a combination of, the approaches.

The same reasoning holds for the use of surveys. Recall that in the introduction, mobile phone based tracking was advanced as a promising alternative for the traditional survey based approach. Many authors agree that this methodology may never completely replace surveys (Gong et al., 2014; Vij and Shankari, 2015; Geurs et al., 2015). Although this disser-

tation mainly stressed the many advantages of actual tracking methods, it goes along with the argument that surveys can be used in addition.

Furthermore, the ground truth for home location in Chapter 3 was defined as the home tower closest to the actual billing address of the customer. Certain people may have more than one home location or their actual home location might differ from the address where they receive their bills. However, this definition is accurate in the majority of the cases and has the advantage that it leads to a ground truth for every individual in the dataset. To make sure that we have the actual (or multiple) home location(s) of each individual in the dataset, one could use a survey based approach, where people are asked about their actual home location (e.g. Ahas et al. (2010a)). That approach however has the downside that obtaining a very large sample becomes very time consuming and costly again. Depending on the goal of the research, one of the methods can be chosen. Moreover, it is definitely possible to implement both methods along each other.

Chapter 4 introduced a two-class expert method in order to transform a one-class into a two-class problem. It was revealed that this did indeed significantly enhance the predictive performance in a case study, dealing with human behaviour. However, it is again necessary to add nuance to the blindfolded use of this approach. The two-class method can fail to identify cases that were not added in the expert scenario creation phase. Even though, the results indicated that certain cases that were not explicitly modelled were still detected, this does not mean that it will be the case for every novelty (or new type of fraud in the selected application). The one-class methods for novelty detection might outperform two-class approaches if there are a large variety of novelties that can not be modelled in advance by domain experts. However, the one-class paradigm is based on a stable normal, non-novel, class. If there is no such class, for example when dealing with human behaviour that is by default varied in nature, this assumption will not hold. Human behaviour can in essence be quite stable since people show many fixed daily routines, usually have one fixed home place and a rather limited amount of other anchor locations (Ahas et al., 2009). Because of this, human behaviour can even be considered rather stable and therefore also predictable (Gonzalez et al., 2008; Song et al., 2010). Nevertheless, this stability refers to inter human stability rather than intra human stability. The behaviour of different people still differs a lot, which is exactly what leads to the so called instability in this case, thereby violating the usual assumptions for the normal or stable class. This aspect needs to be taken into account when considering the choice between a one-class or

two-class approach for novelty detection. Moreover, it might be interesting to investigate how both frameworks could be applied in the sense of ensemble learning (Polikar, 2012; Coussement and De Bock, 2013; Sagi and Rokach, 2018). This way, the final performance could be further improved by combining the strengths of both methods. This provides another interesting choice besides the one- and two-class methods.

In conclusion; the single best method that suits all cases does not exist.

5.1.2 Privacy and GDPR

In 2006, Netflix launched an open competition to improve their collaborative filtering algorithm used to predict user ratings for movies. This Netflix Prize contest led to the disclosure of insufficiently anonymised data about almost 500,000 Netflix customers (Singel, 2009; McSherry and Mironov, 2009). Researchers were able to re-identify users (Narayanan and Shmatikov, 2006). This strongly violated their privacy. The incident led to a payment of 9 million dollars to settle the lawsuit. This example clearly demonstrates the importance of privacy when dealing with data. Other cases such as the Cambridge Analytica case further raised awareness about this sensitive issue. These events led to the introduction of a General Data Protection Regulation (GDPR) by the European Commission (European Commission, 2020). When using data, privacy is always important. When using data captured from tracking human beings, individual personal information is collected and privacy becomes even more important.

The focus of this dissertation has been on the value of tracking data and adequate analytical methods. Although a full legal discussion of privacy concerns is therefore out of scope, it remains a highly relevant topic to discuss. Certain researchers state that respecting the privacy of users is the most essential responsibility when using a system that tracks mobile phones (Lane et al., 2010; Cottrill et al., 2013).

Anonymity was preserved in Chapter 2 by recording only the Media Access Control (MAC) address of the tracked devices. This MAC address allows for individual registration of each device, but can by no means be linked to the identity of the tracked individual. The MAC address can therefore be considered safe to use. Bluetooth devices can also broadcast a ‘friendly name’, a user defined name for the device. As this name is user defined, it can possess information about the true identity of the individual, as some people even use their full name. It should therefore at all times be avoided to record this extra information. Moreover, this ‘friendly name’ does not add any value to the data.

The CDR datasets in Chapter 3 and 4 were anonymised by means of hashing (Knuth, 1973). This means that the identities and actual mobile phone numbers are not used at all for the research purposes.

The GDPR regulation is very strict for personal data. However, anonymous data is not treated as personal data, which makes that no user consent and specific protection is required (Naujokaite, 2018). Nevertheless, as shown by the Netflix Prize example, it remains very difficult to genuinely ensure that the anonymous data is absolutely anonymous. Anonymity should guarantee that it is by no means possible to identify any actual person or identity.

Furthermore, there is a difference between *anonymous* and *confidential* data. Anonymous data is recorded in such a way that it can never be linked to the true identity. As discussed above, by using the MAC address, Bluetooth tracking data can therefore be considered as an example of the strict category of anonymous data. Confidential data on the other hand typically uses a unique identifier that enables linking the anonymised data and the actual identity. This causes users' privacy to be at greater risk and leads to the requirement of protection measures. The CDR data sets fall under this category. The data has been used in this dissertation in a completely anonymous way, as the table that included the hashing information obviously has not been shared. It is strongly advised that the identities are stored separately from the data and the code. The link between the hashed identifiers and the actual identities can only be made by the telecom provider that cooperated in this research.

As a last remark, the interest in this dissertation does not lie in the spatio-temporal behaviour of a single individual. Neither is this the case in typical research, nor in typical practical applications. Although the fact that these analyses start from the individual level, raw data, the ultimate goal is always to elaborate analyses on an aggregated level, leading to insights on this aggregated level. This means that privacy is preserved in the resulting, aggregated outcomes.

5.2 Conclusion and Implications

Recall from the structure of the general introduction that this dissertation aimed at three aspects; (1) investigating data sources for tracking spatio-temporal human behaviour, (2) exploring analytical methods to translate the raw data into meaningful insights and (3) examining this in three different applications. In terms of data sources, Bluetooth tracking and call de-

Contributions		Main Findings		Practical Applications	
Ch.2	Bluetooth Tracking of Humans in an Indoor Environment: An application to Shopping Mall Visits	<ul style="list-style-type: none">– Demonstration of applicability of Bluetooth tracking in an indoor setting.– Value of Bluetooth tracking in a marketing context: update for survey based approach.	<ul style="list-style-type: none">– High data quality achieved, that enabled a variety of analyses.– Low cost of data collection.– Mobile phones and smartphone account for 89% of the registered devices.– A detection ratio of 9,81% was found.	<ul style="list-style-type: none">– Bluetooth tracking can be used as a tool for tracking customers.– Bluetooth tracking can be applied in a large variety of indoor applications, e.g. to improve store lay-out and security.	
Ch.3	Home Location Prediction with Telecom Data: Benchmarking Heuristics with a Predictive Modelling Approach	<ul style="list-style-type: none">– Benchmarking study of existing heuristics used in literature.– Benchmarking study based on strongly improved validation data.– Development of new heuristic method for home detection with CDR data.– Introduction of labelled predictive modelling approach for home detection with CDR data.	<ul style="list-style-type: none">– Labelled predictive modelling outperforms heuristic approaches.– Social network based variables improve the predictive performance.– Heuristic methods achieve 60% accuracy and an average distance error of 4,4 kilometres.– Labelled predictive modelling (including social) achieves 72% accuracy and an average distance error of 2,8 kilometres.	<ul style="list-style-type: none">– Spatio-temporal research based on CDR data can start with a validated approach for the crucial first step: the home location.– Telecom providers can use the insights for the identification of subscription fraud.	
Ch.4	From One-Class to Two-Class Classification by Incorporating Expert Knowledge: Novelty Detection in Human Behaviour	<ul style="list-style-type: none">– New approach for novelty detection.– New methodology that uses human expert knowledge for the conversion of the one-class problem into a two-class problem.– Evidence is provided for the use of CDR data in a telecom subscription fraud setting.	<ul style="list-style-type: none">– Transformation of one-class model into two-class model boosts performance.– Human expert knowledge boosts performance.– Fraud detection rate (FDR) of two-class expert method respectively 15,01% (artificial data set) and 48,72% (expert data set), compared to only 5,89% (artificial data set) and 8,56% (expert data set) for the optimal one-class benchmark.– Successful real-life test on fraud detection application (90% labelled as (likely) fraud).	<ul style="list-style-type: none">– A broad range of one-class / novelty detection problems can be translated into two-class problems, which increases predictive performance.– Telecom providers can use the insights for the identification of subscription fraud.	

