Philipp Schosteritsch

Inferring Road Traffic from Mobile Phone Network Data

Master Thesis

Institute for Knowledge Discovery
Laboratory for Brain-Computer Interfaces
Graz University of Technology
Krenngasse 37, 8010 Graz, Austria
Head: Assoc. Prof. Dipl.-Ing. Dr.techn. Gernot Müller-Putz

Supervisor: Ass.Prof. Dipl.-Ing. Dr.techn. Reinhold Scherer
Dipl.-Ing. Dipl.-Ing Thomas Reiter
Dipl.-Ing. Michael Cik

Evaluator:
Ass.Prof. Dipl.-Ing. Dr.techn. Reinhold Scherer

Graz, August 2012

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am …… ……..
(Unterschrift)

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

…… ……
Date Signature

Danksagung

An dieser Stelle möchte ich mich bei Herrn Dr. Reinhold Scherer bedanken, der sich als Betreuer zur Verfügung stellte und mir das Verfassen dieser Diplomarbeit ermöglichte.

Außerdem bedanke ich mich bei meinen Eltern und bei der Frau meines Vaters Gisela die mir mein Studium ermöglicht haben.

Mit bestem Dank an die ASFiNAG für die Zurverfügungstellung der Zählstellendaten.

ASFiNAG

Inferring on Road Traffic from Mobile Network Data

Kurzfassung

Die Methoden um Verkehrsmodelle zu erstellen die derzeit bekannt sind und angewandt werden sind zeit- und kostenintensiv. Daher werden die Ergebnisse über Jahrzehnte hinweg verwendet ohne in dieser Zeit verifiziert oder angepasst zu werden. In dieser Arbeit soll das Potential von Mobilfunknetzdaten ermittelt werden um Verkehrsmodellierung und Planung zu verbessern. Heutzutage besitzt fast jede Person ein Mobilfunktelefon. Auf Grund der Arbeitsweise von Mobilfunknetzen haben Mobilfunknetzanbieter Verwendungs- und Aufenthaltsdaten ihrer Kunden. Diese Daten erscheinen vielversprechend um auf den Straßenverkehr schließen zu können.

Ein österreichischer Mobilfunknetzanbieter hat zu diesem Zweck die Daten seines Netzes für diese Arbeit zur Verfügung gestellt. Diese Daten sind keine Floating Car Data (FCD) oder Daten die von Mobiltelefonanwendungen kommen. *Bei diesen Daten handelt es sich um pseudonymisierte Ereignismeldungen von den Schnittstellen des Mobilfunknetzes. Daher wird für die in dieser Arbeit beschriebenen Methoden keine Kooperation von Mobiltelefonnutzern benötigt wie z.B. das Installieren oder Verwenden von Anwendungen auf dem Mobiltelefon.*

Im ersten Schritt dieser Arbeit werden die Daten des Anbieters analysiert. Phänomene wie z.B. Anachronismen, die in den Daten auftauchen, werden aufgezeigt. In diesem Schritt wird auch die Wichtigkeit der Verifizierung der Daten erläutert und der Einfluss der Phänomene auf angewandte Methoden und Ergebnisse wird erklärt.

Im zweiten Schritt werden die Daten einer statistischen Analyse unterzogen. Korrelationen von Ereignissen, den empfangenen Daten und der Umwelt werden aufgezeigt. In diesem Schritt werden vier Tageszeiten identifiziert, in denen sich das Verhalten von Mobilfunknutzern ändert.

Für die vier Tageszeiten aus der Statistik werden Markov-Modelle erstellt die Sequenzen aus Ereignisberichten aus dem Mobilfunknetz darstellen sollen. Außerdem werden Markov-Modelle für die Ortswechsel von Mobilfunkteilnehmern in den vier Tageszeiten erstellt. Auf Grund der Ergebnisse der Markov-Modelle wird eine Methode vorgestellt, die nur scheinbare Ortswechsel (also ein Ortswechsel ohne einer tatsächlichen Bewegung) erkennt.

Schließlich werden die Ergebnisse der vorigen Schritte kombiniert und ein Algorithmus wird vorgestellt, der anhand der Mobilfunknetzdaten die Verkehrssituation einer beliebig definierten Route darstellt.

Schlüsselwörter: Mobilfunknetz, GSM, UMTS, Straßenverkehr, Verkehrsmodel

Inferring on Road Traffic from Mobile Network Data

Abstract

The current methods to create road traffic models cost a lot of money and effort. So once a model is created it is used for decades. These models usually base on surveys. They are neither verified nor are they refined because of the costs and the effort. This thesis aims to assist in solving these problems by using mobile phone network data to infer on road traffic.

Today almost everybody has and uses at least one mobile phone. Mobile phone network providers have data from their clients and the usage of their network. Because of the design and the structure of mobile phone networks their providers know about the locations of their clients. This fact makes mobile phone network data interesting for inferences on road traffic.

An Austrian mobile phone network operator provided data of his network for this thesis to have it evaluated and analyzed with regards to infer on road traffic and to find a solution that supports traffic analyzes and traffic surveillance. *The data are neither floating car data (FCD) nor data sent from an application on a mobile phone*. *The data are pseudonymized event reports from interfaces of the mobile phone network* and therefore *do not require any cooperation of users like using special application or devices*.

In the very first step of this thesis the data are analyzed. Detected phenomena and abnormalities of received data are described. In this stage the importance of verification and the influence of phenomena on statistics and methods are shown.

In a second step basic statistical analyzes are performed on the data. Correlations of certain event frequencies with other frequencies as well as with the number of detected devices are shown. In these analyzes four day times are identified where a change in users' behavior as well as their mobility can be seen.

Taking these results a descriptive Markov Model is created to look into the frequencies of certain event types as well as into the change of users' move directions. An algorithm which detects event sequences which were not caused by movements is introduced.

Finally by combining all previous results an algorithm is presented which infers on the situation on a route. The algorithm filters phenomena as well as event sequences not caused by movements. It is inspired by the result of the Markov Model of location changes and evaluates data of users in an area around the route to see how many users are on a road way.

Key words: Mobile Phone Network, GSM, UMTS, Road Traffic, Traffic Model

Table of Content

Glossary

2G	2G means second generation and refers to the GSM technology.
3G	3G means third generation and refers to the UMTS technology.
A	The A interface is between the Base Station Controller and the Mobile Switching Center.
Abis	The Abis interface is between the Base Station and the Base Station Controller.
AMPS	Advanced Mobile Phone Service was an analog, cellular, mobile phone system standard developed by AT&T Bell Laboratories and used for instance by the C-Netz in Germany.
ANPR	Automatic Number Plate Recognition is a surveillance method where registration plates of cars are read automatically.
ASFiNAG	Autobahnen und Schnellstraßen Finananzierungs-Aktiengesellschaft – a publicly owned company in Austria responsible for roadway planning, financing, building and maintenancing.
BS	Base Station or Base Transceiver Station is a term for cell site. It is connected to the Base Station Controller using the Abis interface.
BSC	The Base Station Controller governs a set of Base Stations. It is connected to the Mobile Switching Center via the A interface.
BTS	Base Transceiver Station is a term for cell site.
Cell ID	Identification number of a cell site. The number sometimes tells which antenna of the site is used. The mobile device is aware of the ID of the cell it is using.
Cell Site	A cell site is an element of the 2G GSM network. It is in direct contact with users' mobile devices. It is connected to the Base Station Controller via the Abis interface. One cell site usually serves one GSM cell. In 3G networks the Base Station is called Node B.
CN	The core network (or network core) is the heart of a GSM network. It is operated by mobile switching centers (MSC).
CS	Circuit switching is a method to connect two nodes in a network. A dedicated communication channel (the circuit) is established before the nodes can start communicating. The channel remains even if there is no data flow.
ETSI	European Telecommunications Standards Institute is an independent non-profit organization. It standardized the GSM cell phone system.

ESRI	Environmental Systems Research Institute Inc. is an American software company providing geographic information systems.
Gb	The Gb interface connects the Base Station Controller with a SGSN.
GPRS	General Packet Radio Service is a standard developed to enable packet transmissions in GSM networks. GPRS is a domain in a mobile network and connected to the core network via the Gb interface. It is an integrated part of the core network.
GPS	Global Positioning System is a satellite based navigation system maintained by the government of the United States of America.
GSM	Global System for Mobile Communications or Groupe Spécial Mobile is a standard developed by the ETSI to describe technologies for the second generation of digital cellular networks.
IMEI	The International Mobile Equipment Identity is a code that identifies a mobile station. It is associated with the device.
IMSI	The International Mobile Subscriber Identity is a code that identifies a subscriber. It is associated with the SIM card. It is cloaked by the TMSI.
MS	A Mobile Station is a term for a mobile device. In 3G networks it is sometimes referred as user equipment (UE).
MSC	The Mobile Switching Venter is a service to route calls in the core of a GSM network.
MSISDN	The Mobile Subscriber Integrated Services Digital Network Number is what is commonly known as “phone number”.
LTE	Long Term Evolution is a standard for wireless communication and high speed data transfer. It is marketed as 4G.
PS	Packet switching is a method to connect two network nodes by dividing data into blocks and transmit them as packets.
RNC	The Radio Network Controller manages Node B’s. It is connected to the GPRS domain via the packet switched Iu interface (Iu PS) and to the core network via the circuit switched Iu interface (Iu CS).
SGSN	The Serving GPRS Support Node manages packet delivery from and to mobile devices.
TMSI	The Temporary Mobile Subscriber Identity is a code to cloak the real IMSI.

TLLI	The Temporary Logical Link Identifier is used to route packets to a subscriber. In contrary to the TMSI it has no cloaking purpose.
UE	UE is an abbreviation for user equipment. This term refers to mobile devices in 3G networks.
UMTS	Universal Mobile Telecommunication Standard is developed by the 3GPP (Third Generation Partnership Project) and describes a standard for the third generation (3G) of digital cellular networks.
UTRAN	Universal Terrestrial Radio Access Network is a term for sets of Node B's and Radio Network Controllers. It makes communication with mobile devices and the core network possible.

1 Introduction

1.1 Situation

Today traffic is monitored using external devices like counting stations and other sensors or employing global positioning system (GPS) devices and use floating car data (FCD)[1][2]. Google Maps (http://maps.google.at/) for instance offers the feature "Time in current traffic" to see how much traffic there is on the roads of certain cities and to see how fast travelling should be possible on a given route. An Austrian example is http://www.anachb.at/, a website which offers a traffic service from the ITS Vienna Region. This site shows the traffic in the region of Vienna and updates its results in given intervals.

To estimate the current traffic Google collects anonymized GPS data from persons who use Google Maps from a GPS-enabled phone. These data are then combined with historic traffic data to calculate an estimate of the current traffic situation [3]. This system therefore lives from people who use Google Maps on their mobile phone with GPS support.

The service site http://www.anachb.at/ displays the current traffic situation in the region of Vienna as well as recent reports about construction sites and traffic events. Using this information it offers improved route planning for different types of traffic participants (like bicycles, etc.). Anachb.at does not rely on data from persons who use their services. The site collects FCD information from Taxi cars / yellow caps, information from road surveillance sensors, from public transports and several other sources which report about the traffic situation (like the police, etc.). These data are combined with a traffic model, the Graphen-Integrationsplattform (GIP), to estimate the current traffic situation on the roads [1].

But road traffic isn't only provided to traffic information web sites. The "Autobahnen und Schnellstraßen Finananzierungs-Aktiengesellschaft" - short ASFiNAG – employs several permanent traffic counters and makes the collected data available to everybody. These data contain extended information like direction, vehicle class (personal car, bus, truck) and quality of data [4]. Unfortunately source-destination relations cannot be inferred with certainty from these data.

Beside FCD information and information from road surveillance sensors mobile phone networks can be used. They are used for instance to transmit traffic information. Mobile phone networks cover almost every area in a country. For example the A1 Telekom Austria AG, an Austrian mobile services operator and provider covers almost whole Austria [5]. Today's navigation systems like those from Tom Tom International BV make also use of mobile phone networks to enhance their service for users. Location data of navigation system users are reported to the navigation system provider [2]. The data are evaluated and other users are informed about traffic traffic jams or other obstacles on the route. Such navigation system use mobile phone networks to send and receive data. But as far as it could be verified for this thesis, data from other mobile subscribers and mobile devices like cell phones are not used in Austria yet. Though using this data is subject to current research.

Further information on this topic can be found at the Forschungszentrum Telekommunikation Wien (http://www.ftw.at).

1.2 Motivation

There are several approaches to use mobile phones (or mobile devices in general) for traffic monitoring.

One approach is equipping mobile devices with software to report the current location. For cell phones this could be an application which reports either global positioning system (GPS) coordinates (if available) or the cell site the phone is subscribed to. The id of the cell site is then looked up for its coordinates. These look ups can be done either on open and free databases or on databases. Mapping the cell site to its coordinates only gives a rough idea of the location. In rural areas the deviation may be up to several kilometers. The accuracy may be increased by evaluating the strength of the received signal. Helmut Georg Feiertag researched this topic in his master thesis [6]. He found the different signal strengths not suitable to determine the position or the speed of the subway train but he found a correlation in the variation of the signal strength and in either exiting the train or continuing to the next station.

Another approach is using data from mobile network operators. This is not a new approach but it is still subject of research. In 2000 a paper of researches from the University of Genua the usage of mobile phone network data to estimate traffic parameters on routes covered by mobile cell sites [7]. Unfortunately the authors of this paper suppose a device is detected by a cell site when it enters its area and that devices are forced to signal every change. This assumption is incorrect [8]. A mobile device detects the cell sites it is in but a mobile phone network does not know at which cell site a subscriber is until it actively searches him or her. The only change reported by a mobile device is the change of so called Local Area codes.

Johannes Schlaich researches how to use mobile network data to infer on (motorway) route selection in his doctor thesis “Nutzung von Mobilfunkdaten für die Analyse der Routenwahl” [9]. The data for his thesis are retrieved by so called “network probes”. The area observed is nearby Karlsruhe and Stuttgart. His focus is on using tracking movements between Location Areas and inferring on the route selection by them. He is able to calculate a trajectory (movement path) of a user who moves more than 20 km and passes at least three location areas. It took 14 hours to evaluate an average day.

Danilo Valerio describes improvement of road traffic detection systems using mobile phone network signaling data in his technical report [10]. He mentions active and passive road monitoring techniques. The active technique involves a client application which reports location and position data from the client device. The passive technique involves silent collection of signaling data from one or more points without an impact on the network load. In his results he describes existing correlations of certain events from the mobile network and the traffic flow as well as the correlation of such events with car accidents. Danilo Valerio states that it is possible to infer from mobile network events on road traffic. Mapping the

events of the mobile network to event occurring on the road is still a subject of further research.

This thesis sets up on using data from mobile phone network providers to infer on road traffic.

1.2.1 Objectives

In this thesis the data of the mobile phone network provider are analyzed for their potential to support traffic modeling and inferring on traffic. In several other theses the data from cell phone carriers were similar but still different in detail. Reports about problems with these or obstacles are missing. Therefore the validation and analyses processes are part of this thesis.

The data of the carrier contain information from the mobile phone network. Information about movements has to be extracted from the data. Therefore it is a goal of this thesis to identify the information which can be used to infer on traffic and to find methods to filter information which is not suitable. Further a concrete method is presented that uses the mobile phone network data to infer on an actual traffic situation. Other applications are suggested and discussed.

1.3 Introduction of Mobile Phone Networks

Mobile phone networks (also called cellular networks) are radio networks which are divided into cells. These cells are areas served by at least one cell site. All cells in a whole usually cover large geographic areas like a country. Mobile phone networks are used for communication and data transfer. When in reach of a cell site (when the cell phone can receive the cell site and the cell site can receive the phone) a user can subscribe to the network and use it. In other words she or he may use her or his cell phone.

Predecessors of the today mobile phone networks were analog, circuit switched radio using the Advanced Mobile Phone Service (AMPS) standard developed by AT&T Bell Laboratories. Networks using this standard were the first generation (1G) of mobile phone networks.

Today there are second and third generation networks. GSM (Global System for Mobile Communications or Groupe Spécial Mobile) is a describing standard for digital mobile phone networks. GSM was developed by the ETSI (European Telecommunications Standards Institute)[11]. It is the second generation (2G) of mobile phone networks. GSM is a digital and circuit switched network standard.

General Packet Radio Service (GPRS) - also developed by the ETSI – is a standard to transmit packets in GSM networks. GPRS is not circuit switched anymore, but already packet switched. It is between the second and the third generation (2.5 G) of mobile phone networks.

UMTS (Universal Mobile Telecommunication Standard) is developed by the 3GPP (Third Generation Partnership Project) and describes a standard for the third generation (3G) of digital mobile phone networks. HSDPA (High Speed Downlink Packet Access) is an evolution of the UMTS standard and it is sometimes called third and a half generation (3.5 G).

However HSDPA is no actual technique by itself. It is rather a more efficient way to use UMTS. UMTS and HSDPA are digital, packed switched network standards.

A candidate for the fourth generation (4G) and therefore the successor of UMTS is LTE Advanced (Long Term Evolution) [12].

1.3.1 GSM Network Layout

Figure 1 bellow illustrates the layout of a GSM network, its segments and its network elements. The green hexagons are the cells of the mobile phone network. In reality the covered area is no hexagon, but rather a circle. The actual area covered by a cell site of course depends on geographic features. Mobile phones are referred as mobile stations (MS) in this image. A common synonym for mobile stations is user equipment (UE).

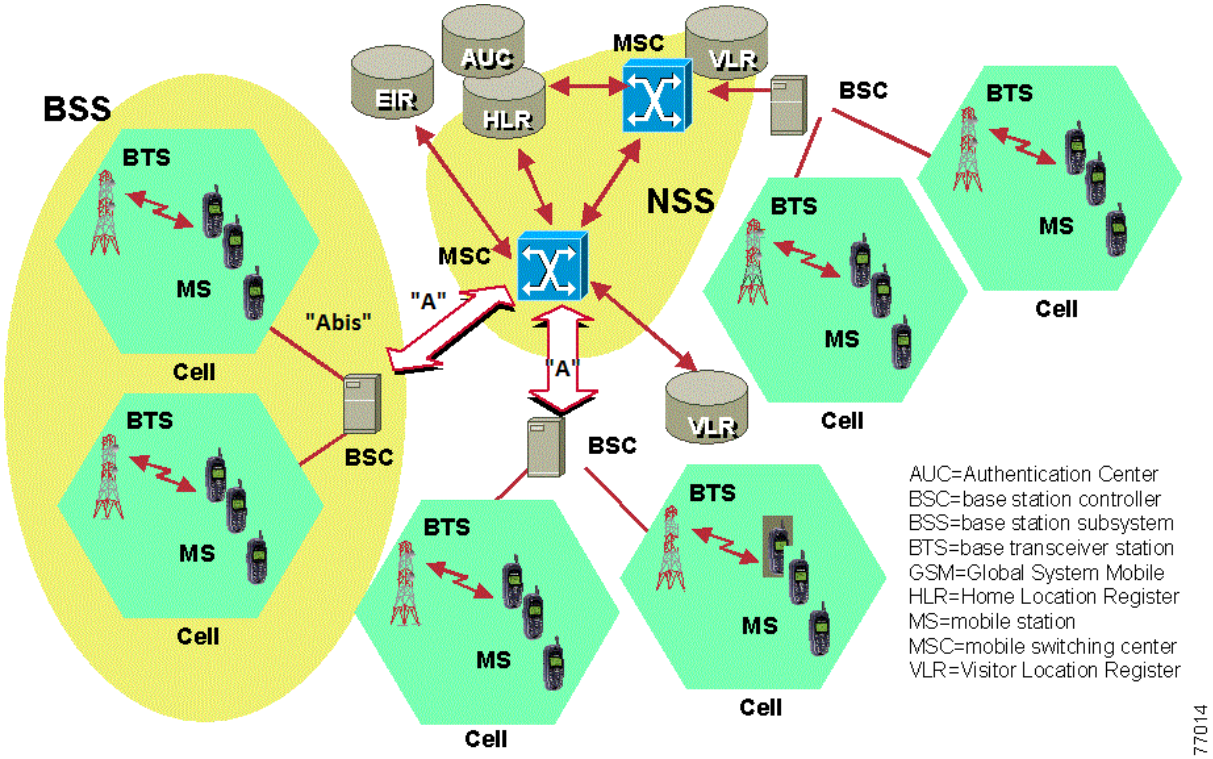


Figure 1: GSM Network Elements
from http://docstore.mik.ua/univercd/cc/td/doc/product/wireless/moblwrls/cmx/mmg_sg/cmxgsm.htm#74011, visited on 11. Feb. 2012

Cell sites are also called base transceiver station (BTS), radio base station or simply base station (BS). In third generation networks they are referred as node B. A crucial part of base stations are antennas. One base station may have one or more antennas. A common set up of base station antennas in the network where the data for this thesis came from was to have up to three antennas. The antennas are usually aligned in 120° angles so the whole area around could be covered. On some base station the antennas are aligned in different angles, usually when no whole area coverage is required. The base station is aware of on which antenna a user is subscribed to. This awareness can then be used to refine the guess of a user's position. Each base station is identified by a Cell ID. The Cell ID is unique within a Location Area (an area with a set of base stations). A Cell ID and the Location Area Code uniquely identify a

base station in a network. The antenna a user is on may be encoded in a Cell ID. If the last digit of the cell id is zero, it may indicate an omnidirectional antenna. If the digit is 1, 2 or 3 it may indicate the other antennas. Antennas may also be referred as sectors.