Table 5.1: Overview of main contributions and findings of this dissertation.

tail record (CDR) data have been investigated and proven to be successful. Bluetooth tracking is advised in an indoor setting, with limited geographical range and high precision. Using CDR data enabled analyses on a much larger geographical scale, where indoor and outdoor applications are possible. However this comes at the cost of a much lower precision, as the precision is related to cell phone towers, that cover a rather large geographical range. On a methodological level, it has been shown that Bluetooth tracking data is straightforward to use and results into a variety of metrics and visualisations. Methodologically, the use of CDR data led to the introduction of a new heuristic for home detection, a labelled predictive modelling approach for the traditionally unlabelled home detection problem and the development of a new expert based method for novelty detection. Bluetooth tracking has been shown to be successful in a indoor marketing application. CDR data has proven its effectiveness in the application of home location detection and in a telecom fraud application.

5.2.1 Benefits of Bluetooth Tracking and Call Detail Records

The research has demonstrated the applicability of both data sources in the selected applications. Two crucial benefits that characterise both data sources are their low cost and their non-participatory aspect.

Due to their labour intensive nature, traditional survey based approaches for collecting data about human spatio-temporal behaviour are characterised by a high cost per respondent. Furthermore, due to the fixed cost per respondent, scaling into larger sample sizes does not lead to cost benefits either. Bluetooth tracking on the other hand incurs a set-up cost, relative to the number of Bluetooth scanners that need to be installed. The initial investment in the scanners and tuning the range of each scanner manually are the main cost drivers. The period of data collection and the number of individuals being tracked does not increase costs. Therefore, the method can be seen as a sound long-term investment. Costs are even more limited when using CDR data for the analyses. As telecom providers capture the CDR data anyway, because they need this data for billing purposes, no extra cost of data collection is incurred. Note however that telecom providers start to realise the potential value of their data and start selling (aggregated) data. Nevertheless, alike Bluetooth data, upscaling the sample is again cost-insensitive.

Furthermore, the fraud detection application in Chapter 4 also leads to cost improvements as a result of the developed methodology. Manual fraud checks are labour intensive and therefore very costly. However, the method

was able to identify fraudulent cases, which provides the fraud department with a prioritised list. This makes the manual fraud checks much more effective. The expert process for generating data for the novelty class is also a much cheaper methodology than manually labelling fraud cases (especially because of the sparse nature of fraud).

A second decisive benefit of both data sources has proven to be their non-participatory nature. The passive data collection method implies that people are largely unaware of being tracked, which enables measurement of their actual, unbiased behaviour. This crucial aspect transformed the possibly biased ‘respondents’ into unbiased tracked individuals. Furthermore, it has been shown in this dissertation that Bluetooth tracking is especially useful in *uncontrollable* settings. These uncontrollable settings have been defined as settings where individuals can freely move in space, without any obvious means of identification (such as a built-in RFID tag in a festival wristband or in a shopping cart for example). A large variety of public places, such as train stations, libraries, shopping malls, and museums can be defined as uncontrollable. The non-participatory aspect can raise privacy concerns as people did not explicitly give permission for being tracked. However, as discussed before, if anonymity can be guaranteed, this should cause no such issues.

5.2.2 Methodological Contributions

The focus in this PhD dissertation gradually shifted from more data oriented towards more methodological over the chapters. It is therefore not surprising that the most important methodological contribution is the two-class expert model, developed in Chapter 4. However, a similar problem was already tackled in Chapter 3 as well. When introducing a labelled predictive modelling approach as an alternative for the unlabelled heuristic methods, the problem of having clearly labelled data for only one class was faced as well. Selecting data for the class of home towers was straightforward due to the ground truth home towers in the data. On the other hand, every tower that was not a home tower could in theory be used as an example of the non-home tower class. Nevertheless, this would have resulted into a model that only learned to distinguish between used (home) towers and not used (non-home) towers as an individual only uses a small subset of all available cell phone towers. A more informative choice needed to be made. This was executed by selecting all used (non-home) towers as examples for the non-home class. This basic idea was further refined in Chapter 4, where human experts were introduced to develop expert scenarios for the creation

or selection of data for the second class. The occurrence of this issue in both chapters, indicates that the problem is not infrequent, nor trivial. It is therefore beneficial that methods have been developed in this dissertation to tackle this issue. In addition, it has been shown that modelling human behaviour with the support of a human expert proves to be a good match.

The human aspect did not only become relevant with the two-class expert methodology. The human input has obviously also been instrumental in the variable creation for the different predictive models. Raw data without input of a data analyst remains unproductive.

5.2.3 Numerical Results of Applications and Implications

Chapter 2 examined the applicability of Bluetooth for customer tracking in a shopping mall. It was found that 89% of the registered devices were mobile phones (including smartphones). Therefore, the approach is able to indeed capture mobile phones, which were suggested as the optimal proxy for tracking human spatio-temporal behaviour. The 11% other devices do not lead to practical issues, as they can easily be filtered out due to the device type information in the MAC address. Furthermore, a detection ratio of 9.81% was found. This number is definitely large enough to quickly generate a large sample size. These numbers indicate that Bluetooth tracking can be considered a beneficial method for a variety of other applications as well. Not only tracking customers, but tracking visitors at various events is possible. The results can be used to improve store lay-out and improve the security at events or in large indoor buildings for example. The methodology is not limited to the passive tracking of individuals, but can for example also be used in the analysis of fire drills.

Literature revealed the importance of home detection for spatio-temporal analyses with CDR data. However, validation of the home detection methods was almost non-existent. Chapter 3 of this dissertation validated the existing methods, that are heuristic approaches. These were able to identify the correct home location in 60% of the cases. The average distance error was found to be 4.4 kilometres. The chapter introduced a labelled predictive modelling approach, that increased these results to 72% and 2.8 kilometres. Although it was shown that this approach should be preferred, it is in many cases not possible to use a labelled procedure. In these cases, the new heuristic method is advised. Furthermore, models that included social network information performed slightly better than the models without. Spatio-temporal research can now start from a more solid base for the often crucial first step of identifying the home location. Telecom providers can

use these results in cases, such as identifying subscription fraud, the topic of the final investigated application in this dissertation.