Several base stations are connected to one base station controller (BSC). The base station controller controls a set of base stations. The controlling tasks cover power control, measurements as well as handling hand-over's (A hand over is the act where the connection of a cell site to a mobile device is handed over to another cell site).

A BSC may be one Location Area or it may be a part of a Location Area. In either ways the Location Area is a set of base stations. A mobile device is aware of which Location Area it is in and it reports on its own to the mobile network if it enters a new Location Area. This is called a Location Area update. The mobile network then knows in which Location Area it has to look for the subscriber if it needs to contact him. In other words if there is an incoming call, text message or other notification for a subscriber only the base stations of his last known Location Area are looking for the mobile device. This also means that location area updates are signals sent to the mobile network without the need of a user's interaction. If a mobile phone is idle for a given amount of time it tells the network about its current Location Area code even without changing the Location Area. This is called a periodical Location Area update.

The base station controller provides several interfaces.

1. The "Abis" interface (shown in Figure 1) connects the base station with the base station controller.

2. The "A" interface (shown in Figure 1) connects the base station controller with the Mobile-Service Switching Center.

3. The "Gb" interface (shown in Figure 2) connects the base station controller with a Serving GPRS Support Node (SGSN). This interface enables connection to the Internet over the General Packet Radio Service (GPRS). The SGSN is responsible for packed switched (PS) operations.

Base stations and base station controllers form the base station subsystem (BSS).

Several base station controllers are connected to one Mobile-Service Switching Center (MSC). The Mobile-Service Switching Center is responsible for circuit switched (CS) operation. Another crucial task is charging and account monitoring. Mobile-Service Switching Centers are not only connected to Base Station Controllers but also to the Short Message Central (SMSC), to Service Control Points (SCPs), Signaling Transfer Points (STPs) and other service end-points. Therefore the MSC is responsible to determine calls leaving the operators network and routing them out to other networks (mobile phone networks as well as land-line phone networks and other networks). Mobile-Service Switching Centers which route data outside the mobile network are called Gateway Mobile-Service Switching Centers.

When looking at the GSM network from a more abstract perspective the network may be divided into the following areas:

The core network (CN): It consists of the MSC to handle circuit switched operations and of the SGSN to handle packet switched operations of any connected mobile device.

The GSM base station subsystem (BSS): The second generation mobile phone network to connect mobile devices to the core network. Connection to the CN uses either the A interface to connect the BSC to the MSC and the Gb interface to connect the BSC to the SGSN.

Not actually part of the GSM network and therefore not included in Figure 1 are the General Packet Radio Service (GPRS) and the Universal Terrestrial Radio Access Network (UTRAN). Both can be "attached" to the core network:

GPRS: A data network that overlays a second-generation GSM network. This data overlay network provides packet data transport at rates from 9.6 to 171 kbps. Additionally, multiple users can share the same air-interface resources simultaneously[13].

UTRAN: The third generation mobile phone network to connect mobile devices to the core network.

1.3.1.1 General Packet Radio Service (GPRS) Overview

GPRS uses most of the GSM network elements to make Internet protocol (IP) based data transfer possible. Some of these elements like base stations need to be upgraded and mobile user devices (cell phones) have to support the enhanced air interface of GPRS. Support nodes like the SGSN are added to the GSM network to do packet switched operations (since prior MSCs do circuit switched operations only).

While the GSM network knows location areas, GPRS has routing areas. "A Location Area is a group of one or several Routing Areas. The Routing Area defines a paging area for GPRS, while the Location Area defines a paging area for incoming circuit-switched calls. Actually, when the network receives an incoming call for a mobile not localized at cell level but localized at Routing Area level, it broadcasts a paging on every cell belonging to this Routing Area." [14]

A GPRS network implements several interfaces to control and support packet transmission over the network. The abbreviations of these names are Ga, Gb, Gc, Gi, Gn, Gp, Gr, Gs and Gf. The event data for this thesis contain information from the Gb interface. This interface connects the SGSN with the BSC. An overview can be found in Figure 2.

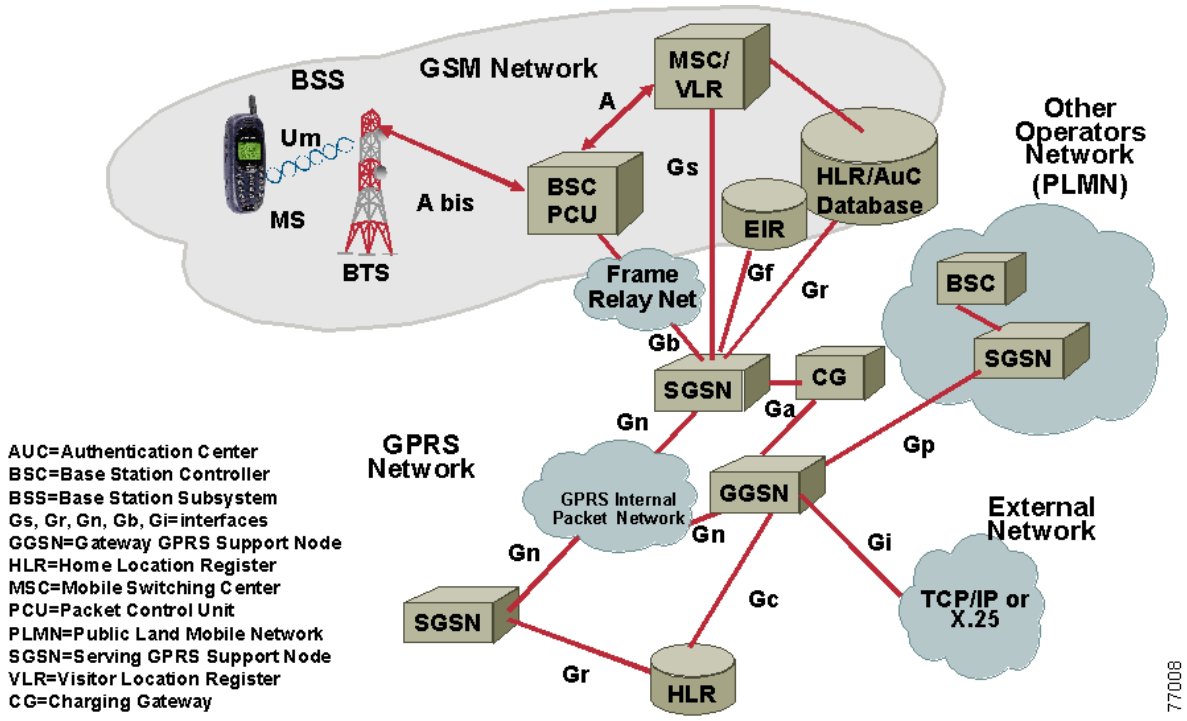


Figure 2: GPRS Interfaces
from http://docstore.mik.ua/univercd/cc/td/doc/product/wireless/moblwrls/cmx/mmg_sg/cmxgsm.htm#68955, visited on 11. Feb. 2012

GPRS is also accessible via UMTS. A mobile device does not necessarily need to be connected via GSM. It may also use be online with UMTS and access GPRS. In this case the connecting interface isn't Gb anymore. The interface for this connection is Iu. More on this is explained in chapter 1.3.1.2.

Beside the network layout and the interfaces the possible states of mobile devices attached to the GPRS need to be briefed. In a survey [15] Danilo Valerio et al. explain the states of mobile devices and the transitions in chapter three (Extended Monitoring Framework). By making use of this information it is possible for them to locate active users at cell granularity and passive users in their routing areas.

1.3.1.2 UMTS Network Layout

UMTS provides higher data transfer rates than GSM. The UMTS network shares the core network with GSM. Only the radio access network is different. Therefore most of the core network structure described in Figure 1 applies for the UMTS network structure too. UMTS does not make use of base stations and the base station controllers of the GSM network. It uses node Bs instead of base stations and radio network controllers (RNCs) instead of base station controllers. The core network is accessed via the "Iu" interface. Figure 3 gives an overview of the architecture. In this image there are no mobile stations but user equipment (UE).

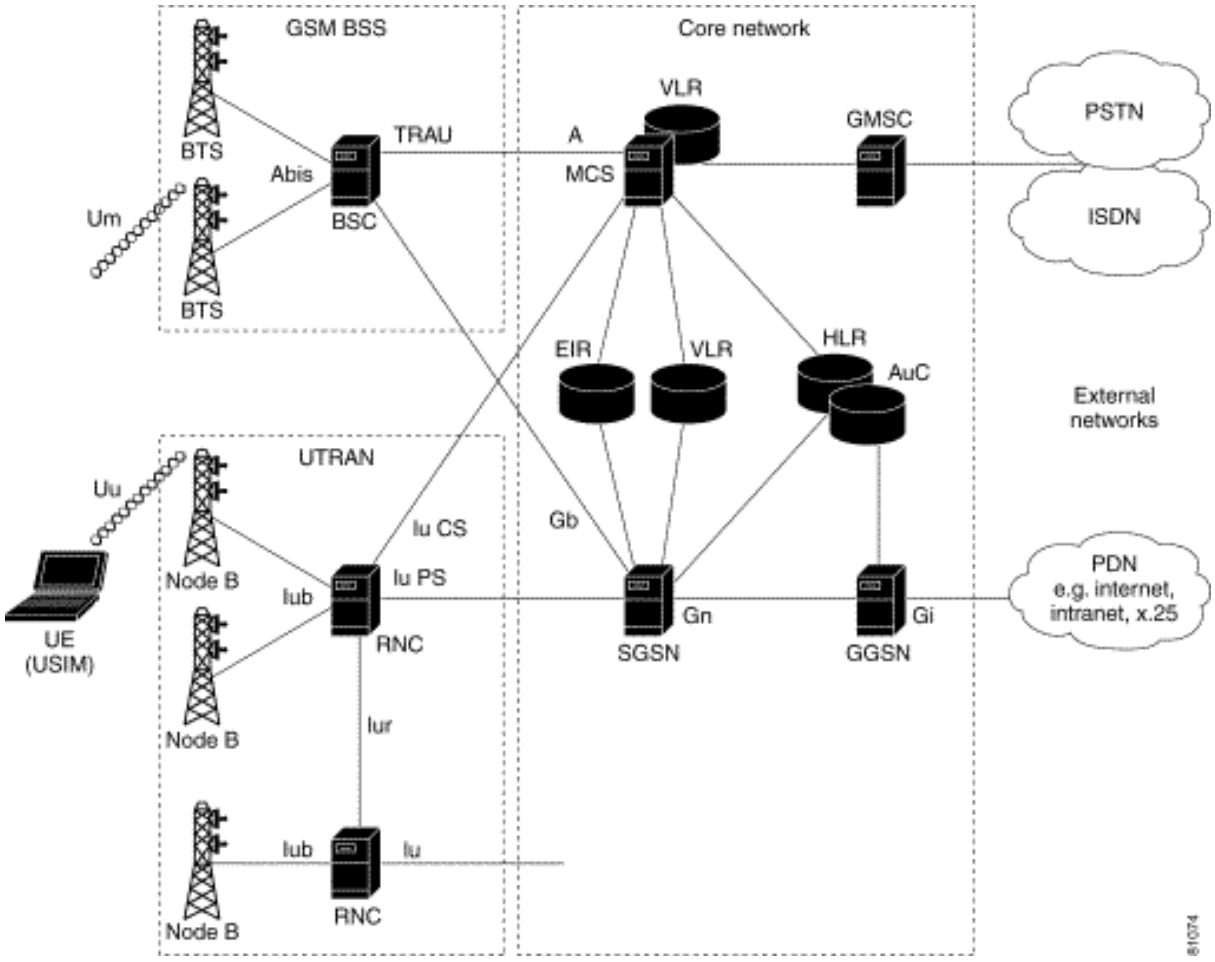


Figure 3: UMTS Architecture
from http://docstore.mik.ua/univercd/cc/td/doc/product/wireless/moblwrls/cmx/mmg_sg/cmxgsm.htm#96688, visited 11. Feb. 2012

The UTRAN covers Node Bs and Radio network controllers. UTRAN is in UMTS basically what the BSS is in GSM. The radio network controller implements the Iu-CS interface to connect to the circuit switched core network (CS-CN). It connects to the packet switched core network (PS-CN) via the Iu-PS interface. The Iu interface carries user traffic and control information.

Most mobile devices support both GSM and UMTS. UMTS is usually preferred by devices if it is available. The user often does not know which network he or she is using.

A detailed explanation of UMTS processes and procedures is described in Chris Johnson's Book "Radio Access Networks for UMTS: Principles and Practice".

1.3.2 User Equipment and IMSI

It makes no difference whether a device is explicitly a mobile phone or explicitly a data modem. When looking at the technical aspects, a person who buys a mobile phone bought a "sophisticated voice modem". Any device which is able to subscribe to a mobile phone network is a mobile station or user equipment. These devices have in common that they are identified by a unique number – the international mobile equipment identity (IMEI) – and they have (at least) limited chip card reading functionality to work with the subscriber identity module card (short SIM card). The SIM card is identified by another unique number - the international mobile subscriber identity (IMSI). The phone number is actually the MSISDN – the Mobile Subscriber Integrated Services Digital Network Number - which identifies the actual subscription.

The IMEI has 14 decimal digits plus an optional check. These 14 digits represented a type allocation code, a final assembly code in old IMEI versions and a serial number. An example for an IMEI is shown in Table 1. Further information can be found at the TS 23.003 specification of the 3GPP.

TAC (plus optional FAC)	Serial Number	Optional Luhn checksum
AA - BB BB BB	CC CC CC	D
35 – 85 37 04	12 34 56	0

Table 1: An example IMEI

An IMSI has 15 digits. It represents a mobile country code (MCC), a mobile network code (MNC) and a mobile subscription identification number (MSIN). An example for an IMEI is shown in Table 2. Further information can be found at the TS 23.003 specification of the 3GPP.

MCC	MNC	MISM
232	01	0123456789

Table 2: An example IMSI

A MSISDN is limited to 15 decimal digits and consists of a country code (CC), a national destination code (NDC) or a number planning area (NPA) and the actual subscriber number (SN). An example for an MSISDN is shown in Table 3. By combining the MCC and the MNC the home network of a SIM card may be identified. The phone number belonging to a SIM card is related to the mobile subscription identification number in the network operator's data base. Further information can be found at the International Telecommunication Unit.

CC	NDC / NDA	SN
43	664	123 45678

Table 3: An example MSISDN

Comparison to the Internet:

The IMEI of mobile devices is somehow comparable to the MAC address of network devices. The IMEI identifies the device but it usually has no relation to the user himself. The user identifies himself by the IMSI of his SIM card. So if a user changes his or her mobile equipment but keeps the SIM card, the mobile network operator won't take much notice. A user could use a mobile phone to do some calls with his SIM card first and then use the USB data stick with the same SIM card to surf in Internet. The network operator would only see that the equipment changed.

The IMSI of a SIM card is somehow comparable to the Internet Protocol (IP) address of a device connected to the Internet. This number identifies the user to the mobile network. However the IMSI is not the phone number. The first three digits of the IMSI are the Mobile Country Code as defined in ITU E.212. The MCC (usually) identifies a geographical country. For an IP address there is no geographical information but a network class (like an IP address with an A-class 127 identifies a local network). The Mobile Network Code identifies the network in the country to which the IMSI belongs. This is comparable to the assignments of IP address space by the Internet Assigned Numbers Authority (IANA).

The MSISDN consists of a country code, a national destination code or the national planning area and the subscriber number. It is somehow comparable to domain names of the internet. A domain name may point to different IP addresses subsequently, so it does not get lost when switching the server or the IP address. Mobile phone numbers may be ported as well. IP addresses in the Internet identify a host uniquely in the whole Internet. The assignment is independent from the underlying equipment – in other words an IP address may be ported from one device to another. The address itself contains routing information. This is similar to IMSIs. IMSIs identify a subscriber uniquely in the GSM network. The IMSI is independent of the underlying device – it may be ported to another phone/modem/device. It contains routing information too. The MAC address of a network device identifies the device uniquely in the local network where it is used. So does the IMEI with the mobile equipment in the operator's network.

Summary:

A user is usually identified by her or his IMSI as well as the MSISDN. The IMSI and the MSISDN give information about the originating country and mobile network of the user. The IMSI gives no information about the phone number of the user or the mobile equipment used by him or her. The MSISDN gives no information about the IMSI or the mobile equipment.

1.3.3 Temporary Mobile Subscriber Identity

Since users are identified by the IMSI it is possible to track a user on the radio interface. In order to avoid that, a temporary mobile subscriber identity (TMSI) is used and changed frequently (usually over an encrypted channel). In theory the actual IMSI is only sent when the mobile device is switched on.

1.3.4 Temporary Logical Link Identifier

The temporary logical link identifier (TLLI) is used in communications between the mobile device and the SGSN. It has no cloaking purpose but uniquely identifies a customer packet data transfer link – in other words it is used for packet routing.

1.4 Introduction to iTraffic Stream Data

The data for this thesis originate from an Austrian mobile network operator. They do not originate from mobile equipment's software. The source technologies are the second and the third generation (GSM and UTMS).

The data are event reports from the mobile network operator's network. These reports inform about events occurring in certain domains of the mobile network (to be more accurate, the events are occurrences at certain interfaces, like the "A" interface). Which events are reported is defined by the mobile network operator, but the events are related to subscribers (i.e. an incoming call or short message is reported).

The current version of event reports is called "Option B". In "Option B" the coordinates are the expected position of a subscriber and no network specific, additional information is provided.

Each Event Block was transmitted from the mobile network operator's computer network to a receiving server over the Internet using the User Datagram Protocol (UDP) protocol. The connection was established by setting a virtual private network up which connected the mobile network operator's network with the network of the university.

To a report several additional information are attached. This package is called "Event Block". An Event Block without its additional information is structured like shown in Table 4. The full Event Block is specified in documents of the mobile network operator[16].

Anonymous ID*	A byte sequence identifying a subscribing device.
Time Stamp	Two long integers representing the date and the time when the event occurred (with millisecond accuracy).
Coordinates**	Two double precision floats representing the latitude and the longitude where the event occurred. The coordinate system is World Geodetic System 1984 (WGS 84)
Data Source	A number indicating the source where the data come from. I.e. 2G network data, 3G network data and GPS Input***.
Event Type	A number indicating what event occurred. The events are explained in the following chapters.

Table 4: The structure of an example Event Block without unused information

*) For a whole day (0-24) the anonymous Id of a subscribing device is the same. This Id is related to the SIM card used, but it is not possible to find anything out of the subscriber/subscribing device behind the anonymous Id. In other words, the data of the subscribing device become pseudonymized.

**) The retrieval of a subscriber's position is explained in chapter 1.4.1.

***) The GPS Input source indicates events from special navigation devices.

Event types may be associated to several networks or domains. In this context they may originate from the BSS if a device is connected via GSM, from UTRAN if a device is connected via UMTS and from the GPRS (independent of the device connection). A more accurate association is to map events to their originating interfaces. These interfaces are the "A" interface (connecting the BSC of the BSS with the MSC in the core network), the Gb interface (connecting the BSC of the BSS with the SGSN in the core network) and the Iu interface (connecting the with the core network). Events from the Iu interface may be further distinguished whether they come from the connection of the RNC with the MSC (Iu CS events) or if they come from the connection of the RNC with the SGSN (Iu PS events).

1.4.1 Positions of Subscribers

The position of a subscriber is the spot where a person actually is located. *When the position is defined with coordinates the person can be found at the coordinates*. The location of a subscriber is an area where the subscriber is in. *When the location is defined with coordinates the person can be found nearby the coordinates*. Coordinates are latitude and longitude values. The coordinates in event blocks describe a location of the mobile device. The location does not need to be the actual position. The location reported is an estimated point. The points are estimated by knowing the cell site where the event occurred as well as the (directed) antenna used to communicate. Within the covered area the mobile operator estimated points with the most likely appearance of users. It is not specified in detail how the points have been estimated.

1.4.2 GPS Input

The third source of events is GPS input. The longitude and the latitude of these events is the actual position of a subscriber. These events are sent by special devices and do not originate from ordinary entities like mobile phones or SIM card modems.

2 Analyzes of the Data

Before any inferences on road traffic or statistical calculations are done the data was inspected. The data were validated if they are in the format of the specification.

Then the detected locations were analyzed for coverage and density. The events which caused a report were mapped to their origin in the network domains.

Finally all provided information in an event block is tested for errors and phenomena.

Basis for these results are records from a mobile network operator covering whole Austria from 21.02.2012 00:00 until 26.02.2012 24:00 and from 01.03.2012 00:00 until 06.03.2012 24:00. The days of week are given in Table 5 bellow.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
	21.02.2012	22.02.2012	23.02.2012	24.02.2012	25.02.2012	26.02.2012
05.03.2012	06.03.2012		01.03.2012	02.03.2012	03.03.2012	04.03.2012

Table 5: Days of week for the test data

The 21st of February was carnival. The 22nd of February was Ash-Wednesday. There was spring break for the states Styria and Upper Austria from 20th of February until 26th of February.

These results are of course no representative example for every day in a year, but they are what someone might call “a look into the box”. On most of the days some invalid data were reported. For the 24th of February and the 1st of March were no invalid data reported. But comparing several classes of events with the amount of events received on other days, incompleteness of the data for the 24th of February and the 1st of March cannot be excluded.

2.1 Methods for Analyses

2.1.1 Validation of the Data

Validation of the data covers verification if the received data are formatted according to the specifications of the mobile network provider and an analysis of the information in the data. Verification of the format is done in the steps:

1. Checking if the value of the latitude is greater or equal to zero degree and smaller or equal to 90°.
2. Checking if the value of the longitude is greater or equal to zero degree and smaller or equal to 180°.
3. Checking if the type code of the received event is one of the specified event codes.
4. Checking if the source type of the received event is one of the specified source type codes.

If all these checks are passed a valid event report is assumed. The information in the report is then checked.

2.1.2 Validation of the Information

Besides checking for formal errors, validation of the information is necessary too. Therefore the current event for each user found is compared to the previous one (as soon as two or more events were received for a user). The information is validated according these rules:

1. All events are assumed to be reported by the probes in the order they occur. Therefore the timestamp of two subsequent events is compared. The information fails the validation if the timestamp of the current event is smaller than the timestamp of the previous event (this would mean the current event occurred before the previous event and was therefore reported delayed).

2. More than one event may happen at the same time at the same location for the same user. But the information fails the validation if more than one event is reported on different locations at exactly the same time for the same user.

3. The location of occurrences over the time is supposed to reflect the movement of a person. The information fails the validation if the user had to move with impossible speed to reach the location where the event occurred. The speed threshold is 300 kM/h for the validations of this thesis. Whenever this validation fails there is a slight possibility for a false negative (e.g. caused by travelling by plane). A false negative may also occur if a user changes to a nearby location due to nervosity (more to this topic in chapter 4.1.2). Therefore there is also a minimum distance of 100 km required to fail the test.

2.1.3 Location and Position

Location and Position are an essential part of the information in an event block. They make it possible to map the event with a geographical position.

Danilo Valerio describes in his technical report “Road Traffic information from Cellular Network Signaling” how data from network probes are used to infer on road traffic[10]. The geographical information is generated by software and refined by data provided from the mobile network operator. Research in his thesis has a special focus on events occurring at location areas and routing areas. In chapter 5.2 “Opposite traffic flows” he inspects the border of two location areas to infer on the traffic flow from location area updates [10].

Johannes Schlaich uses so called “A-Daten” (data from the “A” interface) collected by network probes from a mobile network operator’s network in his doctor thesis. The data contain the Location Area Code and the id of the cell site. These data indicate Location Area updates or handovers of a call from one cell to another. There are no actual geographical positions known from the data, but the data are mapped to known Location Areas by the Location Area code [9].

In the data of this thesis positions and locations are reported in the same field of an event block. The mobile operator’s specification defines that for event blocks from the 2G and 3G network contain the *coordinates of cell sites* for “Option A” and the coordinates of the

subscriber's location for "Option B". Event blocks from GPS input contain the *actual position of a user* for both versions. In this thesis none of the geographical information is refined by default. It is used as provided in the data.

For 2G and 3G networks all possible locations can be retrieved counting all distinct coordinates. These locations are printed on a map. This map is matched the network coverage of the mobile phone operator [5].

2.1.4 Decoding Network Events

The reported events are specified in a company confidential document of the mobile network provider. This specification maps the byte code of the events to an event name but it does explain the events or the origin of the events any further. Therefore the true meaning of the events is not available

Basic information about GSM and GPRS can be found in literature and specification of the ETSI (http://www.etsi.org/WebSite/homepage.aspx - last accessed on 03. Feb. 2012). Basic information about UMTS can be found in literature and specification of the 3GPP (http://www.3gpp.org - last accessed on 03.Feb.2012).

Prior work with data from Austrian mobile network operators can be found in reports from Danilo Valerio[10] who briefs the events he is using in his projects.

2.2 Results of the Analyses

After reception of data for a day, validation of the data was verified as specified in chapter 2.1.1. For some days invalid data were found. The situation has been reported to the mobile network operator.

Invalid data received were always received for a certain time span. For instance all data received for half an hour violated the specifications. *This does not necessarily mean there are no valid events reported from these 30 minutes! It only means any data received during these 30 minutes are invalid*. It is still possible that events which occurred during these 30 minutes are reported delayed. In either ways it has to be expected that the *information for this time span is incomplete*.

For some days a whole class of events was not reported. This means the data are incomplete and certain event types are missing. I.e. there were no data reported from the 2G network at the 3rd of March. There were no Location Area updates received for the 2nd, the 3rd, the 4th and the 5th of March from the GSM network. The same situation is with reports about Handovers.

The validity of data says nothing about the completeness of the data.

Format errors were reported at 09:00 on 21st to 23rd February. 24.855.801 invalid events were reported on 21st, 1.435.992 on 22nd and 24.247.652 on 23rd. On 6th April there were invalid events from 09:00 to 16:00. On 12:00 and 15:00 there were over 10.000.000 invalid events each hour, on the other hours there were less than 20 invalid events for each hour.

Validating the information of the data means to check if the information describes situations which are impossible. During the validation some phenomena were detected. According to the information some users would have moved with supersonic speed or used their mobile device on different locations at the exactly same time. Neither the origin nor the cause has been fully investigated. Some of these phenomena have side effects on usefulness (i.e. for statistical analysis) of the data.

Validating the information also means to process 20 to 60 GB of data per day and keep track of the last event for each of the almost 5 million subscribers in the network.

Cells and their radio coverage are usually pictured as hexagons. When within coverage the mobile network operator knows the cell in which a user is as soon as his device triggers an event (like receiving or doing a call). In the event reports the locations of users are reported as latitude-longitude coordinates. The locations in the event reports are fix points. A user may actually be just nearby this point and not exactly at the point (as explained in chapter 1.4.1). Another conclusion is that cells have different coverage ranges. Therefore equal hexagons are a misleading picture somehow.

No explanation of the event types was provided by the mobile network operator, so their meaning had to be guessed. This was done by studying literature. The results of the event type meanings are therefore only assumptions.

2.2.1 Supersonic User Phenomenon

The term supersonic user describes subscribers who seem to move with supersonic speed. When looking at the data a following up event of a user occurs over hundred kilometer away but just about one or two minutes later than the previous event.

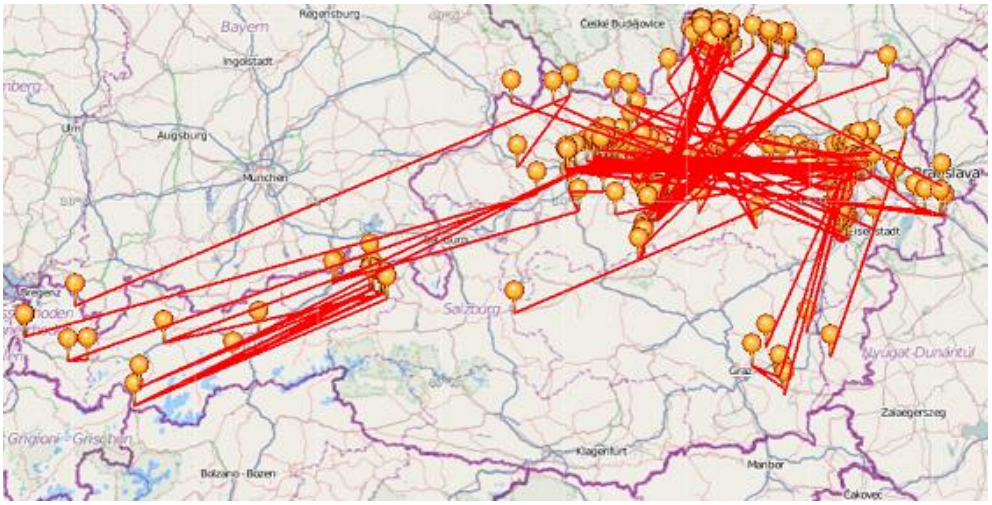


Figure 4: Supersonic users found on 21 and 22.02.2012 around 23:00 and 00:00 of the next day.

In Figure 4 supersonic users from 21st and 22nd of February 2012 are displayed. The orange pins are the locations of the events. The red lines connect these pins if one event follows another. The time frame of the data was only one hour. So it should be impossible for a subscriber to cause one event in Vienna and another one in Salzburg for instance. The possibility of anachronisms (described in chapter 2.2.2) has been considered when this image was made.

In Figure 5 a subscriber was chosen to put a focus on this phenomenon. The chosen subscriber starts his or her trip in St. Pölten at 23:08. According to the route of the trace and the speed the subscriber took the motorway A21 to Vienna. He or she arrived in Vienna at 23:53. On the way two events were triggered in Schlag nearby the Check border on 23:31.

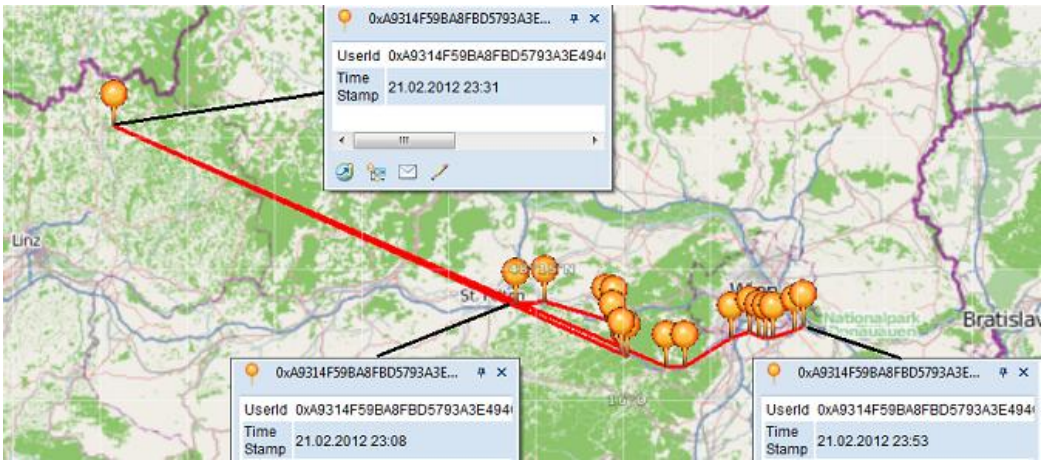


Figure 5: A trace of a subscriber who travels from St. Pölten to Vienna

The supersonic phenomenon has to be considered when trying to estimate the speed of the subscriber or when creating source-destination relations. It is obvious that the user visited Schlag during his or her journey. If the user was ever in Schlag on 21st or 22nd of Feb. 2012 is unknown.

In this thesis this phenomenon is detected by checking two subsequent events occur at different places A and B. If these places are too far away from each others to be reachable at normal (300kM per hour in this thesis) speed, the following event can be suspected to have an invalid location.

Four days were picked as examples to investigate the supersonic phenomenon into more details. In these days the average distance of supersonic events was about 116 kilo Meters. Table 6 shows the number of events occurred on the day (*#E*), the number of found supersonic events (*#SS*), the mean distance in kilo meters of these events (μ) and the standard deviation in kilo meters of the events (σ).

Date	*#E*	*#SS*	μ (in kM)	σ (in kM)
26. Feb. 2012	379538618	92365	100,808	0,174
01. Mar. 2012	548943507	44192	108,362	0,1
02. Mar. 2012	487911804	18367	105,544	0,272
03. Mar. 2012	379386524	15085	159,025	0,461

Table 6: Supersonic events statistics for four days showing the number of events (#E), the number of supersonic users found (#SS), the mean distance (μ) in kM and the standard deviation (σ) in kM.

2.2.2 Anachronism Phenomenon

Anachronisms occur if events are reported delayed and other events are reported in the mean time. Normally events are reported in the order they occur. But sometimes they are not. This phenomenon has to be considered if sequences are analyzed or events should be saved in the order they occur.

Anachronisms were detected by simply comparing the timestamp of a previous event with the timestamp of the following event. If the timestamp of the following event was earlier than *d* seconds the arrival of the event is considered as anachronism.

Four days were picked as examples to investigate anachronisms. For the first analysis in this thesis the parameter *d* was set to 0 seconds (so no tolerance). The mean delay was six seconds with zero seconds standard deviation. All events were reported at least one time delayed. Table 7 shows the number of events occurred on the day (*#E*), the number of found anachronisms (*#A*), the mean delay in seconds of these events (μ) and the standard deviation in seconds of the events (σ). The majority of the delayed events are from the 2G network; mostly incoming and outgoing calls as well as handovers.

Date	*#E*	*#A*	μ (in sec)	σ (in sec)
26. Feb. 2012	379538618	32255495	1	0
01. Mar. 2012	548943507	38065317	7	0
02. Mar. 2012	487911804	948593	15	0,014
03. Mar. 2012	379386524	350546	1	0

Table 7: Anachronisms with no tolerance showing the number of events (#E), the number of anachronisms found (#A), the mean dilatation (μ) in seconds and the standard deviation (σ) in seconds

Table 8 shows anachronisms with one hour tolerance. While 0.0009 % to 8 % of the events were detected as anachronisms with no tolerance, events which arrive one hour late or even later are insignificantly small. Almost only incoming and outgoing calls and handovers from the 2G network are reported. A few (1 to 13) short message events, location updates or supplementary service usage is reported one hour late or later. For the 3rd March 2012, where no 2G network events were reported, no events had one or more hours delay.

Date	*#E*	*#A*	μ (in sec)	σ (in sec)
26. Feb. 2012	379538618	24606	46	0,083
01. Mar. 2012	548943507	9863	19,004	0,403
02. Mar. 2012	487911804	748	2	0,110
03. Mar. 2012	379386524	0	0	0

Table 8: Anachronisms with one hour tolerance showing the number of events (#E), the number of anachronisms found (#A), the mean dilatation (μ) in seconds and the standard deviation (σ) in seconds

It is still unanswered what happens with events which are received after 24:00 PM. The anonymous Ids are supposed to change every day at 00:00 AM. Since this process is unknown it cannot be guaranteed that these events have the same anonymous Ids as those which came in time.

2.2.3 Parallel Events Phenomenon

Parallel events occur if one device triggers more events at exactly the same time at different locations. This phenomenon is observable when a subscriber either does a call or answers a call. Sometimes parallel events occur for handover events too. It cannot be excluded for other events. The distance of both locations is usually less than 10 kilometers. A parallel event for a subscriber is shown in Figure 6.

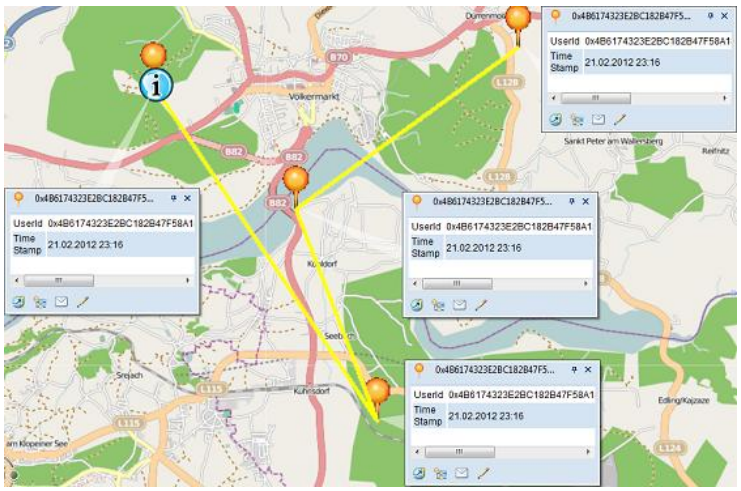


Figure 6: Parallel occurring events of a user in Völkermarkt at 23:16. The events were triggered by an incoming call

The Parallel Events phenomenon has to be considered when tracing a user's route. Since several locations are reported for a user at exactly the same time it is uncertain which location is the closest to the user's actual position.