The subscription fraud application in Chapter 4 led to the development of a new methodology for novelty detection in general. The human expert based two-class model increased the fraud detection rate (FDR) of the optimal one-class benchmark method from 5.89% to 15.01% on an artificial data set and from 8.56% to 48.72% in an expert data set. The approach was further validated in a real-life test, that examined the top predicted fraud cases. 90% of this selected set could be labelled as fraud (43%) or likely fraud (47%). A broad range of one-class novelty detection problems could benefit from this positive results. Nevertheless, it needs to be examined how generalisable the method is in other contexts, which brings us to the last section of this dissertation.

5.3 Limitations and Future Research

An important limitation of the implemented Bluetooth tracking set-up in Chapter 2 is its unautomated set-up, that hinders the full potential of this method. Every point of interest needs a manual set-up of a Bluetooth scanner. In the investigated case this means that this set-up was required for every individual store. The range needs to be fine tuned in order to avoid false positives: registrations for people that visited a neighbouring store for example. Although this practical problem causes a burden at the start of each installation, this is merely a start-up cost. However, the unautomated set-up also refers to the fact that the scanners in this research were not connected to each other, nor to a network. Data storage happens at each individual scanner. The data needed to be read from every scanner separately. It is of course strongly advised to use a connected approach as this not only facilitates easier data collection, but also gives the possibility to identify possible problems a lot quicker.

Sensors (e.g. Bluetooth sensors) might also capture data insufficiently. In the practical shopping mall setting, it is possible that people that did visit a store (with a Bluetooth device in visible mode), were not captured after all. These false negatives can not be counted. Further experimental research is needed to assess this aspect. On the other hand, false positives (e.g. customer of shop 2, registered in shop 1 is a false positive for shop 1) will occur as well and are also not possible to identify as false positive afterwards. Furthermore, factors such as whether a phone is in a pocket or not may affect the registration. Certain sensors may need more time to cap-

ture a device, which makes that for applications where people only quickly pass by sensors, this aspect needs to be examined. The CDR data does not have this limitation, as every call and SMS will be captured. However, this also means that the location can only be registered if there has been CDR activity. The data is therefore being generated in a non-continuous way.

The type I and type II errors when detecting mobile phones with Bluetooth tracking are related to the rather low positional accuracy of Bluetooth, a technology that was originally not designed with location purposes in mind. Nevertheless, recently a new Bluetooth standard (Bluetooth 5.1) has been launched. Bluetooth 5.1 supports radio direction finding (RDF), which means that it can now achieve an effective accuracy of less than one metre (Suryavanshi et al., 2019). Further research is needed to examine how well this works in practical applications.

Whether a Bluetooth device is being tracked is also dependent on the visibility setting. This detection ratio was found to be 9.81%, which is similar to ratios found in previous research (O'Neill et al., 2006; Hagemann and Weinzerl, 2008; Versichele et al., 2012a,b, 2014b). However, it is unclear how this detection ratio will evolve over time. Abedi et al. (2015) reported a higher detection ratio for Wi-Fi tracking when compared to Bluetooth tracking. However, with the introduction of Android version Q, mobile phones automatically generate different MAC addresses for every Wi-Fi network, which strongly reduces the value of Wi-Fi for tracking purposes. This does not affect visitor counts for example, but constructing paths and trajectories becomes impossible. These privacy measures taken in order to limit tracking might unfortunately strongly affect this type of research in the future.

The same question about the detection ratio can be raised for CDR data. Will standard CDR data, generated by standard call and text behaviour, remain the best data source, or will the enhanced CDR data or even mobile apps take over? Enhanced CDR data (as introduced in Table 1.1) does add observations to the standard CDR data by also including mobile data sessions and pings. However, future research will need to investigate the ultimate impact on the data quality. When people make more and more use of their mobile data connection, their amount of traditional calls and SMS might strongly diminish. Does the benefit of having more observations outweigh this adverse affect? The social network information that can be derived from standard CDR data might therefore be reduced, which would favour the use of mobile phone apps to study the social aspect in human spatio-temporal behaviour. Human behaviour changes over time and so

will the data that is being generated by their spatio-temporal behaviour. Methods and the preferred way of tracking are therefore far from static.

Bluetooth tracking might also introduce a sample bias as some people are more likely than others to enable Bluetooth on their mobile phone. The same holds for CDR data and mobile phone use in general. Certain groups might be difficult to track with mobile phones. This problem grows bigger with research that requires smartphones. In the research of Gadziński (2018) for example, only 42% of the people in the 65+ age group owned a smartphone. This makes that the results might be biased towards younger people. However, it might be equally difficult to reach the same groups with a survey approach.

This dissertation investigated three specific applications. As holds for all application based research, the question remains whether the results and the applied methods will hold in other settings as well. The home detection application in Chapter 3 might be most sensitive to this aspect. The results need further validation in other datasets in other countries. Furthermore, the applications focussed on rather short term analyses (almost 3 weeks of Bluetooth tracking data, 5 weeks of CDR data for home detection and two 5-week periods of CDR data for the fraud application). It is advised to investigate how longer periods of data will affect the results.

In the chapters using CDR data, the value of including social network information was touched upon. Compared to Bluetooth tracking data, standard CDR data has the advantage of the embedded social information. However, it remains to be investigated to what extent social can play a role in Bluetooth research as well. The Coronalert app discussed in the general introduction already hints at the social potential of Bluetooth, as Bluetooth is being used to assess the social contact of individuals. As discussed before, enhanced CDR data might have a detrimental effect on the social data in standard CDR. Nevertheless, the ubiquity of mobile phone data connections spirals the use of social smartphone applications. Therefore the data source that needs to be considered might change, but it likely will not completely prevent social network research. It can be concluded that both in the cases of Bluetooth and CDR, more devoted research in the field of social networks can be of interest.

Chapter 2 was positioned mainly as a proof of concept and the main goal was to investigate the value of Bluetooth tracking in an indoor setting. Therefore, the analytical aspect in that chapter has been limited to some fundamental insights. Nevertheless, it is important to stress again that the Bluetooth tracking data enables sequence analysis. It was mentioned in

Chapter 2 that clustering customers based on their behaviour leads to different insights when using methods that take the sequence of the visited stores in account. Especially interesting in this respect are sequence alignment methods (SAM). These methods have been used in prior marketing research to predict the next-product-to-buy (NPTB models) since standard classification methods are typically unable to effectively capture sequential patterns (Prinzie and Van den Poel, 2006, 2007). Sequence alignment has also been used already to effectively analyse sequences originating from Bluetooth tracking data (Delafontaine et al., 2012). Further research and applications can explore the many possible avenues for the analysis of (Bluetooth) tracking data.

An interesting avenue for future research constitutes the idea of ensemble learning (as mentioned in Section 5.1.1) for the one and two-class approach. Both methods separately have advantages and disadvantages. The implicit assumption that all positive cases are modelled in the one-class model might lead to new positive cases falsely being classified as negative. On the other hand, if a new negative case was not defined by a scenario in the two-class expert method, it might be that this case will not be classified as negative either. The two-class method still has the advantage that it can learn from two classes in order to correctly classify a new case. Nevertheless, it might be beneficial to combine the two-class model with predictions made by one-class models in order to end up with a more comprehensive approach. A similar idea can be implemented for Bluetooth and Wi-Fi tracking. Whereas until now both methods are mostly used separately, a combination of both into a hybrid approach makes perfect sense (Kao et al., 2017).