In this thesis this phenomenon was detected by saving the last event of each subscriber and comparing the timestamp of the current event with the timestamp of the last event. If the timestamp of both events are the same (millisecond accuracy) but the locations are different the phenomenon is reported.

Table 9 shows absolute numbers of parallel events occurred on four days. *#E* is the number of total events for a day. *#N* is the number of parallel events. In columns with "*n*" are absolute numbers of occurrences with *n* parallel events at the same time. There are hardly more than 100 parallel events per day where more than five events happen at the same time for the same user. There is no significant preference for certain events types to be parallel. And the occurrence varies from day to day.

Date	*#E*	*#N*	*n=2*	*n=3*	*n=4*	*n=5*
26. Feb. 2012	379538618	263700	181372	55753	4321	126
01. Mar. 2012	548943507	543710	339938	143544	11365	901
02. Mar. 2012	487911804	121213	63199	34668	3705	661
03. Mar. 2012	379386524	93545	48389	26360	2740	548

Table 9: Absolute numbers of parallel events

2.2.4 Location and Positions

Figure 7 illustrates locations and their corresponding cell site. The locations are expected positions of users. How the expected positions are defined is not available. In this thesis it is assumed that the expected positions are the center of a signal cone. A cell site may have several (usually up to three) directed antennas. These directed antennas cover an area – usually 120° each antenna if three antennas are on one site. Depending on the type of the cell site the signal strength of the antennas may by different. Therefore cell site with huge signal strength cover bigger areas than those with smaller signal strength. Coverage of cell sites may be overlapping.

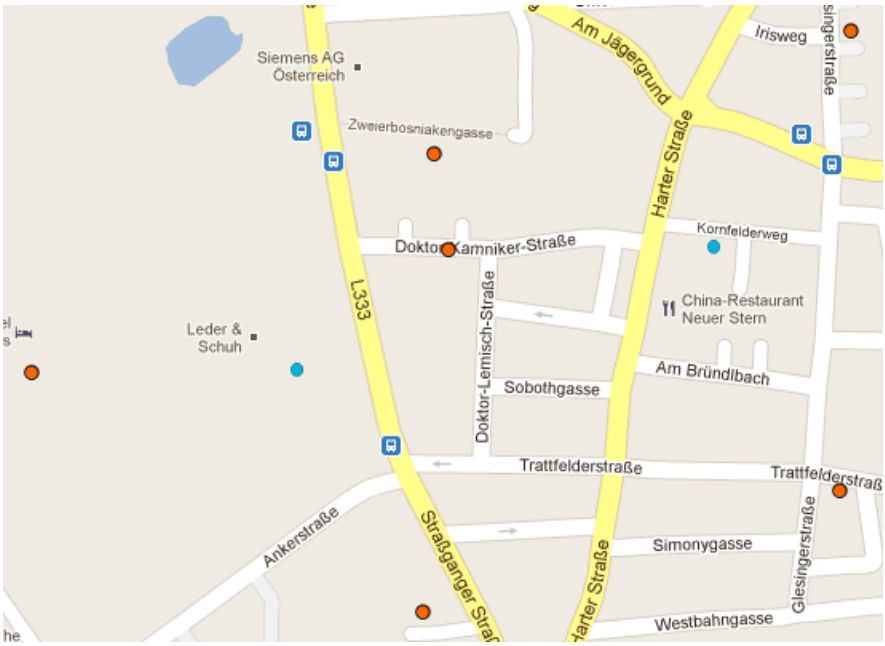


Figure 7: Two cell sites (blue dots) and expected positions of subscribers (orange dots)

During the sixteen inspected days there were 11450 distinct locations from 2G sources (11450 longitudes and 11297 latitudes) and 18502 distinct locations from 3G sources (18502 longitudes and 18122 latitudes).

In total 18677 different locations plus an invalid location (latitude and longitude are zero) were found for whole Austria. The total number of locations is smaller than the sum of 2G and 3G locations because a cell site may serve for 2G and 3G connections.

Since the positions from the GPS Input source are actual positions of subscribers the positions change accordingly.

In Figure 8 the 18677 different locations are painted as orange dots on the map of Austria.

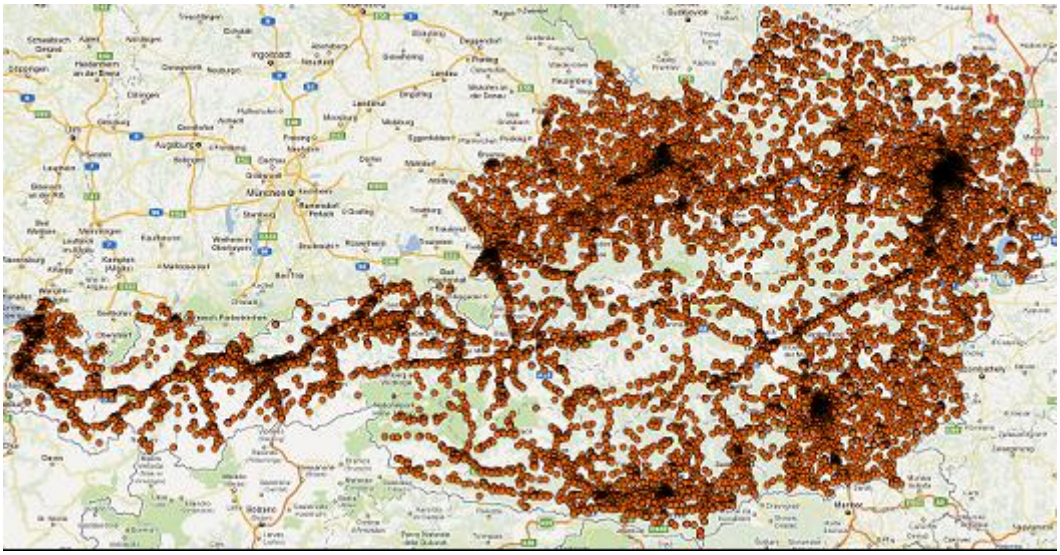


Figure 8: The 18678 locations (without GPS locations) displayed on Austria

The black patches among the orange dots indicate a very high density. These patches are especially visible in city areas like the area of Vienna, Graz, Linz and Salzburg. The domestic areas around these cities are also covered well while the alpine area is rather sparse.

Some motorways are also covered very well by the locations. On the picture the route S6 from Vienna to Judenburg as well as the A10 from Salzburg to Bischofshofen can be identified. When comparing Figure 8 with a map of Austria, which is except for routes blank, most of the other routes become visible.

2.2.5 Reported Events

The reports of the events are specified by the mobile network provider. They may be direct forwards of occurrences or a summary of events sent as one event. In the specification provided by the mobile network operator the event codes are mapped to event names. Neither the meaning of the code nor the meaning of the event name is explained any further. The results in this chapter are therefore assumptions and conclusions which are backed up by readings as explained in chapter 2.1.4.

2.2.5.1 Events from the GSM Network

The following events were reported for the GSM network. The events mobile originated and mobile terminated calls as well as using supplementary services were reported explicitly for the GSM network.

Detaching: A device is logging off from the GSM network. This may happen when the devices changes networks (i.e. a roaming user connects to a network with a better signal).

Short Message Service: A device either sends or receives a short message. From the GSM interface we do not know whether it was an incoming message or an outgoing one.

Terminating Calls: A call is reaching a device (the user is being called by someone).

Originating Calls: A call is outgoing from a device (the user is calling someone).

Location Updates: A user leaves a prior location area and enters a new location area. In this case his device sends a signal so the mobile network operator knows the new location area the user is now in.

Handovers: While the user is calling (a connection is established) the mobile device switches cell sites (most likely because the user is moving).

Emergency Calls: An emergency number is dialed. These are special numbers because usually every mobile device may dial them no matter if the device belongs to the network or not.

Supplementary Services: This term covers services like call redirection, holding a call, forwarding a call as well as advices of charge or other services. What services exactly are covered is mobile network provider depended. It was not specified by the provider providing the data for this thesis.

2.2.5.2 Events from the GPRS

Events reported from the GPRS may either originate from the GSM network or from the UMTS network. Which network they come from depends on the network the mobile device is subscribed to. As a matter of fact the GPRS related events reported in the data for this thesis originate all from the UMTS network. *Even events which are supposed to originate from the GSM network, like reports from the Gb interface, are reported to come from the UMTS network.*

Attaching: The mobile device contacts the SGSN and requests a so called context. After obtaining such a context an IP address is assigned to the mobile device. The attach process puts a device from idle state (not attached for GPRS) to ready state (just sent a packet and/or attached for GPRS). Such an attach may be done via the Gb interface or via the Iu interface.

Cell Updates: When a mobile device is in ready state and it detects a new cell it may perform a cell update.

Routing Area Updates: In general a mobile device updates its routing area when entering a new routing area. When the device performs a circuit switched operation (like doing a call) while changing the routing area it will update the new routing area after completing the operation.

Other Events: Which activities are covered by this event is not specified by the mobile network operator.

2.2.5.3 Events from the UMTS Network

Attaching to the GPRS: As stated in the previous chapter a mobile device may either attach via Gb interface if it is connected to the BSS network or via the Iu interface if it is connected to UTRAN. This event indicates an attaching of a mobile device to the GPRS via Iu.

Detaching from the GPRS: Mobile devices that attach to the GPRS over the Iu interface do not switch to idle state when detaching. They definitely detach. So this event indicates such a switch to detached state.

Detaching over CS: A detach request was received over the CS channel.

Paging: Paging messages via the circuit switched or the packet switched Iu interface are reported in two separate events. One event is for circuit switched messages; another one is for packet switched messages.

Temporary Logical Link Identifier: A new TLLI was assigned to a mobile device.

Temporary Mobile Station Identifier: A new TMSI was assigned to a mobile device.

Routing Area Updates: When a mobile device changes the routing area it reports to the SGSN. The routing area is then updated for the device. Such an update is reported when the mobile device performs an update because of moving. Periodic updates are reported too. Combined updates like routing and location area updates are also reported (in a separate event type).

Short Messages: Short messages sent or received over the circuit switched Iu interface are reported in two events. One event is for sent messages another is for received messages.

Setup: A connection for an outgoing call is requested.

Connection Acknowledge: A connection for an incoming call is accepted / acknowledged.

Disconnect: A connection for a call is being disconnected.

3 Statistical Analyses

The data are inspected for statistical moments to find event types or classes of events which seem to be promising for inferences on road traffic. As statistical model a Markov Model is created to discover sequences of events and movements.

3.1 Statistical Methods

Because of the large amount of data received each day (between 20 and 40 Giga Bytes) and the possibility to receive and process the data stream online performance is important Therefore optimal processing of these methods was desired.

If there is a finite set of n values the mean value can be calculated in one iteration, the variance is calculated in another iteration. The complexity therefore is $O(2n) = O(n)$. For an infinite set of values or the amount of values is not known the mean value and the standard deviation need to be estimated.

The **mean value** μ of a set of values X_n is defined as the sum of all values ($\sum x_n$ i.e. $x_1 + x_2 + \ldots + x_n$) divided by the amount of values n. So $\mu = (1/n) * \sum x_n$.

The **deviation** of a value is the difference of the value to the mean value ($x_n - \mu$).

The **variance** of a set of values X_n is the sum of the square of the deviations of all values divided by the amount of values: $Var(X) = (1/n) * \sum (x_n - \mu)^2$.

The **standard deviation** of a set of values X_n is the square root of the variance of the set $\sigma = sqrt(Var(X_n))$.

To calculate the mean value and the standard deviation of a set with an unknown amount of items, an online algorithm is needed. This algorithm computes these values for every new item in the set. The textbook formula ("computational formula for the variance" or "Satz von Steiner" in German) for calculating the mean value and the standard deviation online is $sqrt(1/(n-1) * [(\sum x_k^2) - (1/n) * \sum (x_k)^2])$; k is $1 \leq k \leq n$. This formula makes it neither necessary to store all values ($x_1 \ldots x_n$) nor to do iterations. The problem with the computational formula for the variance is its numerical instability which may lead to taking the square root of a negative number because of rounding errors.

A numerically stable algorithm is described by Donald Ervin Knuth [17]. The mean value is initiated with the first received value $\mu = x_1$. The sample is initiated with zero $S = 0$. For every k^{th} value the mean value is updated with $\mu_k = \mu_{k-1} + (x_{k-1} - \mu_k) / k$. The sample is updated with $S_k = S_{k-1} + (x_k - \mu_{k-1}) * (x_k - \mu_k)$. At every moment the standard deviation can be retrieved with $\sigma = sqrt(S_n / (n-1))$ where n is the actual amount of values. Tony F. Chan et al. give a parallel algorithm in one their reports [18]. This algorithm has been used to do statistical analyzes on phenomena as described in chapter 2.2.

3.1.1 Correlation Coefficient

The correlation coefficient measures the degree of correlation between two data sets. The Pearson product-moment correlation coefficient gives a correlation coefficient (from 1 to -1)

for linear depended data. The coefficient then refers to the linear dependency of both data sets. The closer the coefficient is at zero, the loser is the linear dependency of the data. The closer the coefficient is either at 1 or at -1 the higher is the linear dependency. E.g. there are two sets of data *A* and *B*. Both sets can be plotted as a graph. If the correlation coefficient of the sets is 1 then the graph for set *B* increases when the graph for set *A* increases. If the coefficient is -1 then the graph for set *B* decreases when the graph of set *A* increases.

The Pearson product-moment correlation requires the data to be normal distributed, scaled and linear depended.

The Pearson product-moment correlation coefficient ρ of two data sets *A* and *B* is the co-variance of the sets *Cov(A, B)* divided by the product of the standard deviations σ_A and σ_B of the two sets as shown in Table 10.

$$\rho = \frac{Cov(A,B)}{\sigma_A * \sigma_B}$$

Table 10: The Pearson product-moment correlation coefficient formula

In this thesis the Pearson product-moment correlation coefficient is used to check the linear dependency of statistical results. The coefficient does not explain the reason for the dependency or the direction. So if there is a strong positive correlation between two sets *A* and *B* it only means the set *A* describes *B* and vice versa. It does not necessarily mean because *A* increased also *B* increased.

There is no standard scale of strength and weakness of a correlation coefficient. Rossitza Setchi gives a scale in his book "Knowledge-based and Intelligent Information and Engineering Systems"[19] where coefficients about 0 indicate a weak correlation, coefficients about ±0.5 indicate a moderate correlation and coefficients about ±1 indicate a strong correlation. Timothy C. Urdan suggests in his book "Statistics In Plain English" [20] to take such scales with "a grain of salt". In other words the interpretation of the coefficient depends on the data and the case. He advices once again that a correlation does not mean one causes the other.

The actual correlation coefficient was calculated using Microsoft Excell.

3.1.2 Markov Model

The sequence of events and the movements of users from one location to another are modeled in Markov chains. A Markov chain has states *S* and transitions from one state to another *T*. Further there is an initial probability for each state and a probability for each transition to happen. In case of the reported events, the states are the event types. The initial probability for each state is the likelihood that an event type is the first event reported for a user in the evaluated set of data. The transition probability is the likelihood that one state follows another in the evaluated set of data.

Both, the initial state probability and the transition probability are expressed as $n \times n$ matrices (where n is the number of events). The initial state probability P_x for each event x is the number of users for which the reported event is the first event $\#s_x$, divided by the count of all reported events *#events*: $P_x = \#s_x / \#events$. The likelihood of the x^{th} event is then inserted in the cell of the x^{th} column in the x^{th} row. The probability of the transition T_{xy} of each event x to any other event y is the number of subsequent occurrences of the two events *#(event x followed by event y)*, divided by the number of total subsequent occurrences *#(event x followed by other events)*. T_{xy} = *#(event x followed by event y) / #(event x followed by other events)*. The sum of all likelihoods that event x is followed by another event (T_x) is the likelihood that an event takes part in a sequence. On the other hand, the difference of $1\text{-}T_x$ is the likelihood that the event is not followed by an event and therefore it is either the last event of sequences or a stand-alone event. When the likelihood is inserted into a matrix, the first event is in the row and the subsequent events are found in the columns.

In a similar fashion the Markov model is created for locations and location changes. Each location is a state. And each location has a likelihood to be the first location for users. This likelihood is the initial state probability. Further there are probabilities for each location that a user may move from it to another. This is the transition probability.

In this basic model the time between the events or locations has not been respected! Therefore subsequent events where one event occurs hours after another event is treated equally as an event that occurs only seconds after another. This is the same for changes of the location.

Location changes of every cause are treated equally. There is no difference made between location changes caused by movement and location changes caused by anything else.

Events have to be checked for phenomena as described in chapter 2.1.2. Anachronisms may have a heavy impact on the probabilities. Delayed events alter the sequences if not ordered according to the occurrence time.

3.2 Results from Statistics

The amount of data received per hour varies. In general during night time there are fewer events reported - most likely because fewer events occur - and during the day time more events are reported. The data received were saved in separate files for each hour. Figure 9 shows a graph of the amounts. The blue lines show the days Thursday, Friday, Monday and Tuesday during the week. The green lines show the weekend days Saturday and Sunday. The red line is a graph of the mean value. The actual numbers to this figure as well as to all other figures of this chapter can be found in the appendix.

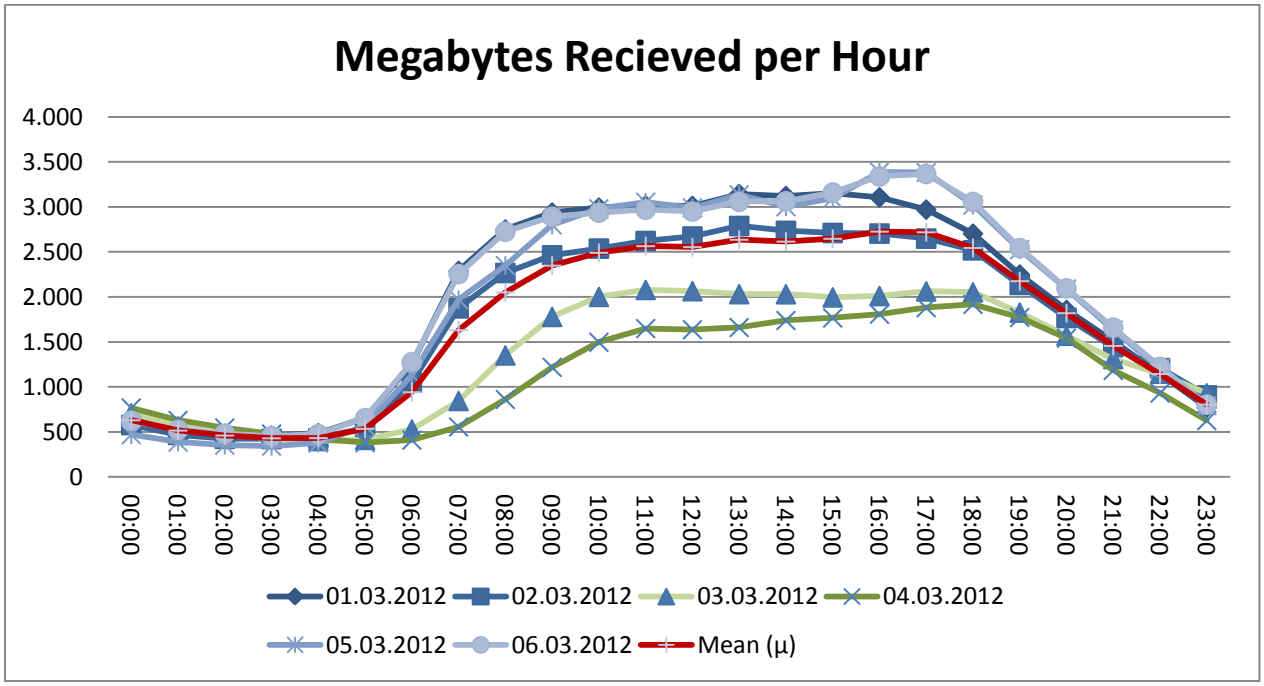


Figure 9: Amount of data in Megabytes (MB) received per hour

The characteristic results of this graph are:

4. There are more events on days during the week than on the weekend.
5. There is a significant increase of events at about 05:00 AM on days during the week compared to the hours from 00:00 AM to 05:00 AM. On weekends the events start to increase between 06:00 AM and 07:00 AM.
6. The increase throttles at about 09:00 AM on days during the week and on 10:00 AM on days at weekend.
7. There is a significant decline of events at about 17:00 PM on days during the week and on weekends compared to the hours from 09:00 AM to 17:00 PM.

The time from 00:00 AM until 01:00 AM starts with about 630 MB on average. There is a slight but steady decline in the graph from 00:00 AM to 04:00 AM. Between 03:00 AM and 04:00 AM is an inflection point. The graph starts a sharp incline at 05:00 AM. In numbers about 540 MB are received from 05:00 AM to 06:00 AM.

There is sharp incline from 06:00 AM to 09:00. In numbers about 940 MB are received from 06:00 AM until 07:00 AM and 2.000 MB are received from 08:00 AM until 09:00 AM.

From 09:00 AM to 18:00 PM there is a plateau with a slight increase. From 09:00 AM until 10:00 AM there are 2.350 MB received on average. From 16:00 PM to 17:00 PM there is about 2.730 MB received on average. Then there is another inflection point at about 17:00 PM.

The average transfer rates start with 2.720 MB from 17:00 PM to 18:00 PM and end with 800 MB from 23:00 PM until 24:00 PM.

Considering the graph and considering common working hours the day can be divided into four times: Night time, morning time, day time and evening time. The time from 00:00 AM to 06:00 AM will be referred as **night time**. From 06:00 AM to 09:00 AM will be the **morning time**. **Day time** refers to the time from 09:00 AM to 18:00 PM. **Evening time** refers the time from 17:00 PM to 24:00 PM.

For comparison Figure 10 shows the number of unique devices found per hour on the days. In general the lines match with the lines from Figure 9. Valleys in the graphs (i.e. on 06.03.2012) are a result of invalid data received for this time. The green lines are days on a weekend.

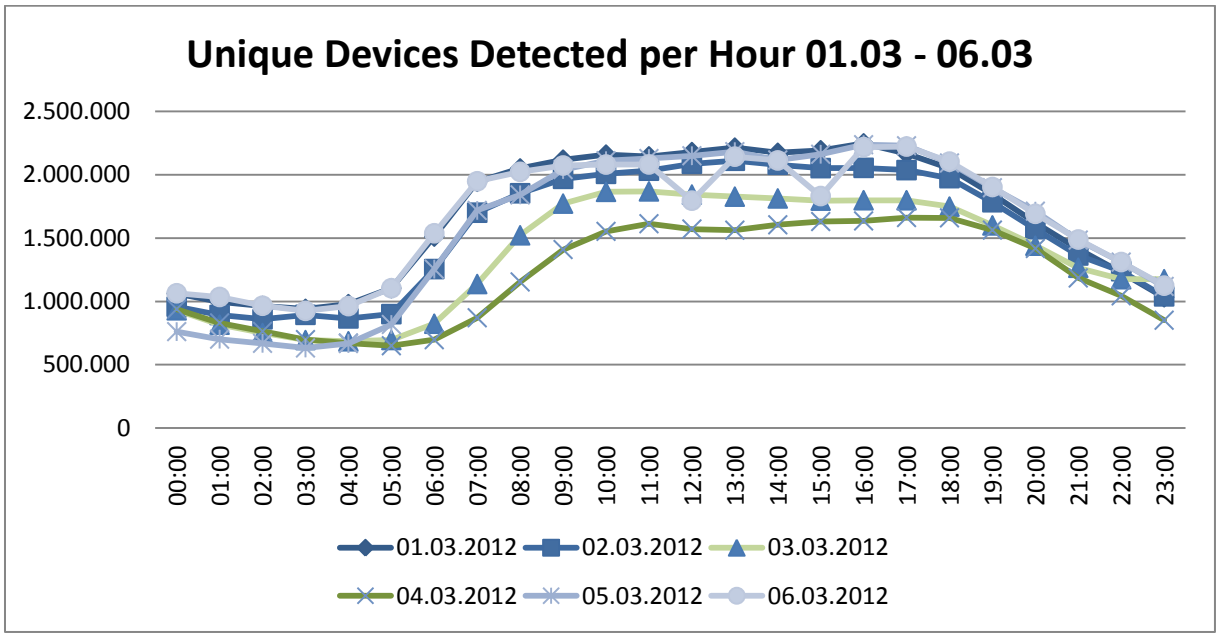


Figure 10: Unique devices found per hour on all days from 01.03.2012 until 06.03.2012

The received MB per hour and the detected devices per hour have a correlation coefficient of 0.9779. So, over the day times the number of events triggered by one device changes in the same fashion as the number of devices does.

The trend of detected devices is the same when looking at the days from 21.02.2012 to 26.02.2012 in Figure 11. The valley at 09:00 is again a result of invalid data received for this hour on 21st, 22nd and 23rd February.

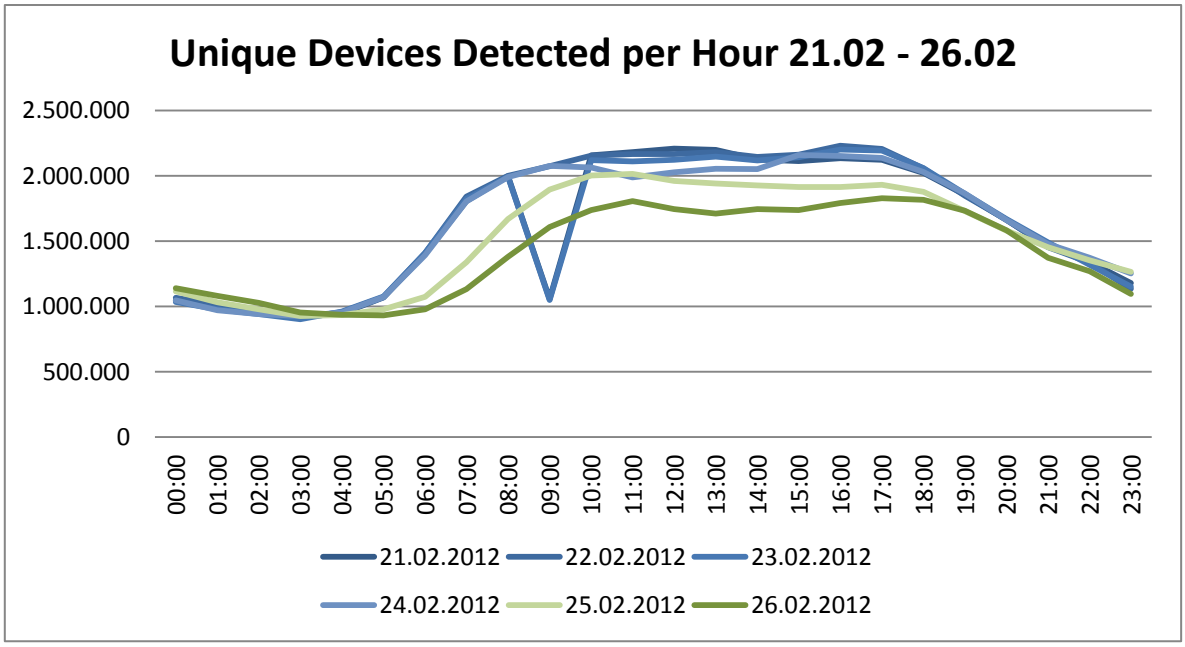


Figure 11: Unique devices found per hour on all days from 21.02.2012 until 26.02.2012

The unique devices found per hour on all days from 23.06 until 26.02 correlate with those devices found per hour on all days from 01.03 until 04.03 with a coefficient of 0,9492 (weekdays correlate with the corresponding day of week).

Figure 12 shows the mean value of Kilobytes (kb) received per device every hour of the day (blue) and the standard deviation in kb (red). As a matter of fact the graph also expresses the amount of events triggered by a device (considering that each event is reported in 88 bytes).

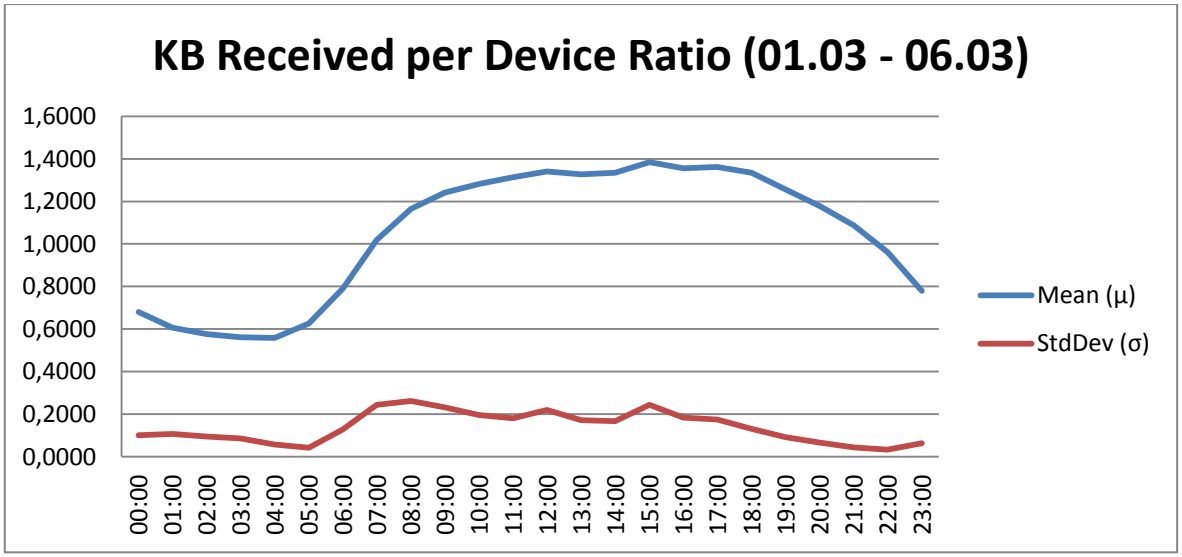


Figure 12: The ratio of megabytes received per device over the day

The peaks in the standard deviation at 12:00 and at 15:00 are a result of the invalid data received on 06.03.2012 for these hours. The graph shows that the amount of data received

does not only increase because there are more devices detected during the day time. *There are also more events triggered by each device during the day time.*

In the morning time (from 21.02 until 26.02) each device caused between 5 and 8 events. In the morning time (from 21.02 until 26.02) the events increase. During the week each device caused between 9 and 12 events. On the weekend (25.02 and 26.02) the number is not increasing. During the day time (from 21.02 until 26.02) about 16 events are sent per device during the week. On weekends the number is about 13. In the evening time (from 21.02 until 26.02) the number of events is declining each hour until it reaches about 8 at 24:00. Every hour the standard deviation of triggered events is about 5 to 7 events. This trend is similar for the time from 01.03 to 06.03. Actual numbers can be found in the appendix.

Among all events there are certain event types which are of special interest when trying to infer on road traffic. At first there are Location Area updates and cell handovers because these events are connected to movements as explained in chapter 1.3.

3.2.1 Location Area Updates

A Location Area update is triggered when a mobile device is turned on (no matter if it is idle or active) and it enters a new Location Area. Therefore the ratio of reported Location Area updates to active devices give a degree of subscribers' mobility. The degree does not tell anything about the number of users in the observed area. There are periodical updates, but the time span between the updates is too large (about 6 hours) to be significant. The observed area must be large enough. It has to cover at least two Location Areas so the updates can happen.

Figure 13 and Figure 14 show the number of Location Area updates each hour of the day from 21.02.2012 until 26.02.2012 for whole Austria.

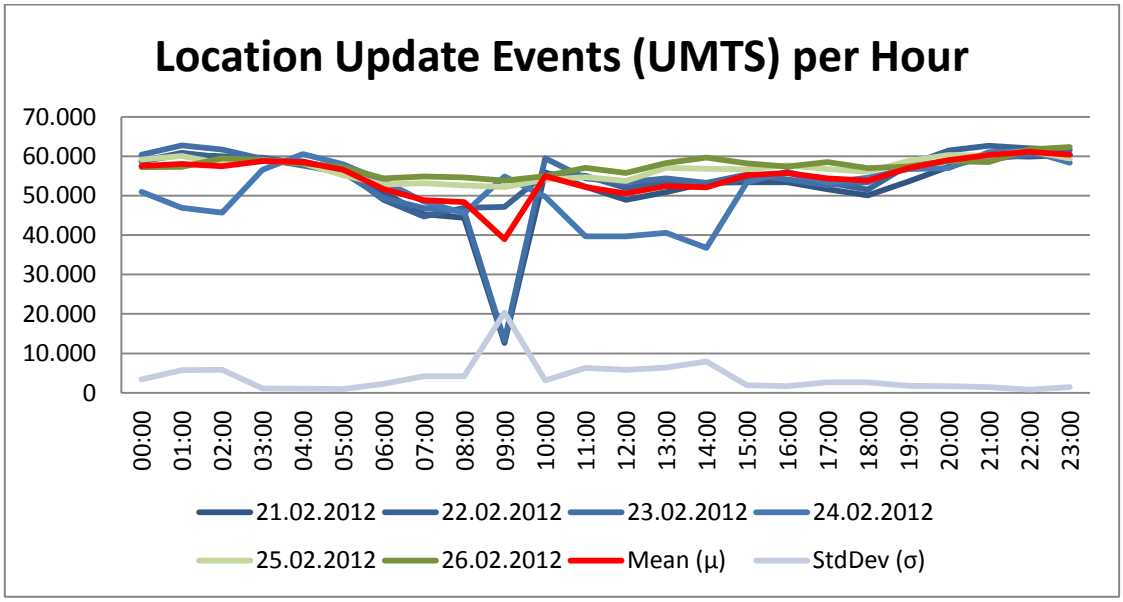


Figure 13: Location Area updates reported from the UMTS domain from 21.02.2012 until 26.02.2012

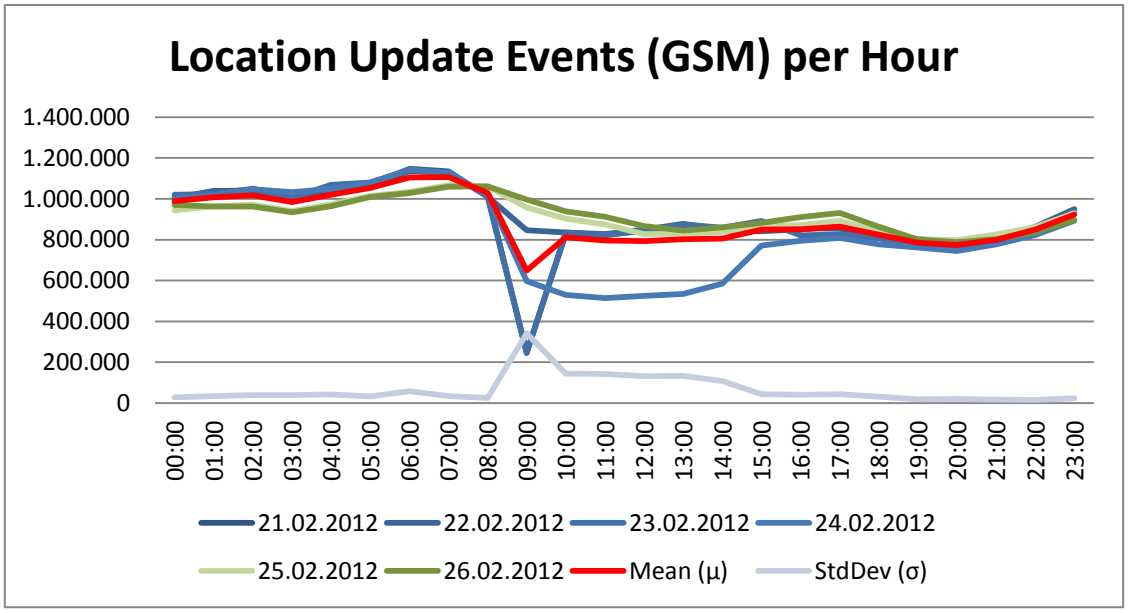


Figure 14: Location Area updates reported from the GSM domain from 21.02.2012 until 26.02.2012

The graphs do not follow the trend of received devices or amounts of data received. The amounts of Location Area updates vary in a small range (compared to the ranges of previous graphs). For updates reported from the UMTS network it is between 50.000 and 60.000 updates. For updates reported from the GSM network it is between 800.000 and 1.200.000 updates. The data of the UMTS network correlate with data of the GSM network with coefficient of only 0.4369.

The graph of Location Area updates from the GSM domain is flat in the Night time and slightly increases in the Morning time. After the morning time there is a significant decrease in the beginning of the Day time. During the day time the graph is again flat. The number of Location Area updates rises again in the later Evening time.

The reason for the increase in the Morning time could be the situation that many people are on the way to their working places. The small peak in the Evening time might be caused by people who are on the way back home. The increase short before midnight could indicate moving people like trucks. Unfortunately the real reason cannot be determined by the data themselves. But since Location Area updates are caused by long distance movements a higher number of these updates feature an increased mobility rate.

There is no increase in the morning time of the graph of Location Area updates from the UMTS domain. And in general the number of Location Area updates in the UMTS domain is rather constant compared to the number in the GSM domain.

A possible explanation could be the mobile network infrastructure. If more UMTS cells are in urban areas or areas with little long distance movements (compared to GSM covered areas) this might explain the constant number of Location Area updates. The only common trend is the increase of Location Area updates in the Evening time. The number increases a few hours earlier than in the GSM graph.

3.2.2 Handovers

Handovers are triggered when a user is on a call and his connection is handed from one cell over to another. This may have different causes. One (and for this thesis interesting) cause is a user moved.

Figure 15 shows the graph of handovers each hour from 21.02.2012 until 26.02.2012. Figure 16 shows the graph of handovers each hour from 01.03.2012 until 06.03.2012. Since handovers are only reported from the GSM domain there is a big gap in the time from 01.03.2012 until 06.03.2012. Still the peak of the graphs in the evening time can be seen.