The focus of this dissertation lies on the applications, data sources and the methodologies. However the relative importance of (data) privacy is much greater than the percentage of text actually devoted to this aspect. This dissertation has demonstrated that is for example possible to achieve a reasonable estimate of people's home location based on anonymous CDR data. This raises the question how anonymous anonymous data really is. Every human being is unique and generates a unique trail of location data. Hence, if no measures - such as aggregation of the results - are taken, it becomes difficult to ensure the privacy of each individual. Research has already indicated that anonymously sharing your location might nonetheless give away your identity (Pyrgelis et al., 2018). It is clear that further research into the anonymity of tracking data is needed. Is the use of anonymous identifiers (such as the MAC address and hashed phone numbers)

actually sufficient? There is a growing need for new and more encompassing methods to actually safeguard people's privacy in many applications. Furthermore, changing privacy legislation further pushes the need for this type of methods. Developing new methods and standards to ensure privacy is urgent in order to prevent undermining spatio-temporal research.

Traditional mobile phones are being (or have been) replaced by smartphones throughout the world. These smartphones include a large set of embedded sensors, such as an accelerometer, proximity sensor, digital compass, gyroscope, GPS, microphone and camera. All these sensors possess the potential to further spiral the ongoing and growing research about human spatio-temporal behaviour. This PhD dissertation definitely does not mark the end of this research.

Bibliography

- Abe, N., Zadrozny, B., Langford, J., 2006. Outlier detection by active learning, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 504–509.
- Abedi, N., Bhaskar, A., Chung, E., 2014. Tracking spatio-temporal movement of human in terms of space utilization using media-access-control address data. *Applied Geography* 51, 72 – 81.
- Abedi, N., Bhaskar, A., Chung, E., Miska, M., 2015. Assessment of antenna characteristic effects on pedestrian and cyclists travel-time estimation based on bluetooth and wifi MAC addresses. *Transportation Research Part C: Emerging Technologies* 60, 124 – 141.
- Abualigah, L.M., Khader, A.T., Hanandeh, E.S., 2018a. Hybrid clustering analysis using improved krill herd algorithm. *Applied Intelligence* 48, 4047–4071.
- Abualigah, L.M., Khader, A.T., Hanandeh, E.S., 2018b. A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *Journal of Computational Science* 25, 456–466.
- Abualigah, L.M.Q., 2019. Feature selection and enhanced krill herd algorithm for text document clustering. Springer.
- Agostaro, F., Collura, F., Genco, F., Sorce, S., 2004. Problems and solutions in setting up a low cost bluetooth positioning system. *WSEAS Transactions on Computers* 3, 1102 – 1106.
- Ahas, R., Aasa, A., Silm, S., Tiru, M., 2010a. Daily rhythms of suburban commuters' movements in the tallinn metropolitan area: Case study with mobile positioning data. *Transportation Research Part C: Emerging Technologies* 18, 45–54.

- Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M., 2010b. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of urban technology* 17, 3–27.
- Ahas, R., Silm, S., Saluveer, E., Järv, O., 2009. Modelling home and work locations of populations using passive mobile positioning data, in: *Location based services and TeleCartography II*. Springer, pp. 301–315.
- Al-Habaibeh, A., Parkin, R., 2003. An autonomous low-cost infrared system for the on-line monitoring of manufacturing processes using novelty detection. *The International Journal of Advanced Manufacturing Technology* 22, 249–258.
- Alpaydin, E., 1999. Combined 5×2 cv f test for comparing supervised classification learning algorithms. *Neural computation* 11, 1885–1892.
- Anastasi, G., Bandelloni, R., Conti, M., Delmastro, F., Gregori, E., Mainetto, G., 2003. Experimenting an indoor bluetooth-based positioning service, in: *Distributed Computing Systems Workshops, 2003. Proceedings. 23rd International Conference on, IEEE*. pp. 480–483.
- Andres, L., 2012. *Designing and doing survey research*. Sage.
- Apple, 2016a. ibeacon for developers - apple developer. URL: <https://developer.apple.com/ibeacon>.
- Apple, 2016b. Maps for developers - apple developer. URL: <https://developer.apple.com/maps>.
- Ashouri, F., 1993. An expert system for predicting gas demand: A case study. *Omega* 21, 307–317.
- Axhausen, K.W., 2005. Social networks and travel: Some hypotheses. *Social dimensions of sustainable transport: transatlantic perspectives*, 90–108.
- Axhausen, K.W., Zimmermann, A., Schönfelder, S., Rindsfuser, G., Haupt, T., 2002. Observing the rhythms of daily life: A six-week travel diary. *Transportation* 29, 95–124.
- Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., Puchinger, J., 2019. Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies* 101, 254–275.

- Backstrom, L., Sun, E., Marlow, C., 2010. Find me if you can: improving geographical prediction with social and spatial proximity, in: *Proceedings of the 19th international conference on World wide web*, ACM. pp. 61–70.
- Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J., 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society* 54, 627–635.
- Baesens, B., Van Vlasselaer, V., Verbeke, W., 2015. *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. John Wiley & Sons.
- Barbosa, H., Barthelemy, M., Ghoshal, G., James, C.R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J.J., Simini, F., Tomasini, M., 2018. Human mobility: Models and applications. *Physics Reports* 734, 1–74.
- Barse, E.L., Kvarnstrom, H., Jonsson, E., 2003. Synthesizing test data for fraud detection systems, in: *Computer Security Applications Conference, 2003. Proceedings. 19th Annual*, IEEE. pp. 384–394.
- Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi: An open source software for exploring and manipulating networks. URL: <http://www.aaii.org/ocs/index.php/ICWSM/09/paper/view/154>.
- Benoit, D.F., Van den Poel, D., 2012. Improving customer retention in financial services using kinship network information. *Expert Systems with Applications* 39, 11435–11442.
- Bensky, A., 2007. *Wireless positioning technologies and applications*. Artech House.
- Bharti, K.K., Singh, P.K., 2015. Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Systems with Applications* 42, 3105–3114.
- Bhaskar, A., Kieu, L.M., Qu, M., Nantes, A., Miska, M., Chung, E., 2013. On the use of bluetooth mac scanners for live reporting of the transport network, in: *10th International Conference of Eastern Asia Society for Transportation Studies*, Taipei, Taiwan.
- Blondel, V.D., Decuyper, A., Krings, G., 2015. A survey of results on mobile phone datasets analysis. *EPJ data science* 4, 10.

- Blondel, V.D., Esch, M., Chan, C., Clérot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., Ziemlicki, C., 2012. Data for development: the d4d challenge on mobile phone data. arXiv preprint arXiv:1210.0137 .
- Bluenion, 2016. Indoor positioning and tracking system - bluenion. URL: <http://www.bluenion.com/solutions.php?id=8>.
- Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S., Ratti, C., 2015. Choosing the right home location definition method for the given dataset, in: International Conference on Social Informatics, Springer. pp. 194–208.
- Bonne, B., Barzan, A., Quax, P., Lamotte, W., 2013. Wifipi: Involuntary tracking of visitors at mass events, in: World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2013 IEEE 14th International Symposium and Workshops on a, IEEE. pp. 1–6.
- Breiman, L., 2001. Random forests. Machine learning 45, 5–32.
- Brockmann, D., Hufnagel, L., Geisel, T., 2006. The scaling laws of human travel. Nature 439, 462–465.
- Bullock, D., Haseman, R., Wasson, J., Spitler, R., 2010. Automated measurement of wait times at airport security. Transportation Research Record: Journal of the Transportation Research Board 2177, 60 – 68.
- Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C., 2011. Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Computing 10, 36–44. doi:doi: 10.1109/MPRV.2011.41.
- Carino, J.A., Delgado-Prieto, M., Zurita, D., Millan, M., Redondo, J.A.O., Romero-Troncoso, R., 2016. Enhanced industrial machinery condition monitoring methodology based on novelty detection and multi-modal analysis. IEEE access 4, 7594–7604.
- Carrasco, J.A., Miller, E.J., 2006. Exploring the propensity to perform social activities: a social network approach. Transportation 33, 463–480.
- Celikkan, U., Somun, G., Kutuk, U., Gamzeli, I., Cinar, E., Atici, I., 2011. Capturing supermarket shopper behavior using smartbasket, in: Snasel, V., Platos, J., El-Qawasmeh, E. (Eds.), Digital Information Processing and Communications. Springer Berlin Heidelberg. volume 189 of *Communications in Computer and Information Science*, pp. 44–53.