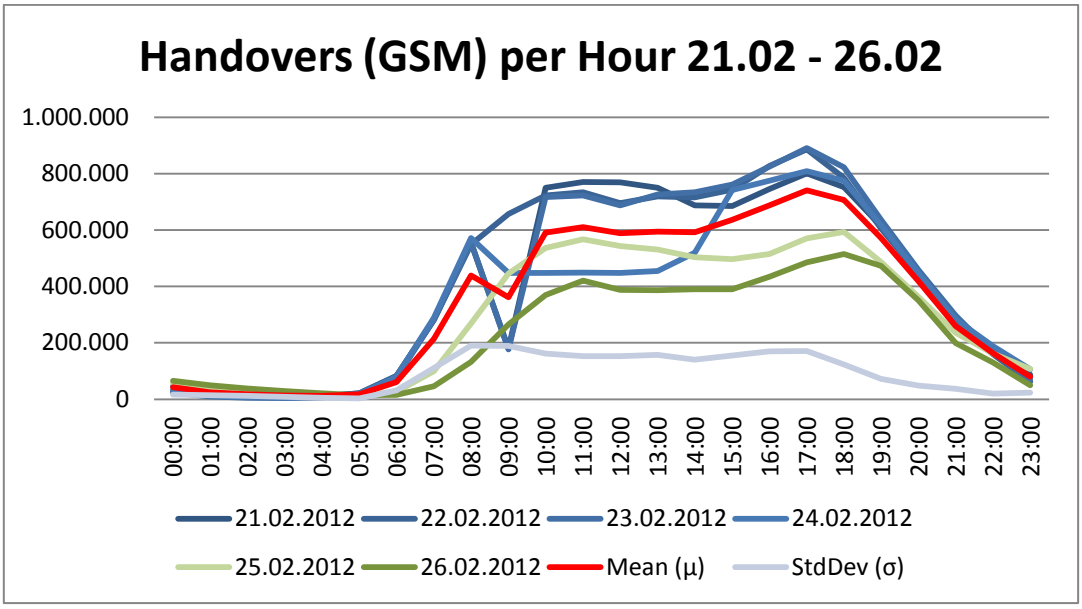


Figure 15: Handovers per hour from 21.02.2012 until 26.02.2012

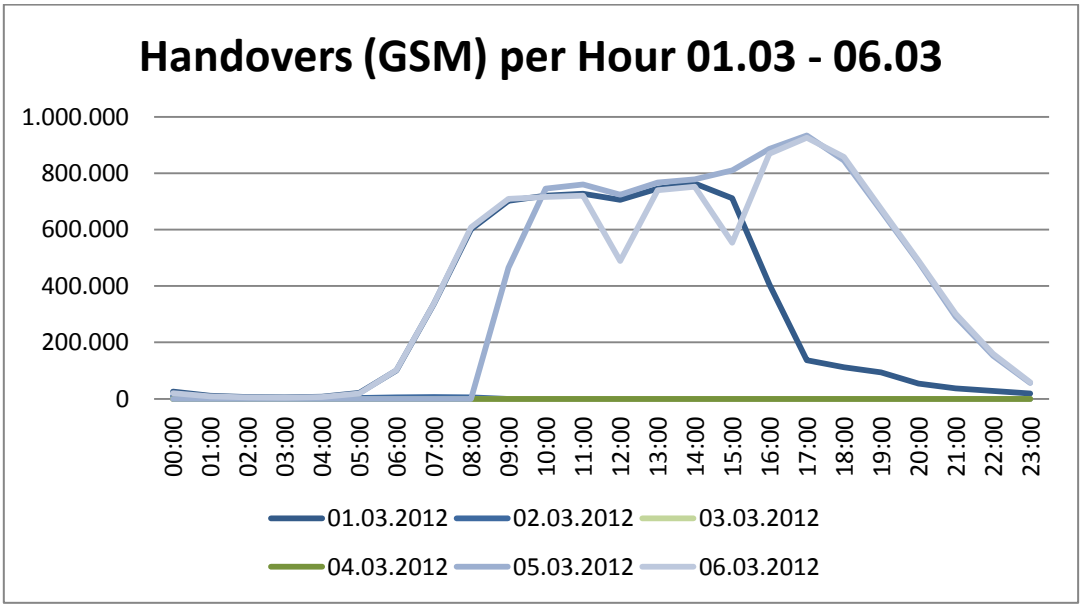


Figure 16: Handovers per hour from 01.03.2012 until 06.03.2012

3.2.3 Statistics Summary

To sum the results up, there are significant changes in the amount of data received each hour of the day and in the amount of devices detected each day. Both, the amount of data and the number of devices detected devices show a similar trend for each hour (correlation coefficient > 0.9). There are not more data just because there are more devices. Each device seems to send more data in the morning and during the day and fewer data in the evening and during the night.

Because of the trend changes in the amount of received data and the number of detected devices the hours of the day are allocated to four day times: Night time, Morning time, Day time and Evening time. The allocation is according to the trend changes of the graphs (like increases, plateaus and declines).

A different people's behavior can be assumed for each day time. Also a different people's mobility can be assumed. In the Night time most people won't move and sleep. In the Morning time most people will be awake and move. During the day time even more people will be awake. In the evening time more people will move than during the day time but this mobility will decline as it becomes later. These changes should become better visible when looking at the sequences of triggered events and to the location changes.

3.3 Markov Models

Markov Models are stochastic models of a system where the Markov property is assumed. This means the stochastic process has no memory (in other words the current transition probabilities do not depend on previous states). Using Markov Models it is possible to model sequences as transitions of states [21].

The Markov Models were created for each of the day times described in chapter 3.2. Four Markov Models show events sequences of whole Austria at each day time. Another four models show likelihoods of changes from one location to another in the city of Graz. The Markov Model for location changes was limited to Graz because of the large amount of locations and their changes to each other.

It was assumed people have a different behavior during the different day times from chapter 3.2. During night time most people are supposed to be asleep and not using their mobile phones. In the morning people are supposed to be moving and slightly using their phones. During the day time most people are supposed to be moving and using their phones. In the evening time people are supposed to be moving again. For each day time, the data over the time span were considered and analyzed according to the method described in chapter 3.1.2.

To see if the expected behavior can be described by the data only one day was chosen. Data source for the models are data from Thursday the 1st March 2012. The Markov Model of sequences has to be seen in relation to absolute numbers of event occurrences. Further the interpretation of the sequence model is difficult because the true meaning of the events is unknown. Still an interpretation of the sequence analyses can be found in the appendix (chapter 8.1). The next chapter will only present the Markov Model of location changes since these results were used to create the Road Surveillance Algorithm of chapter 5.

3.3.1 Markov Model of Location Changes

To create a Markov model of location changes a radius of 12 kilo meters around Graz (47.67° as latitude and 15.43° longitude for the center) was observed for event occurrences from the 2G and 3G network over the four day times during one day. The result is two matrices with all possible locations within the radius as rows and columns. One matrix holds the initial probability that a user will show up the first time during a day time. The other matrix holds the transition probabilities that a user “moves” from one location to another. Actually it is only assumed the user moves. As a matter of fact it is only known that the user triggers an event at one location and then triggers an event at another location.

In either ways the model was supposed to show movement preferences independent of the number of users. The model yielded the expected results and in addition further features were discovered.

Figure 17 shows the likelihood of several locations to find a new subscriber during the night time of the 1st March 2012. Great red circles show a high likelihood. Orange medium circles an average likelihood and green small circles a small likelihood. Most orange and red spots can be found in the north of Graz, while in the South there are mostly white and yellow spots.

Most spots in the east are green or yellow while they rather red in the west. In the west of Graz there is the A9 road way leading through the Plabutsch tunnel (with no radio access inside the tunnel). Parallel to the tunnel is the L333 as well as the B67 (Wienerstraße). The L333 is connected to the L301 which goes from Thal to Graz. The Wienerstraße has an exit from and a ramp to the A9 motorway. It has a fork to B67a (Grabenstraße) at northern Graz.

As stated before, the model shows the probability for a location to find a new subscriber. In other words the subscriber appeared the first time at this location and not at any other location before (over the observed time span). Red spots therefore mean at this location are many new users while white or yellow spots mean there are no new users. Spots do not say anything about the amount of users. A spot with a few but distinct users will always be darker than a spot with a lot of users which already appeared somewhere else before. The result of the night time therefore suggest there are many users moving from the north and west of Graz, coming from the A9 to the south or to the east.

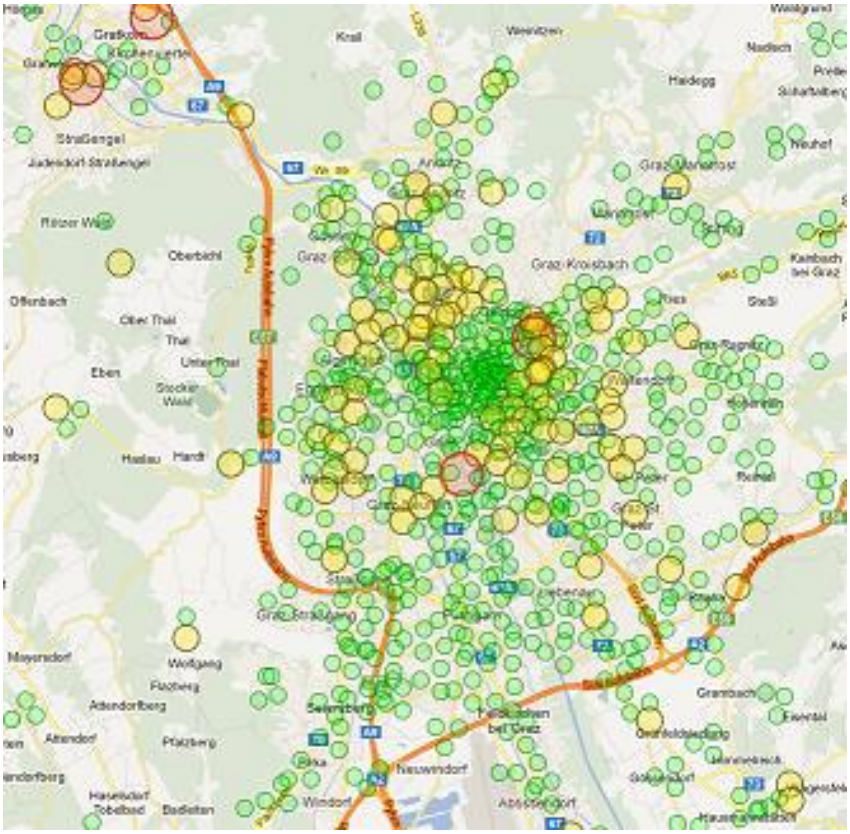


Figure 17: Initial probabilities of locations during night time for the 1st March 2012

Figure 18 shows the transition probabilities during the night time of the 1st March 2012. The deeper red an edge is the higher is the probability that one location leads to another. The edges on the map are undirected although the values in the matrix yield directed edges.

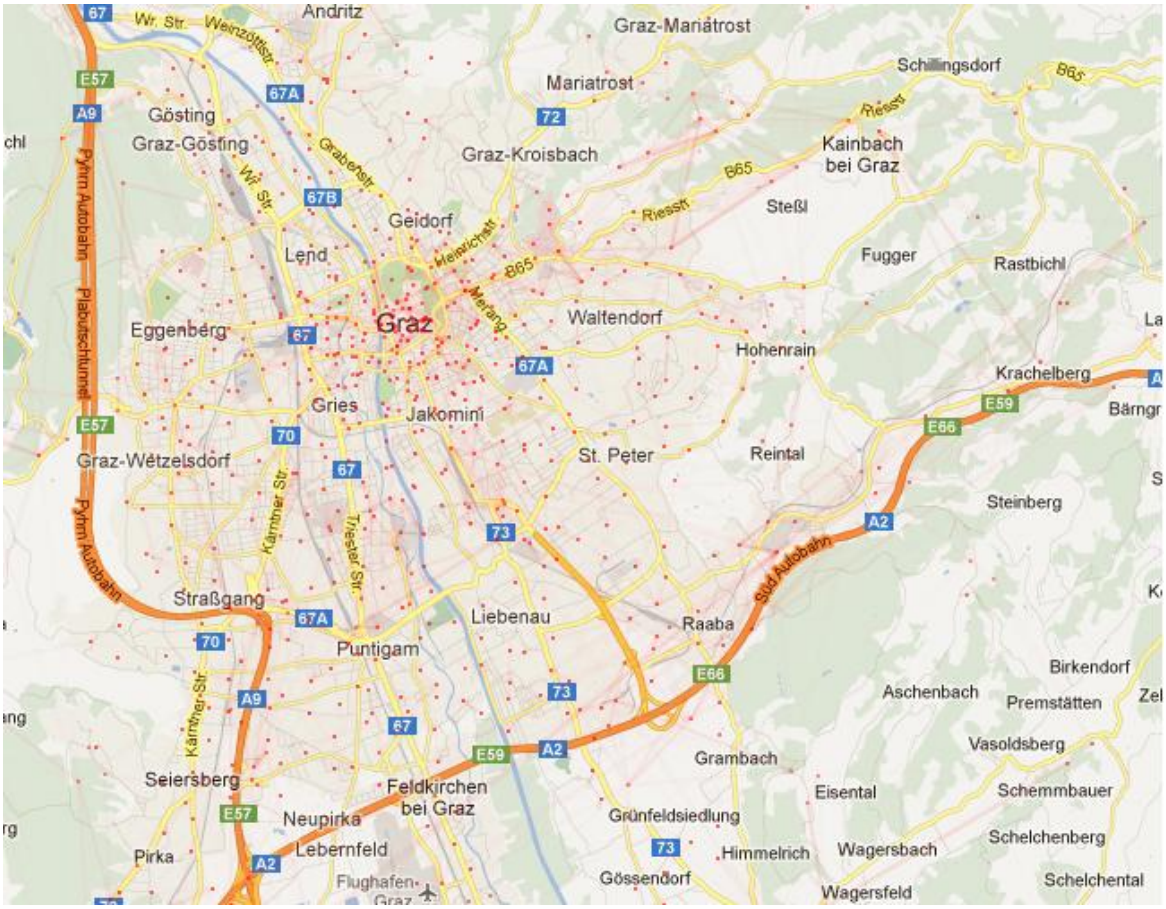


Figure 18: Transition probabilities of locations during night time for the 1st March 2012

The color of the edges is very opaque. Therefore it seems like there is little mobility. Further of interest is the length of the edges. The distances between two events are often very big. When looking at the road way A2 (Südautobahn) distances reach from Raaba to Kachelberg and longer.

Figure 19 shows the transitions for the day time. The first difference of Figure 19 compared to Figure 18 is the color of lines and the density of lines. During the day time, where more subscribers are around, there are more transitions for every location to other locations. When looking again at the road way A2 the road way now seems more like an opaque red stripe. A similar intense can be seen at B65 (Riesstraße). At the north eastern end of the B65 at Schillingsdorf there were almost no transitions between the locations there during night time. During the day time the transitions between these locations can now be clearly seen. This result features the assumption that there was less traffic in the north east of Graz during night time than during day time for this day. Further of interest are transitions at both ends of the Plabutsch tunnel. The Pyhrn Autobahn (motorway A9) has the form of an L when looking at

Graz-Gösting and Graz-Strassgang. The transitions from one end to another (one end at Graz-Gösting, another at Graz-Strassgang) gained intense during day time.

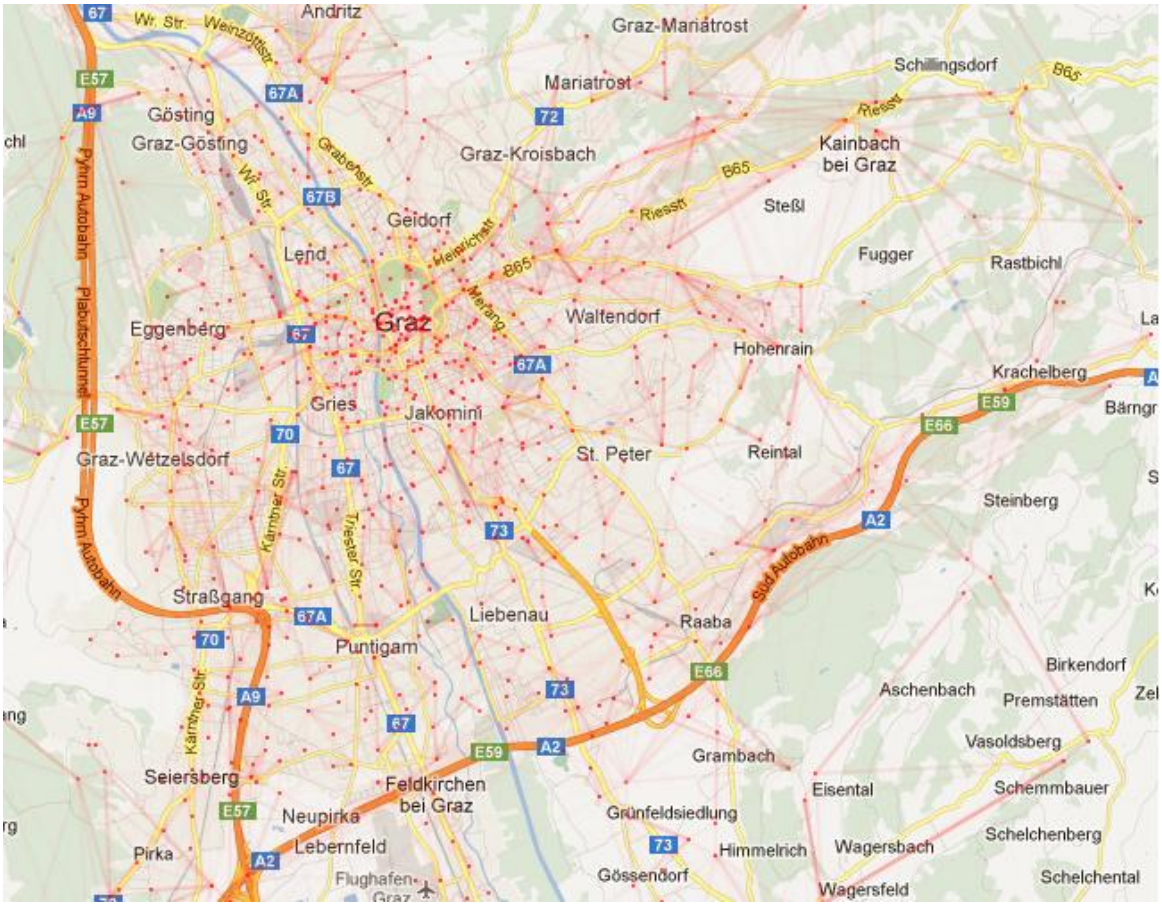


Figure 19: Transition probabilities of locations during day time for the 1st March 2012

In Figure 20 are the likelihoods for all locations to see a new subscriber during the day time. There are now more new users directly in the city center. The chance to see new users also increase at the south of Graz on the B73 (Liebenauer-Hauptstraße). The likelihoods decreased for the east of Graz (Graz-Gösting to Graz-Wetzelsdorf) on the L333.

In contrast to the likelihoods Figure 21 shows the accumulations in percent (number of users at a location divided by the number of all users) over the day time. Looking at B67 in north of Graz it can be seen there are fewer users than in the city center for example. Another contrast can be seen when looking at the districts Graz-Wetzelsdorf, Graz-Eggenberg and Graz-Gösting. The probability to see new users is higher (in percents) than the accumulation of users (also in percents). On the other hand the images are similar when looking at the eastern center of Graz.

In either case locations with a high probability to see new devices are either locations where people accumulate or they are nearby such a location.