- Chang, C.C., Lin, C.J., 2011. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 27.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al., 2000. *Crisp-dm 1.0: Step-by-step data mining guide*. SPSS inc 9, 13.
- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16, 321–357.
- Chen, J., Liu, Y., Zou, M., 2014a. From tie strength to function: Home location estimation in social network, in: *2014 IEEE Computers, Communications and IT Applications Conference*, IEEE. pp. 67–71.
- Chen, X., Wang, X., Xu, Y., 2014b. Performance enhancement for a gps vector-tracking loop utilizing an adaptive iterated extended kalman filter. *Sensors* 14, 23630–23649.
- Cheung, K.C., Intille, S.S., Larson, K., 2006. An inexpensive bluetooth-based indoor positioning hack, in: *Proceedings of UbiComp 2006 Extended Abstracts*.
- Cho, E., Myers, S.A., Leskovec, J., 2011. Friendship and mobility: user movement in location-based social networks, in: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM. pp. 1082–1090.
- Clifton, L., Clifton, D.A., Watkinson, P.J., Tarassenko, L., 2011. Identification of patient deterioration in vital-sign data using one-class support vector machines, in: *2011 federated conference on computer science and information systems (FedCSIS)*, IEEE. pp. 125–131.
- Clifton, L., Clifton, D.A., Zhang, Y., Watkinson, P., Tarassenko, L., Yin, H., 2014. Probabilistic novelty detection with support vector machines. *IEEE Transactions on Reliability* 63, 455–467.
- Cottrill, C.D., Pereira, F.C., Zhao, F., Dias, I.F., Lim, H.B., Ben-Akiva, M.E., Zegras, P.C., 2013. Future mobility survey: Experience in developing a smartphone-based travel survey in singapore. *Transportation Research Record* 2354, 59–67.

- Coussement, K., De Bock, K.W., 2013. Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of Business Research* 66, 1629–1636.
- Crandall, D.J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., Kleinberg, J., 2010. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences* 107, 22436–22441.
- Cranshaw, J., Toch, E., Hong, J., Kittur, A., Sadeh, N., 2010. Bridging the gap between physical location and online social networks, in: *Proceedings of the 12th ACM international conference on Ubiquitous computing*, ACM. pp. 119–128.
- Crosscan, 2016. Real-time data solutions for retail - crosscan gmbh. URL: <http://crosscan.com/en>.
- Danalet, A., Farooq, B., Bierlaire, M., 2014. A bayesian approach to detect pedestrian destination-sequences from wifi signatures. *Transportation Research Part C: Emerging Technologies* 44, 146–170.
- Das, S., Wong, W.K., Dietterich, T., Fern, A., Emmott, A., 2016. Incorporating expert feedback into active anomaly discovery, in: *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, IEEE. pp. 853–858.
- Dash, M., Nguyen, H.L., Hong, C., Yap, G.E., Nguyen, M.N., Li, X., Krishnaswamy, S.P., Decraene, J., Antonatos, S., Wang, Y., et al., 2014. Home and work place prediction for urban planning using mobile network data, in: *2014 IEEE 15th International Conference on Mobile Data Management*, IEEE. pp. 37–42.
- Dayanik, A., Lewis, D.D., Madigan, D., Menkov, V., Genkin, A., 2006. Constructing informative prior distributions from domain knowledge in text classification, in: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM. pp. 493–500.
- Delafontaine, M., Versichele, M., Neutens, T., Van de Weghe, N., 2012. Analysing spatiotemporal sequences in bluetooth tracking data. *Applied Geography* 34, 659–668.
- Desai, P., Purohit, D., Zhou, B., 2018. Allowing consumers to bundle themselves: The profitability of family plans. *Marketing Science* .

- Ding, X., Li, Y., Belatreche, A., Maguire, L.P., 2014. An experimental evaluation of novelty detection methods. *Neurocomputing* 135, 313–327.
- Dugundji, E.R., Walker, J.L., 2005. Discrete choice with social and spatial network interdependencies: an empirical example using mixed generalized extreme value models with field and panel effects. *Transportation Research Record* 1921, 70–78.
- Eagle, N., Pentland, A.S., Lazer, D., 2009a. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences* 106, 15274–15278.
- Eagle, N., Pentland, A.S., Lazer, D., 2009b. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences* 106, 15274–15278.
- El-Yaniv, R., Nisenson, M., 2007. Optimal single-class classification strategies, in: *Advances in Neural Information Processing Systems*, pp. 377–384.
- European Commission, 2020. Data protection in the eu. URL: https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en. accessed: 2021-01-14.
- Farley, J., Ring, L., 1966. A stochastic model of supermarket traffic flow. *Annals of Operations Research* 14, 555 – 567.
- Farvaresh, H., Sepehri, M.M., 2011. A data mining framework for detecting subscription fraud in telecommunication. *Engineering Applications of Artificial Intelligence* 24, 182–194.
- Fawcett, T., Provost, F., 1997. Adaptive fraud detection. *Data mining and knowledge discovery* 1, 291–316.
- Feldmann, S., Kyamakya, K., Zapater, A., Lue, Z., 2003. An indoor bluetooth-based positioning system: Concept, implementation and experimental evaluation., in: *International Conference on Wireless Networks*, pp. 109–113.
- Fernandes, T., 2011. Indoor localization using bluetooth, in: *6th Doctoral Symposium in Informatics Engineering*, pp. 480–483.

- Junqué de Fortuny, E., Martens, D., Provost, F., 2013. Predictive modeling with big data: Is bigger really better? *Big Data* 1, 215–226.
- Friedman, J., Hastie, T., Tibshirani, R., et al., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics* 28, 337–407.
- Fujino, T., Kitazawa, M., Yamada, T., Takahashi, M., Yamamoto, G., Yoshikawa, A., Terano, T., 2014. Analyzing in-store shopping paths from indirect observation with RFID-tags communication data. *Journal on Innovation and Sustainability. RISUS ISSN 2179-3565* 5, 88–96.
- Gadziński, J., 2018. Perspectives of the use of smartphones in travel behaviour studies: Findings from a literature review and a pilot study. *Transportation Research Part C: Emerging Technologies* 88, 74–86.
- Geurs, K.T., Thomas, T., Bijlsma, M., Douhou, S., 2015. Automatic trip and mode detection with move smarter: First results from the dutch mobile mobility panel. *Transportation research procedia* 11, 247–262.
- Giannotti, F., Pedreschi, D., 2008. *Mobility, Data Mining and Privacy: A Vision of Convergence*. Springer.
- Gong, L., Morikawa, T., Yamamoto, T., Sato, H., 2014. Deriving personal trip data from gps data: A literature review on the existing methodologies. *Procedia-Social and Behavioral Sciences* 138, 557–565.
- Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L., 2008. Understanding individual human mobility patterns. *Nature* 453, 779.
- Google, 2016. Indoor maps - about - google maps. URL: <https://www.google.com/maps/about/partners/indoormaps>.
- Görnitz, N., Kloft, M.M., Rieck, K., Brefeld, U., 2013. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research*.
- Gosset, P., Hyland, M., 1999. Classification, detection and prosecution of fraud in mobile networks. *Proceedings of ACTS mobile summit, Sorrento, Italy*.
- Grimm, P., 2010. Social desirability bias. *Wiley international encyclopedia of marketing*.

- Gu, Y., Lo, A., Niemegeers, I., 2009. A survey of indoor positioning systems for wireless personal networks. *Communications Surveys Tutorials*, IEEE 11, 13–32.
- Hagemann, W., Weinzerl, J., 2008. Automatische erfassung von umsteigern per bluetooth-technologie. *Nahverkerspraxis*. Springer, Heidelberg .
- Hallberg, J., Nilsson, M., Synnes, K., 2003. Positioning with bluetooth, in: *Telecommunications, 2003. ICT 2003. 10th International Conference on*, IEEE. pp. 954–958.
- Hartigan, J.A., 1975. *Clustering algorithms* john wiley & sons. Inc., New York, NY .
- Harwood, M., 2009. *CompTIA network+ N10-001*. Pearson Education.
- Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C., 2014. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41, 260–271.
- Hay, S., Harle, R., 2009. Bluetooth tracking without discoverability, in: *Location and context awareness*. Springer, pp. 120–137.
- Helen, M., Latvala, J., Ikonen, H., Niittylahti, J., 2001. Using calibration in RSSI-based location tracking system, in: *Proceedings of the 5th world multiconference on circuits, systems, communications and computer*.
- Hempstalk, K., Frank, E., Witten, I.H., 2008. One-class classification by combining density and class probability estimation, in: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer. pp. 505–519.
- Hilas, C.S., Mastorocostas, P.A., 2008. An application of supervised and unsupervised learning approaches to telecommunications fraud detection. *Knowledge-Based Systems* 21, 721–726.
- Hironaka, S., Yoshida, M., Umemura, K., 2016. Analysis of home location estimation with iteration on twitter following relationship, in: *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, IEEE. pp. 1–5.
- Hodge, V., Austin, J., 2004. A survey of outlier detection methodologies. *Artificial intelligence review* 22, 85–126.

- Hollmén, J., Tresp, V., 1999. Call-based fraud detection in mobile communication networks using a hierarchical regime-switching model. *Advances in Neural Information Processing Systems* , 889–895.
- Hornik, K., Meyer, D., Karatzoglou, A., 2006. Support vector machines in r. *Journal of statistical software* 15, 1–28.
- Hurjui, C., Graur, A., Turcu, C., 2008. Monitoring the shopping activities from the supermarkets based on the intelligent basket by using the RFID technology. *Electronics and Electrical Engineering* 83, 7 – 10.
- Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies* 40, 63–74.
- Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A., 2011. Identifying important places in people’s lives from cellular network data, in: *International Conference on Pervasive Computing*, Springer. pp. 133–151.
- Japkowicz, N., Myers, C., Gluck, M., et al., 1995. A novelty detection approach to classification, in: *IJCAI*, pp. 518–523.
- Jung, I.C., Kwon, Y.S., 2011. Grocery customer behavior analysis using RFID-based shopping paths data. *World Academy of Science, Engineering and Technology* 59, 2011.
- Jyothisna, V., Prasad, V.R., Prasad, K.M., 2011. A review of anomaly based intrusion detection systems. *International Journal of Computer Applications* 28, 26–35.
- Kanda, T., Shiomi, M., Perrin, L., Nomura, T., Ishiguro, H., Hagita, N., 2007. Analysis of people trajectories with ubiquitous sensors in a science museum, in: *Robotics and Automation, 2007 IEEE International Conference on*, IEEE. pp. 4846–4853.
- Kaneko, Y., Yada, K., 2016. Fractal dimension of shopping path: influence on purchase behavior in a supermarket. *Procedia Computer Science* 96, 1764–1771.
- Kao, C.H., Hsiao, R.S., Chen, T.X., Chen, P.S., Pan, M.J., 2017. A hybrid indoor positioning for asset tracking using bluetooth low energy and wi-fi, in: *2017 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, IEEE. pp. 63–64.

- Karikoski, J., Soikkeli, T., 2013. Contextual usage patterns in smartphone communication services. *Personal and ubiquitous computing* 17, 491–502.
- Khan, S.S., Madden, M.G., 2014. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review* 29, 345–374.
- Kholod, M., Nakahara, T., Azuma, H., Yada, K., 2010. The influence of shopping path length on purchase behavior in grocery store, in: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Springer. pp. 273–280.
- Knuth, D.E., 1973. *Sorting and searching* .
- Krings, G., Calabrese, F., Ratti, C., Blondel, V.D., 2009. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment* 2009, L07003.
- Kung, K.S., Greco, K., Sobolevsky, S., Ratti, C., 2014. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one* 9, e96180.
- Lambiotte, R., Blondel, V.D., De Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., Van Dooren, P., 2008. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications* 387, 5317–5325.
- Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T., 2010. A survey of mobile phone sensing. *IEEE Communications magazine* 48, 140–150.
- Larichev, O., Asanov, A., Naryzhny, Y., 2002. Effectiveness evaluation of expert classification methods. *European Journal of Operational Research* 138, 260–273.
- Larson, J.S., Bradlow, E.T., Fader, P.S., 2005. An exploratory look at supermarket shopping paths. *International Journal of Research in Marketing* 22, 395 – 414.
- Lauer, F., Bloch, G., 2008. Incorporating prior knowledge in support vector machines for classification: A review. *Neurocomputing* 71, 1578–1594.

- Lessmann, S., Voß, S., 2009. A reference model for customer-centric data mining with support vector machines. *European Journal of Operational Research* 199, 520–530.
- Li, H., Li, Z., Li, L.X., Hu, B., 2000. A production rescheduling expert simulation system. *European Journal of Operational Research* 124, 283–293.
- Liao, I.E., Lin, W.C., 2007. Shopping path analysis and transaction mining based on RFID technology, in: *RFID Eurasia, 2007 1st Annual*, pp. 1–5.
- Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A., 2005. Geographic routing in social networks. *Proceedings of the National Academy of Sciences* 102, 11623–11628.
- Liebig, T., Wagoum, A., 2012. Modelling microscopic pedestrian mobility using bluetooth. *ICAART* 2, 270 – 275.
- Liu, F., Janssens, D., Wets, G., Cools, M., 2013. Annotating mobile phone location data with activity purposes using machine learning algorithms. *Expert Systems with Applications* 40, 3299–3311.
- Ly, Y., Duan, Y., Kang, W., Li, Z., Wang, F.Y., 2014. Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems* 16, 865–873.
- Lymberopoulos, D., Liu, J., Yang, X., Choudhury, R.R., Handziski, V., Sen, S., 2015. A realistic evaluation and comparison of indoor location technologies: Experiences and lessons learned, in: *Proceedings of the 14th international conference on information processing in sensor networks*, ACM. pp. 178–189.
- Madhavapeddy, A., Tse, A., 2005. A study of bluetooth propagation using accurate indoor location mapping, in: *UbiComp 2005: Ubiquitous Computing*. Springer, pp. 105–122.
- Mahmud, J., Nichols, J., Drews, C., 2014. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 47.
- Manevitz, L.M., Yousef, M., 2001. One-class svms for document classification. *Journal of machine Learning research* 2, 139–154.