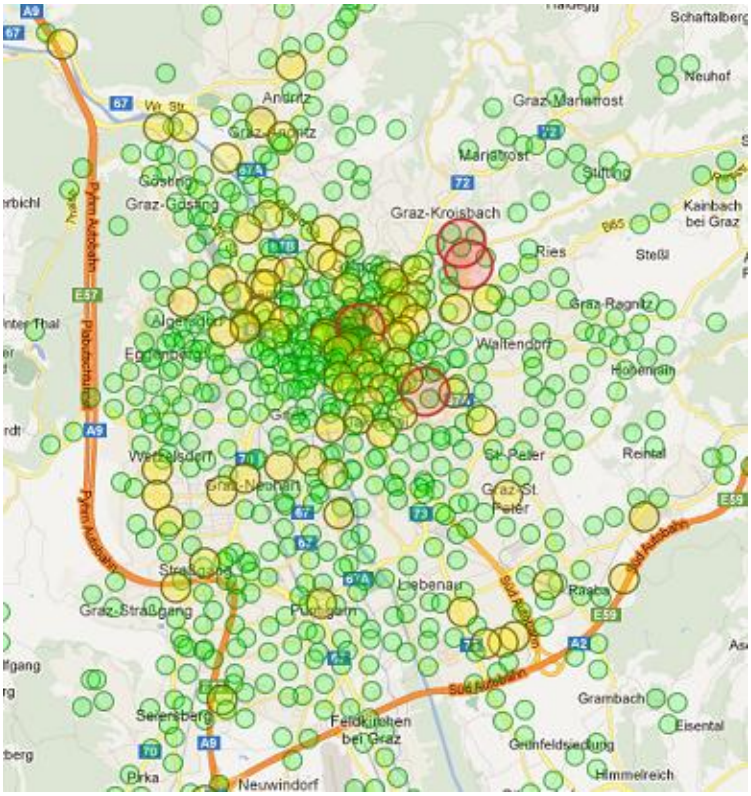


Figure 20: Initial probabilities of locations during day time for the 1st March 2012

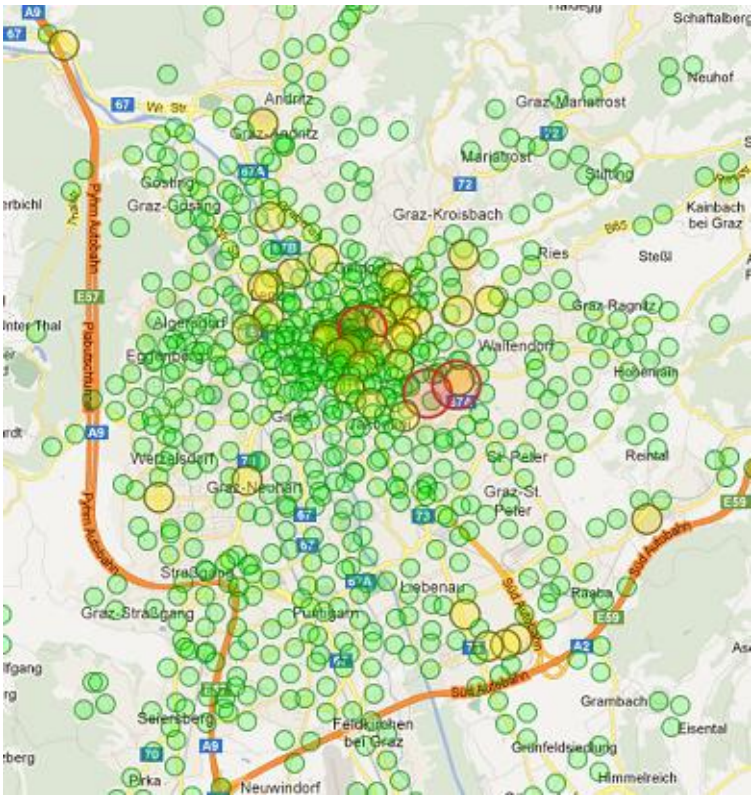


Figure 21: Accumulations of users at locations during day time for the 1st March 2012

3.3.2 Markov Model Summary

The results of the Markov Models (both, the model for location changes in the chapter above and the model for sequences in the appendix) can be summarized to the following bottom lines. The models were only created for one day as well as for one location. The sequence model was created for whole Austria; the location change model was created for Graz. Inspecting only one day is too little to infer on behavior as well as inspecting only one location is too little to evaluate the suitability of the model to infer on traffic. Still the model for location changes contributed to the nervosity/circle detection algorithm and the road surveillance algorithm of the next chapters. From the model it can be assumed that if more people go into a certain direction the transition probabilities for this direction will increase. This assumption is exploited in the road surveillance algorithm.

4 Inferring on Road Traffic

4.1 Inferences Methods

4.1.1 Geographical Algorithms

4.1.1.1 Distance Algorithms

In a flat Euclidian plane the distance between two points P_1 and P_2 can be calculated by the Pythagorean Theorem. This distance is also called second-norm distance or Euclidian Distance.

In a curved plane the curvature has to be respected for an accurate result of the distance. Further if the plane is the surface of a sphere or an ellipsoid the plane is a closed manifold and two points have two distances. The shortest distance of two points on a sphere is called **orthodromic distance** or **great-circle distance**. A widely spread formula to calculate the orthodromic distance of two points defined by latitude and longitude is the **Haversine formula** [22]. The formula needs the Earth's radius *R* (in the desired output dimension – i.e. kilo meters) and the latitudes *la* and longitudes *lo* of two points: $P_1(la_1, lo_1)$ and $P_2(la_2, lo_2)$. The distance *D* is then:

$$D = 2 * R * \arcsin\left(\sin^2((la_1 - la_2) / 2) + \cos(la_1) * \cos(la_2) * \sin^2((lo_1 - lo_2) / 2) \right)$$

Table 11: The Haversine formula

An example implementation can be found in the Microsoft's article "Bounding Box, Radius, and Polygon Search"[23]. If speed is of essence and accuracy may be neglected the Pythagorean Theorem is sufficient to estimate distances between two points.

4.1.1.2 Initial Bearing Algorithm

The bearing is the direction of a route. When following a line on a curved surface from a start point A to an end point B the bearing will change for each arbitrary point between A and B because the line is curved. Therefore the initial bearing (also called forward azimuth) is taken which if followed in a straight line along a great-circle arc leads from A to B. The returned angle from the formula in Table 12 is between 0° to 180° (north to south) for eastern directions and between -180° to 0° (south to north) for western directions. Finally is normalized to an angle from 0° to 360° relative to north (as usual compass bearings are). The variables la_1, lo_1 and la_2, lo_2 are the latitudes and longitudes of two points which define a line/direction.

$$\theta = \operatorname{atan2}\left([\sin(lo_1 - lo_2) * \cos(la_2)], [\cos(la_1) * \sin(la_2) - \sin(la_1) * \cos(la_2) * \sin(lo_1 - lo_2)] \right)$$

Table 12: The initial bearing formula

4.1.1.3 Point in Polygon Test

For a flat plane there are several known methods to test whether a point is in a polygon or not. Such methods are for instance the winding number test or the even-odd rule which is applied

in the ray-casting algorithm. The concept of these methods can be applied for curved surfaces too.

The algorithm in the statistics application uses the even-odd rule. The point Q to test is defined by longitude and latitude. The (complex) polygon P is defined by a set of n points P_1 … P_n which are also defined by longitude and latitude. The actual algorithm is given in an article on Microsoft's Developers Network[23] and can be also found in the article "Point-in-Polygon Algorithm—Determining Whether a Point Is Inside a Complex Polygon" by Darel Rex Finley [24]. An explanation can be found in the report "Some Algorithms for Polygons on a Sphere" by Chamberlain & Duquette [22].

4.1.2 Nervosity and Circle Detection Algorithm

When a user's device changes its location it does not necessarily mean the user moved. The reason why a location changed is usually because another site provides better service to a mobile device. This may have several causes. One cause - and for this thesis the interesting cause - is a user moved and because of his movement his device has a better connection to another antenna/cell site than before. Other causes - and for this thesis less interesting causes - are for instance weather influences on the radio connection, environmental and geographical influences as well as any other cause than the movement of a subscriber.

Johannes Schlaich deals with such irrelevant location changes in chapter 4.2 "Fahrtenidentifikation" of his thesis [9]. To filter location changes which are assumed to have not been caused by a user's movement he introduces the "Nervositätsregel" ("nervosity rule" in English). The nervosity rule gives a ratio of the number of Location Area changes to the number of different Location Areas. Since Johannes Schlaich focuses on Location Areas his rule is applied in a different way. The term "nervosity" in the context of the method described here refers to visiting the same location again right after visiting only one different location. The term "circle" refers to visiting the same location again right after visiting two different locations. The terms are pictured in Figure 22 where the green arrows show a nervosity and the blue arrows show a circle.

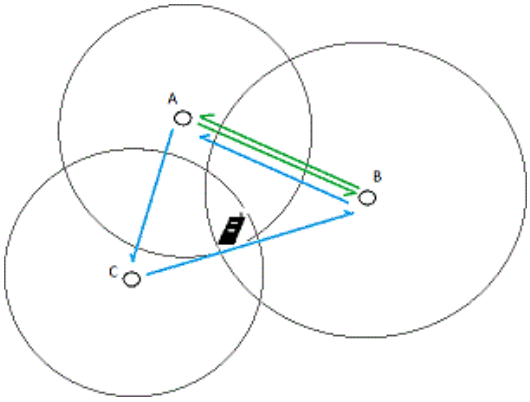


Figure 22: Nervosities and Circles caused by a mobile station

Instead of differentiating between Location Areas, locations are differentiated. Instead of calculating a factor of the location changes, the location changes are classified. Another

significant difference is that only the last four distinct locations of a user are considered. Filtering non-distinct locations is necessary because this algorithm shall work on Location Area updates as well as other events. Applying the rule set bellow sequences like A – B – A, A – B – C – B and A – B – C – A should be detected and classified. A, B and C are distinct locations.

There is a variable ε which defines a maximum distance for circles and nervosities. It can be assumed that nervosities and circles occur within a certain area. If a location sequence like A – B – A happens where the distance between A and B are more than ε meters it is very likely the user moved from A to B and back instead of causing an event at A and then at B because for instance of a better connection to the more than ε meters away B.

The actual value of ε has to be chosen by the user. A good value depends on several factors. The most important factor is the density of cell sites in the observed area and the cell site types in the observed area. Several cell sites with a small range could be backed up by an umbrella cell site (a cell site that covers other cell sites) with a large range for example.

The rule set described bellow shall detect the nervosities and circles of Figure 22.

1. The locations of every user are tracked in a list. Every user starts with an empty list. There are three counters: A nervosity counter, a circle counter and a movement counter. All counters are initialized to zero.

2. For every event seen from a user the location is inspected. If the location is not on the user's list yet it is added to it.

3. If the list only contains one location and the current location is already on the list nothing is done (a user caused two or more events at one and the same location).

4. If the list only contains one location and the current location is not on the list the current location is added to the list and the movement counter is increased by one.

5. If the list already contains two locations and the current location is the same as the last location on the list nothing is done (a user caused two or more events at one and the same location after moving to it).

6. If the list already contains two locations and the current location is the same as the first location on the list, the distance between the first and the current location is calculated. If the distance is smaller than ε the nervosity counter is increased by one, the movement counter is decreased by one and the user's list is cleared. Otherwise the user's list is cleared and the movement counter is increased by one.

7. If the list already contains two locations and the current location is distinct from the two locations, the current location is added to the list and the movement counter is increased by one.

8. If the list already contains three locations and the current location is the same as the first location on the list, the distance between the first and the second location as well as the distance of the second and the current location is calculated. If the sum of the distances is smaller than 2ε the circle counter is increased by one, the movement counter is decreased by two and the user's list is cleared. Otherwise the user's list is cleared and the movement counter is increased by one.

9. If the list already contains three locations and the current location is the same as the second location on the list, the distance between the second and the current location is calculated. If the distance is smaller than ε the nervosity counter is increased by one, the movement counter is decreased by one and the second as well as the current location are removed from the list. Otherwise the second as well as the current location are removed from the list and the movement counter is increased by one.

10. If the list already contains three locations and the current location is the same as the last location on the list, nothing is done (a user caused two or more events at one and the same location after moving to it).

11. If the list already contains three locations and the current location is distinct from all locations on the list, the first element is removed and the movement counter is increased by one.

It does not matter why an event occurred. Movements which do not match the pattern above are voted by increasing the movement counter by one. Since it takes two movements to create the sequence A – B – A the movement counter has to be decreased only by one. Instead of increasing the movement counter for the second movement the nervosity counter is increased. In a similar fashion sequences like A – B – C – B are dealt with. It takes three movements to create this sequence so the previous two movements were invalid and the third movement isn't count anyways (the circle counter is increased instead).

When this method is applied the information of the event block data has to be analyzed for phenomena first using the techniques described in chapter 2.1.2. Anachronisms may have a heavy impact on the numbers because nervosities and circles depend on the sequence events occur.

4.1.3 Nervosity and Circle Detection Results

The statistics and the Markov model above were created without considering nervosities and circles (both are location changes without subscribers' movements). In chapter 4.1.2 a method was introduced to detect them. This method was applied on the city of Graz within a radius of 12 kM (the same location and radius used to create the Markov model) and on the route A2 from Gleisdorf to Mooskirchen where the roadway itself is limited by polygons.

In the first application the parameter ε was set to 24 km. So every sequence that matches the nervosity or the circle pattern would be counted, regardless to its distance. The deviations were small. The mean value for distances of nervosities was 466 meters, for circles it was

1081 meters. The method is applied on the same region with a ε according to this mean value: ε = 500 meters.

The red line shows the classification result for circles, the blue line shows the classification result for nervosities and the green line shows the classification result for location changes caused by subscribers' movements. The purple line is the sum of all classification values.

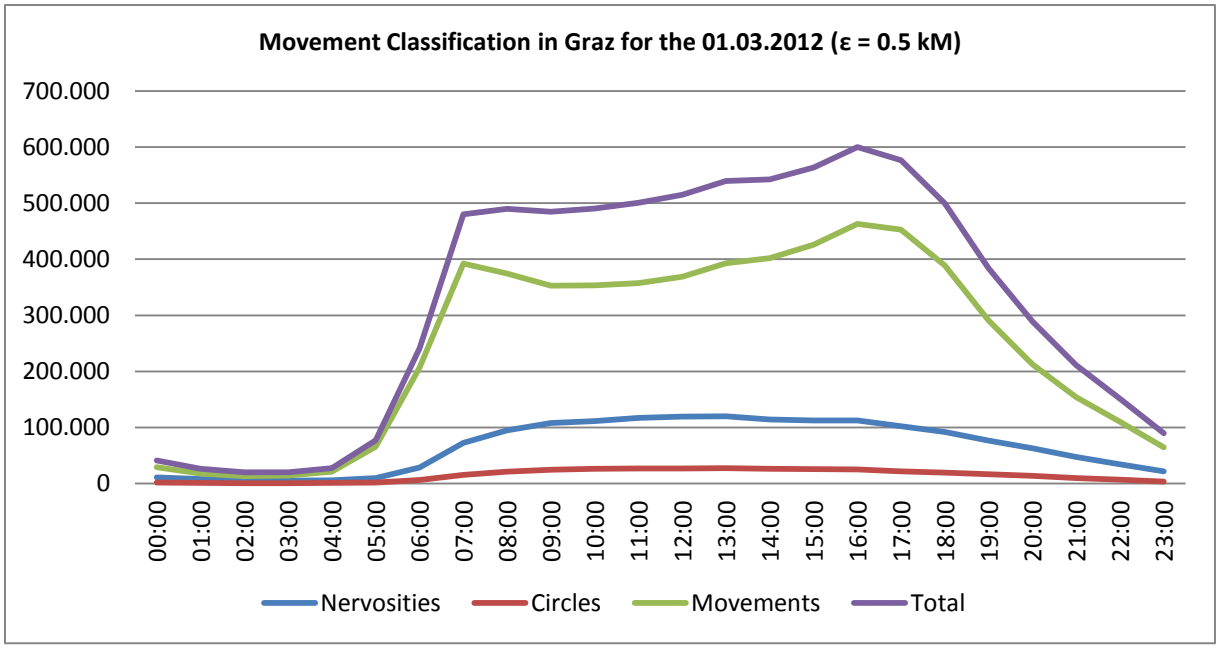


Figure 23: Movement classification in Graz on the 1st of March 2012 with ε = 0.5 km

The numbers for nervosities, circles and movements are different to those of the previous application with ε = 24 km. This is due to the fact that every detected nervosity reduces the movements by two and every detected circle reduces movements by three. On the other hand, the fewer nervosities and circles there are, the more movements are featured. Now the standard deviations for nervosities and for circles are zero. This indicates that ε was set to a value which only lets common nervosities and circles pass.

Not visible by the diagrams, but visible in the tables is the increased number of detected circles. *The number of detected circles is higher for a smaller value for ε*. This can be explained with the way the algorithm works.

Before a circle can be detected, a sequence has to be negative on the nervosity check. It can fail this check either by not being a sequence or by having a distance greater than ε. The minimum distance for a circle is 2ε, derived from the distance of nervosities. So a sequence like A – B – A – B – C – A can be classified to a circle, if the distance of A – B is greater than ε, but the distance of A – B – C is smaller than 2ε. A, B and C are different locations.

The idea behind setting the minimum distance for circles to twice of the distance to nervosities is the assumption that they both are related to each other. Figure 24 illustrates this idea. The blue arrows show a nervosity with a distance within ε. The green arrows show a

circle with a distance within 2 ε because the distance of A to B and the distance of B to C are small enough. While the red arrows show a circle which is not classified as circle because the distance of A to B and B to C are too large.

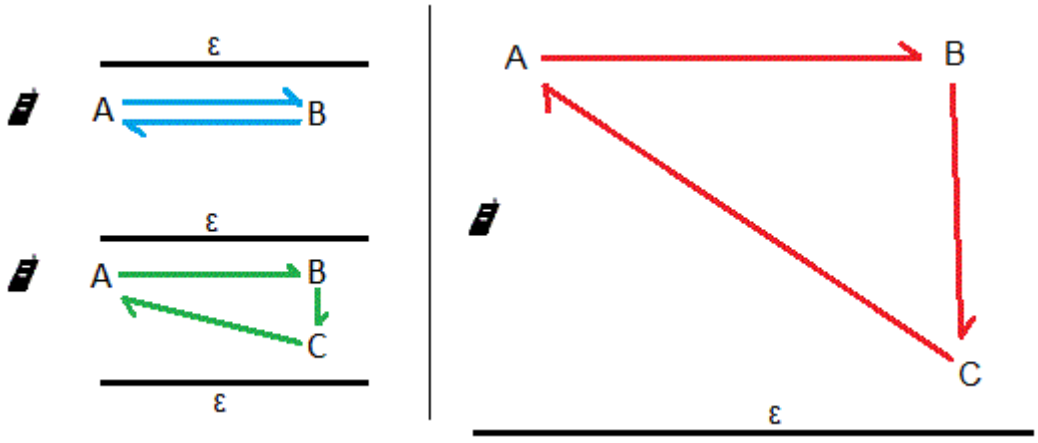


Figure 24: The correlation of Epsilon, nervosities and circles

The algorithm could of course be modified to take two epsilon parameters: One for nervosities and one for circles. But since one epsilon is enough to remove standard deviations in both, nervosities and circles the algorithm in this thesis is left to one parameter ε.

Again, the method applied is a classification algorithm. It can only give a clue about amounts of nervosities, circles and especially about movements. Assuming A, B, C, D, E and F are different locations, a subscriber which causes location changes in the sequence A – B – C – B – D – E – A – B – C – D – E – F is classified as 1 nervositiy (B – C – B), 1 circle (A – D – E – A) and 5 movements (A – B – C – D – E – F).

Especially the movements have to be put in relation to the number of devices detected when interpreting the results. A high number of devices and a low number of movements mean few movements by each device. A low number of devices but a high number of movements mean a lot of movements by each device. To make a high number of movements per device possible, the observed area needs a high number of cell sites. Therefore the conclusion for the algorithm is, it is a classifier that can detect location change pattern which are most likely not caused by a subscriber's movement. Movements are featured by the algorithm. How these detections are interpreted and used are up to the user of the algorithm.