- Mazuelas, S., Bahillo, A., Lorenzo, R.M., Fernandez, P., Lago, F., Garcia, E., Blas, J., Abril, E.J., et al., 2009. Robust indoor positioning provided by real-time rssi values in unmodified wlan networks. *Selected Topics in Signal Processing*, IEEE Journal of 3, 821–831.
- McSherry, F., Mironov, I., 2009. Differentially private recommender systems: Building privacy into the netflix prize contenders, in: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 627–636.
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2017. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. URL: <https://CRAN.R-project.org/package=e1071>. r package version 1.6-8.
- Meyners, J., Barrot, C., Becker, J.U., Bodapati, A.V., 2017. Reward-scrounging in customer referral programs. *International Journal of Research in Marketing* 34, 382–398.
- Miguéis, V.L., Camanho, A.S., Borges, J., 2017. Predicting direct marketing response in banking: comparison of class imbalance methods. *Service Business* 11, 831–849.
- Millonig, A., Gartner, G., 2008. Shadowing-tracking-interviewing: How to explore human spatio-temporal behaviour patterns., in: *BMI*, pp. 1–14.
- von Mörner, M., 2017. Application of call detail records-chances and obstacles. *Transportation research procedia* 25, 2233–2241.
- Moya, M.M., Koch, M.W., Hostetler, L.D., 1993. One-class classifier networks for target recognition applications. Technical Report. Sandia National Labs., Albuquerque, NM (United States).
- Murtagh, F., 1985. Multidimensional clustering algorithms. *Compstat Lectures*, Vienna: Physika Verlag, 1985 .
- Musa, A., Eriksson, J., 2012. Tracking unmodified smartphones using wi-fi monitors, in: *Proceedings of the 10th ACM conference on embedded network sensor systems*, pp. 281–294.
- Nakahara, T., Uno, T., Yada, K., 2010. Extracting promising sequential patterns from RFID data using the lcm sequence, in: *Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, pp. 244–253.

- Narayanan, A., Shmatikov, V., 2006. How to break anonymity of the netflix prize dataset. arXiv preprint cs/0610105 .
- Naujokaite, R., 2018. Is gdpr consent required for the use of anonymous data? URL: <https://www.chino.io/blog/what-is-anonymous-data-according-to-gdpr/>. accessed: 2021-01-13.
- Newzoo, 2020. Global mobile market report 2020. URL: <https://newzoo.com/insights/trend-reports/newzoo-global-mobile-market-report-2020-free-version/>. accessed: 2021-01-13.
- Nitzan, I., Libai, B., 2011. Social effects on customer retention. *Journal of Marketing* 75, 24–38.
- Nolte, T., Lynch, N., 2007. A virtual node-based tracking algorithm for mobile networks, in: *Distributed Computing Systems, 2007. ICDCS'07. 27th International Conference on*, IEEE. pp. 1–1.
- Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., Mascolo, C., 2012. A tale of many cities: universal patterns in human urban mobility. *PloS one* 7.
- Onnela, J.P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., Barabási, A.L., 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences* 104, 7332–7336.
- Oosterlinck, D., Baecke, P., Benoit, D.F., 2021. Home location prediction with telecom data: Benchmarking heuristics with a predictive modelling approach. *Expert Systems with Applications* 170, 114507. doi:doi: <https://doi.org/10.1016/j.eswa.2020.114507>.
- Oosterlinck, D., Benoit, D.F., Baecke, P., 2020. From one-class to two-class classification by incorporating expert knowledge: Novelty detection in human behaviour. *European Journal of Operational Research* 282, 1011 – 1024. doi:doi: <https://doi.org/10.1016/j.ejor.2019.10.015>.
- Oosterlinck, D., Benoit, D.F., Baecke, P., Van de Weghe, N., 2017. Blue-tooth tracking of humans in an indoor environment: An application to shopping mall visits. *Applied Geography* 78, 55 – 65. doi:doi: <https://doi.org/10.1016/j.apgeog.2016.11.005>.

- OSM, 2016. Indoor mapping - openstreetmap wiki. URL: http://wiki.openstreetmap.org/wiki/Indoor_Mapping.
- O'Neill, E., Kostakos, V., Kindberg, T., Penn, A., Fraser, D.S., Jones, T., et al., 2006. Instrumenting the city: Developing methods for observing and understanding the digital cityscape, in: *UbiComp 2006: Ubiquitous Computing*. Springer, pp. 315–332.
- Pashigian, B.P., Gould, E.D., 1998. Internalizing externalities: The pricing of space in shopping malls. *The Journal of Law and Economics* 41, 115–142.
- Patcha, A., Park, J.M., 2007. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer networks* 51, 3448–3470.
- Pearson, R.K., 2005. Mining imperfect data: Dealing with contamination and incomplete records. volume 93. Siam.
- Pfeiffer, D., Robinson, T.P., Stevenson, M., Stevens, K.B., Rogers, D.J., Clements, A.C., et al., 2008. Spatial analysis in epidemiology. volume 142. Oxford University Press Oxford.
- Phithakkitnukoon, S., Smoreda, Z., 2016. Influence of social relations on human mobility and sociality: a study of social ties in a cellular network. *Social Network Analysis and Mining* 6, 42.
- Phithakkitnukoon, S., Smoreda, Z., Olivier, P., 2012. Socio-geography of human mobility: A study using longitudinal mobile phone data. *PloS one* 7, e39253.
- Phua, C., Lee, V., Smith, K., Gayler, R., 2010. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.
- Pimentel, M.A., Clifton, D.A., Clifton, L., Tarassenko, L., 2014. A review of novelty detection. *Signal Processing* 99, 215–249.
- Platt, J., et al., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 61–74.
- Polikar, R., 2012. Ensemble learning, in: *Ensemble machine learning*. Springer, pp. 1–34.

- Porter, M.F., et al., 1980. An algorithm for suffix stripping. *Program* 14, 130–137.
- Prinzie, A., Van den Poel, D., 2006. Investigating purchasing-sequence patterns for financial services using markov, mtd and mtdg models. *European Journal of Operational Research* 170, 710–734.
- Prinzie, A., Van den Poel, D., 2007. Predicting home-appliance acquisition sequences: Markov/markov for discrimination and survival analysis for modeling sequential information in nptb models. *Decision support systems* 44, 28–45.
- Pyrgelis, A., Kourtellis, N., Leontiadis, I., Serrà, J., Soriente, C., 2018. There goes wally: Anonymously sharing your location gives you away, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE. pp. 1218–1227.
- Quinlan, E., 2008. Conspicuous invisibility: shadowing as a data collection strategy. *Qualitative Inquiry* 14, 1480 – 1499.
- Quinn, J.A., Williams, C.K., 2007. Known unknowns: Novelty detection in condition monitoring, in: *Iberian Conference on Pattern Recognition and Image Analysis*, Springer. pp. 1–6.
- R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- Rizos, C., Roberts, G., Barnes, J., Gambale, N., 2010. Experimental results of locata: A high accuracy indoor positioning system, in: *Indoor Positioning and Indoor Navigation (IPIN)*, 2010 International Conference on, IEEE. pp. 1–7.
- Rodriguez, M., Pece, J.P., Escudero, C.J., 2005. In-building location using bluetooth, in: *International Workshop on Wireless Ad-hoc Networks*.
- Roelens, I., Baecke, P., Benoit, D.F., 2016. Identifying influencers in a social network: The value of real referral data. *Decision Support Systems* 91, 25–36.
- Sagi, O., Rokach, L., 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, e1249.

- Saxena, S., Brémond, F., Thonnat, M., Ma, R., 2008. Crowd behavior recognition for video surveillance, in: *Advanced Concepts for Intelligent Vision Systems*, Springer. pp. 970–981.
- Scherrer, L., Tomko, M., Ranacher, P., Weibel, R., 2018. Travelers or locals? identifying meaningful sub-populations from human movement data in the absence of ground truth. *EPJ Data Science* 7, 19.
- Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C., 2000. Support vector method for novelty detection, in: *Advances in neural information processing systems*, pp. 582–588.
- Sciensano, 2020. Coronalert hoe werkt het? URL: <https://coronalert.be/nl/hoe-werkt-het/>. accessed: 2021-01-08.
- Senion, 2016. Senion indoor positioning system indoor navigation. URL: <https://senion.com/>.
- Singel, R., 2009. Netflix spilled your brokeback mountain secret, lawsuit claims. *Threat Level (blog)*, *Wired* .
- Sofman, B., Neuman, B., Stentz, A., Bagnell, J.A., 2011. Anytime on-line novelty and change detection for mobile robots. *Journal of Field Robotics* 28, 589–618.
- Soh, W.S., et al., 2007. A comprehensive study of bluetooth signal parameters for localization, in: *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on, IEEE*. pp. 1–5.
- Song, C., Qu, Z., Blumm, N., Barabási, A.L., 2010. Limits of predictability in human mobility. *Science* 327, 1018–1021.
- Sorensen, H., 2003. The science of shopping. *Marketing Research* 15, 30 – 35.
- Steinwart, I., Hush, D., Scovel, C., 2005. A classification framework for anomaly detection. *Journal of Machine Learning Research* 6, 211–232.
- Surace, C., Worden, K., 2010. Novelty detection in a changing environment: a negative selection approach. *Mechanical Systems and Signal Processing* 24, 1114–1128.

- Suryavanshi, N.B., Reddy, K.V., Chandrika, V.R., 2019. Direction finding capability in bluetooth 5.1 standard, in: International Conference on Ubiquitous Communications and Network Computing, Springer. pp. 53–65.
- Takai, K., Yada, K., 2010. Relation between stay-time and purchase probability based on RFID data in a japanese supermarket, in: Knowledge-Based and Intelligent Information and Engineering Systems. Springer, pp. 254–263.
- Tang, J., Liu, F., Wang, Y., Wang, H., 2015. Uncovering urban human mobility from large scale taxi gps data. *Physica A: Statistical Mechanics and its Applications* 438, 140–153.
- Tarassenko, L., Hayton, P., Cerneaz, N., Brady, M., 1995. Novelty detection for the identification of masses in mammograms .
- Tax, D.M., Duin, R.P., 2001. Uniform object generation for optimizing one-class classifiers. *Journal of machine learning research* 2, 155–173.
- Tax, D.M., Duin, R.P., 2004. Support vector data description. *Machine learning* 54, 45–66.
- Tizzoni, M., Bajardi, P., Decuyper, A., King, G.K.K., Schneider, C.M., Blondel, V., Smoreda, Z., González, M.C., Colizza, V., 2014. On the use of human mobility proxies for modeling epidemics. *PLoS computational biology* 10, e1003716.
- Turner, A., 2020. How many phones are in the world?
URL: <https://www.bankmycell.com/blog/how-many-phones-are-in-the-world>. accessed: 2021-01-13.
- Utkin, L.V., Zhuk, Y.A., 2014. Imprecise prior knowledge incorporating into one-class classification. *Knowledge and information systems* 41, 53–76.
- Van Vlasselaer, V., Meskens, J., Van Dromme, D., Baesens, B., 2013. Using social network knowledge for detecting spider constructions in social security fraud, in: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM. pp. 813–820.

- Vanhoof, M., Lee, C., Smoreda, Z., 2018a. Performance and sensitivities of home detection from mobile phone data. arXiv preprint arXiv:1809.09911 .
- Vanhoof, M., Reis, F., Ploetz, T., Smoreda, Z., 2018b. Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics* 34, 935–960.
- Vanhoof, M., Reis, F., Smoreda, Z., Plötz, T., 2018c. Detecting home locations from cdr data: introducing spatial uncertainty to the state-of-the-art. arXiv preprint arXiv:1808.06398 .
- Vazquez-Prokopec, G.M., Bisanzio, D., Stoddard, S.T., Paz-Soldan, V., Morrison, A.C., Elder, J.P., Ramirez-Paredes, J., Halsey, E.S., Kochel, T.J., Scott, T.W., et al., 2013. Using gps technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. *PloS one* 8.
- Verbeke, W., Martens, D., Baesens, B., 2014. Social network analysis for customer churn prediction. *Applied Soft Computing* 14, 431–446.
- Versichele, M., 2014. Sensing and making sense of crowd dynamics using Bluetooth tracking: an application-oriented approach. Ph.D. thesis. Ghent University.
- Versichele, M., Delafontaine, M., Neutens, T., Van de Weghe, N., 2010. Potential and implications of bluetooth proximity-based tracking in moving object research, in: 1st Int. workshop on movement pattern analysis (MPA) in conj. with the 6th Int. conf. on Geographic Information Science, pp. 111–116.
- Versichele, M., de Groote, L., Bouuaert, M.C., Neutens, T., Moerman, I., de Weghe, N.V., 2014a. Pattern mining in tourist attraction visits through association rule learning on bluetooth tracking data: A case study of ghent, belgium. *Tourism Management* 44, 67 – 81.
- Versichele, M., Neutens, T., Claeys Bouuaert, M., Van de Weghe, N., 2014b. Time-geographic derivation of feasible co-presence opportunities from network-constrained episodic movement data. *Transactions in GIS* 18, 687–703.
- Versichele, M., Neutens, T., Delafontaine, M., Van de Weghe, N., 2012a. The use of bluetooth for analysing spatiotemporal dynamics of human

- movement at mass events: A case study of the ghent festivities. *Applied Geography* 32, 208–220.
- Versichele, M., Neutens, T., Goudeseune, S., Van Bossche, F., Van de Weghe, N., 2012b. Mobile mapping of sporting event spectators using bluetooth sensors: tour of flanders 2011. *Sensors* 12, 14196–14213.
- Vij, A., Shankari, K., 2015. When is big data big enough? implications of using gps-based surveys for travel demand analysis. *Transportation Research Part C: Emerging Technologies* 56, 446–462.
- Vukovic, M., Lovrek, I., Kraljevic, H., 2012. Discovering shoppers' journey in retail environment by using RFID., in: *KES*, pp. 857–866.
- Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabasi, A.L., 2011. Human mobility, social ties, and link prediction, in: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, Acm. pp. 1100–1108.
- Wang, W., Zhang, W., 2008. An asset residual life prediction model based on expert judgments. *European Journal of Operational Research* 188, 496–505.
- Wang, Z., He, S.Y., Leung, Y., 2018. Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society* 11, 141–155.
- Yamin, M., Ades, Y., 2009. Crowd management with rfid and wireless technologies, in: *Networks and Communications, 2009. NETCOM'09. First International Conference on*, IEEE. pp. 439–442.
- Yesuf, A.S., Wolos, L., Rannenberg, K., 2017. Fraud risk modelling: requirements elicitation in the case of telecom services, in: *International Conference on Exploring Services Science*, Springer. pp. 323–336.
- Zagatti, G.A., Gonzalez, M., Avner, P., Lozano-Gracia, N., Brooks, C.J., Albert, M., Gray, J., Antos, S.E., Burci, P., zu Erbach-Schoenberg, E., et al., 2018. A trip to work: Estimation of origin and destination of commuting patterns in the main metropolitan regions of haiti using cdr. *Development Engineering* 3, 133–165.
- Zhao, F., Pereira, F.C., Ball, R., Kim, Y., Han, Y., Zegras, C., Ben-Akiva, M., 2015. Exploratory analysis of a smartphone-based travel survey in singapore. *Transportation Research Record* 2494, 45–56.

Zhou, S., Pollard, J.K., 2006. Position measurement using bluetooth. *Consumer Electronics, IEEE Transactions on* 52, 555–558.