4.1.3.1 The Application on a Route

The above method was used on a part of the roadway A2. The observed area is within the limits of the green polygon of Figure 25. The red line highlights the route. The orange dots are possible locations for the 2G and 3G network of the mobile network provider. The area was chosen to include locations which are assumed to be accessed by users who actually are on the roadway. The transition probabilities from chapter 3.3.1 were the base. If a point is within the polygon was determined with the algorithm from chapter 4.1.1.3.

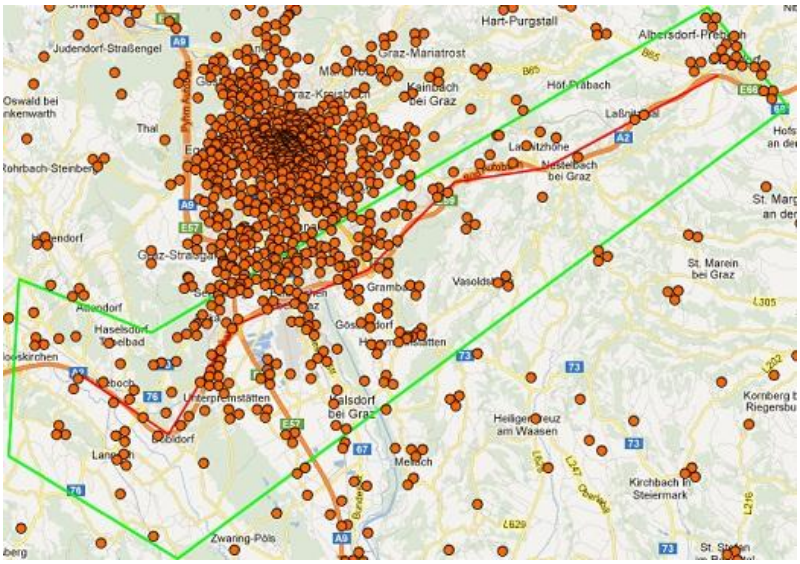


Figure 25: the polygon limiting the route A2 from Gleisdorf to Mooskirchen

In the first application the parameter ε was again set to 24 km so nervosities and circles were not limited to minimum distances. In the second application ε was set to 2.6 km. This values is the mean distance of nervosities found in the previous application with ε = 24 km. Using ε = 2.6 km still yielded standard deviations. Therefore the algorithm was applied a third time with ε = 0.5 km. The third and last application yielded no standard deviation. This means the detected nervosities and circles always had the same distance and therefore they are supposed not to be caused by movements (i.e. it is unlikely that many subscribers would move in the same circle).

During the application 200.380 different devices were found in the area of interest. Each device caused about 62 events on average with about 20 events ± standard deviation.

The first application on the A2 with ε = 24 km:

The corresponding graph to the application can be found in Figure 26. The green line is the classification results for movements, the blue line is the classification result for nervosities and the red line is the classification result for circles. The puple line is the sum of the classification values from detected circles, nervosities and movements.

The application yielded similar classification as when applying it on whole Graz as in chapter 4.1.3. There are again two peaks. But this time the peaks are more featured than in the application on whole Graz. The first peak is in the morning time, the second peak is in the evening time. Circles and Nervosities also increase in the morning time and then are stuck on a plateau with only slight increases. Both circles and nervosities decrease in the evening. Again the nervosities catch up with movements when ε is set to 24 km.

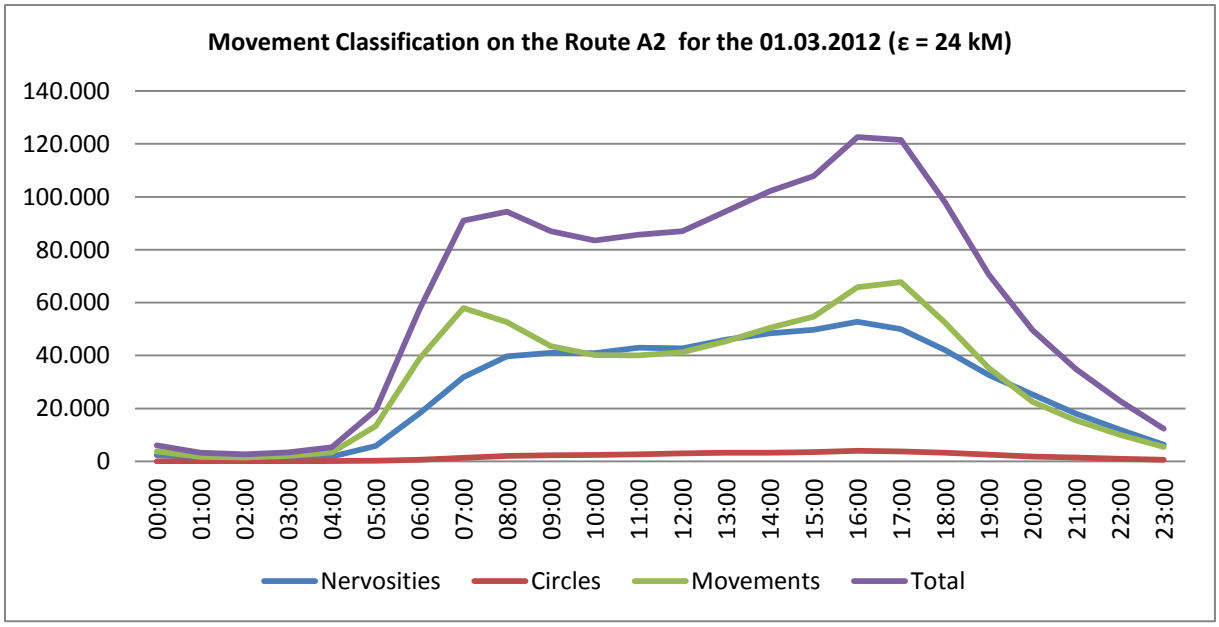


Figure 26: Movement classification for the A2 on the 1st of March 2012 with ε = 24 km

Table 20 in the appendix has the actual figures for Figure 26. There are only two hours with no standard deviation, but almost twice as much hours with deviations for nervosities and Circles as for Graz. The mean distance for nervosities is about 2.5 km and the mean distance for circles is 2.1 km. These means are higher than those found in Graz. Because of the mean for nervosities ε = 2.6 km is used for the second application.

The second application on the A2 with ε = 2.6 km:

The result for a smaller epsilon is fewer nervosities, more circles and a higher value for movement classifications as shown in Figure 27. Although ε was adjusted according to the mean value of the previous application there were still deviations in the current application. The number of deviations decreased and the number of rows with no deviations increased as shown in Table 21 in the appendix.

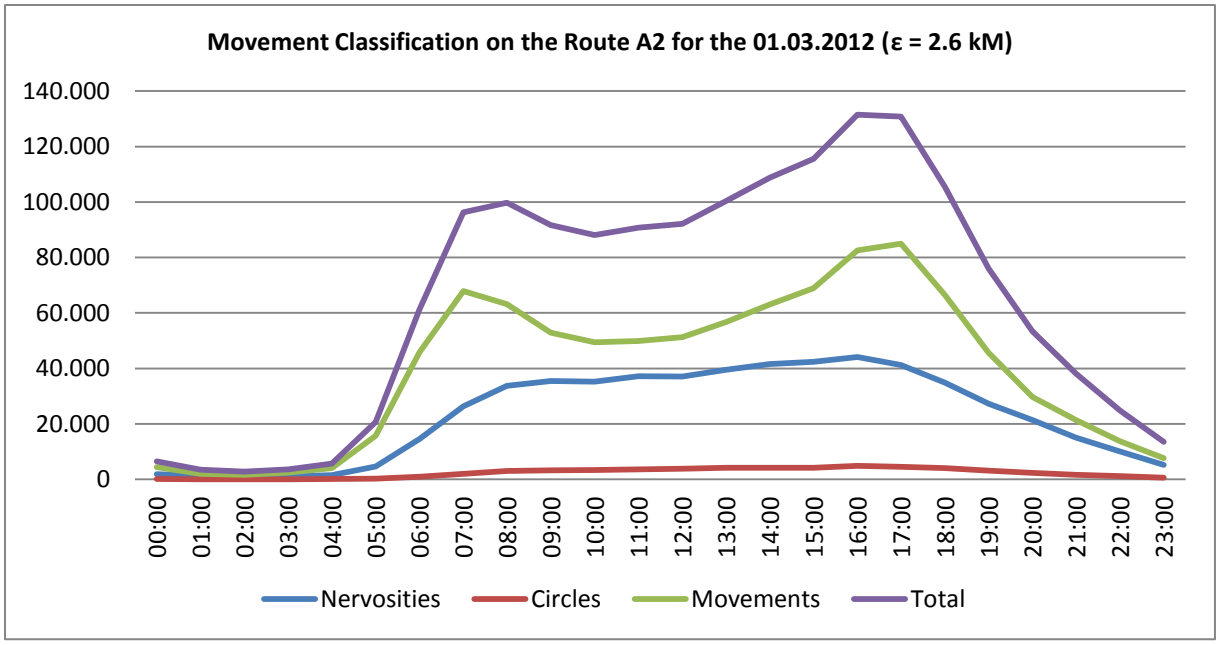


Figure 27: Movement classification for the A2 on the 1st of March 2012 with ε = 2.6 km

The third application on the A2 with ε = 0.5 km:

Decreasing the value for epsilon again results into a detection of fewer nervosities and a higher movement classification. There are slightly more circles detected on some hours. With ε = 0.5 km there were no more deviations. The results can be seen in Figure 28 the actual numbers are in Table 22 in the appendix.

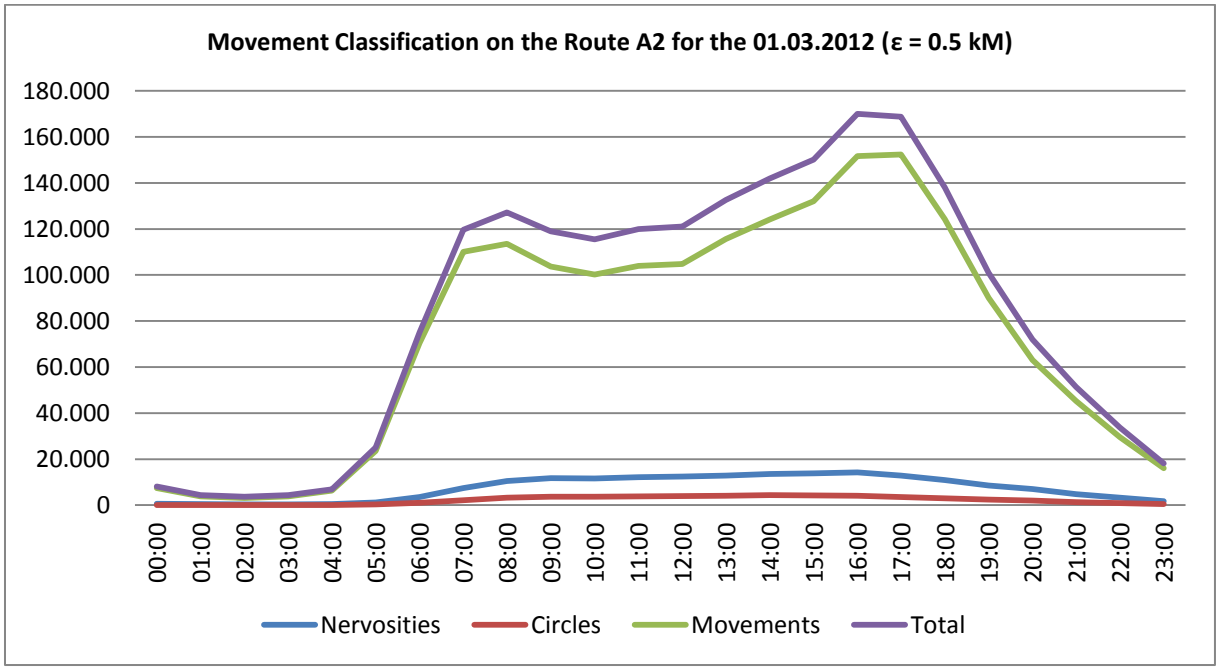


Figure 28: Movement classification for the A2 on the 1st of March 2012 with ε = 0.5 km

To sum the nervosity and circle detection up it is a further filtering of data (if movement data are sought). The algorithm filters events which are supposed to have not been caused by movements. The algorithm can also be used to feature users for their mobility. Since more movements are rated with a higher classification value user who have more moves without circles and nervosities will be featured higher than users who may have more moves at all or left a greater distance behind. The algorithm can also contribute to evaluate the infrastructure of the mobile phone network and to find out cell sites and locations which are rather subject to circles or nervosities than other sites.

5 Road Surveillance

5.1 The Algorithm

By combining the results from the previous chapters an algorithm is derived to observe a route and tell about the situation on a specified route in a certain interval. The algorithm needs preliminary knowledge of the route as well as of the cell sites/locations where events are supposed to occur. Therefore the user has to define the route, an area of interest, a maximum distance to detect nervosities and circles and an interval to report about the situation. The algorithm will then inspect events occurring in the area of interest to infer about the traffic situation on the route and report the situation after a given interval.

A route R is defined as directed, connected graph with x nodes but without cycles. The nodes of the graph are the points on the route $N = \{n_1, \ldots, n_x\}$, $N \in R$. Each edge of the graph connects two nodes. The edges are also the Euclidian distance between two nodes $E = \{e_1, \ldots, e_{x-1}\}$, $E \in R$. Each point on the route (or node of the graph) is a tuple consisting of longitude and latitude $n = \{lo, la\}$. The tuples of all points are distinct. The first point n_1 in the set of points is the start of the route. The last point n_x in the set of points is the end point. The direction of the route is given by walking the points subsequently from the start point n_1 to the end point n_x.

The route is enclosed by a polygon P. This polygon defines an area of interest. The polygon is defined by a set of y points $P = \{p_1, \ldots, p_y\}$. Each point is a tuple consisting of longitude and latitude $p = \{lo, la\}$. The points ordered clockwise and when connected the shape of the polygon is formed.

The route as well as the area of interest has to be defined by a user. The output of the algorithm is depending on their definitions. The key for a good route definition is to place way points nearby possible locations because the neighborhood is of importance. In a similar fashion the area of interest has to be defined in a way, that it includes all locations where the user expects events which are related to the traffic situation of the route.

While the algorithm is running a data structure “User Information” *UI* is used to store certain user information. This user information contains the following information about a user which is used for analyses:

1. The anonymous Id of the user.
2. A first-in-first-out queue with the last three locations visited by the user sorted ascending by their occurrence time (to track the last two moves of the user).
3. Another list with all locations visited by the user ordered ascending by the time of visit (to track all locations of the user on demand).
4. A feature value which indicates the impact of the user on the route traffic.
5. A feature value which indicates the direction of the user.
6. The closest node of the route to the user's first known location.
7. The closest segment of the route to the user's last move.

The result of the algorithm is a set of user information that describes the situation on the route at a given time. In the result set are only users which are assumed to contribute to the traffic on the given route. The impact on the traffic is expressed by the feature of the user's last move as well as by the user's first known location.

By allocating the closest way point to the user's first location the part of the route is featured where a new users just appeared. It is not guaranteed that a user who triggers an event in the area of interest triggers another event in the time span of interest. If many users appear at a way point and many users move away from this point, the way point is most likely a starting point. If many users appear at a way point but don't move along the route, the point may be on another route not in the area of interest or it could be a destination point for users.

The feature of a user's movement indicates how much its direction matches the direction of a route. The user does not necessarily be on the route. If the route is according to a road a user could be on a parallel road (depending on the set up of the area of interest).

It is finally up to the user of the algorithm to interpret the number of movements detected, the features assigned to the movements and the ratio of movements to users detected in the area of interest. Since the algorithm assigns feature values to each user which are continuous values instead of discrete values the algorithm may fail as counting algorithm.

The algorithm is prone to some phenomena of chapter 2.2. Parallel events may add invalid features to a route. So do supersonic user events. In addition they may start the dismissal of a user actually who never left the area of interest. Anachronisms can be neglected since the events are sorted by their occurrence time.

5.1.1.1 Preparation and Route Set Up

The algorithm needs to know the route and an area of interest and a maximum distance for circles and nervosities. Secondary parameters are limits to detect phenomena and an expiration time for users who do not trigger further events.

The polygon for the area should limit the route in a shape that contains those locations where users who are on the route may trigger an event. But the polygon should also be small enough to exclude routes which are parallel to the observed route or which may affect the results in any other negative way. The graph defining the route should match the actual route on the one hand but also have its nodes nearby a cluster of possible locations on the other hand.

An idea for a route definition and the shape of the polygon can be obtained by creating a Markov model of location changes as described in chapter 3.1.2. Locations which show likelihoods to see users along the route should be definitely included in the area of interest. Figure 29 shows the likelihoods of locations changes along the roadway A2 from Gleisdorf to Graz. Figure 30 illustrates a setup to observe the route A2 from Gleisdorf to Mooskirchen.

In the north of the roadway in Figure 29 location changes away from the route can be seen. In the south the location changes are along the route. In Figure 30 the polygon excludes the locations leading away from the route in the north and includes those in the south.

The route is simply set along the roadway. In Figure 30 it can be seen as red line. The way points (marked as red figures along the route) are nearby locations. The black dots are permanent road way counters of the ASFiNAG. The black numbers next to the black dots are the number of the counting station.

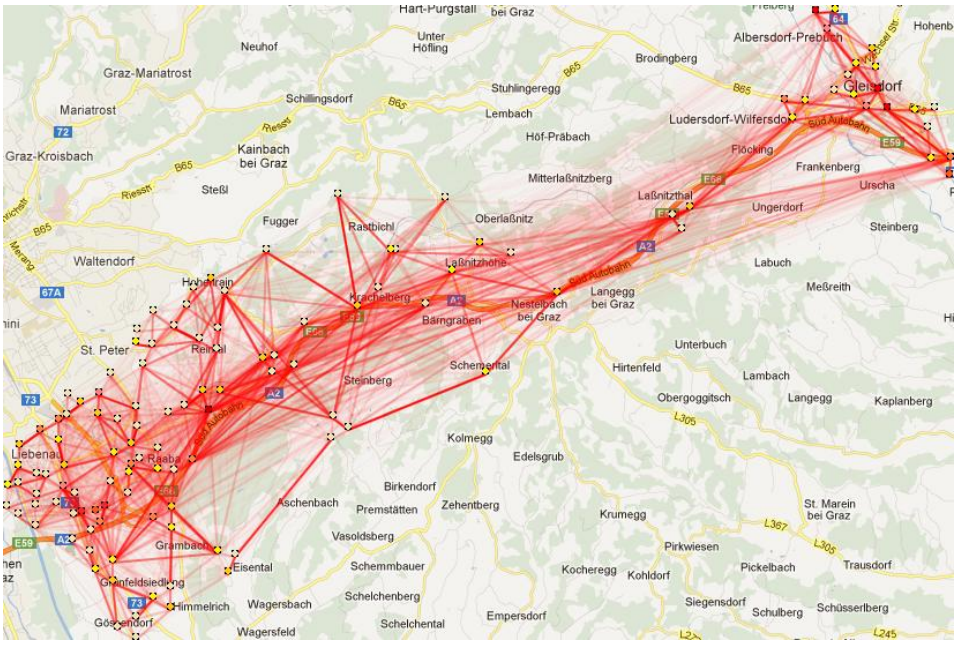


Figure 29: A zoomed part of the Markov Model from chapter 3.3.1 limited to the road way A2 from Gleisdorf to Graz

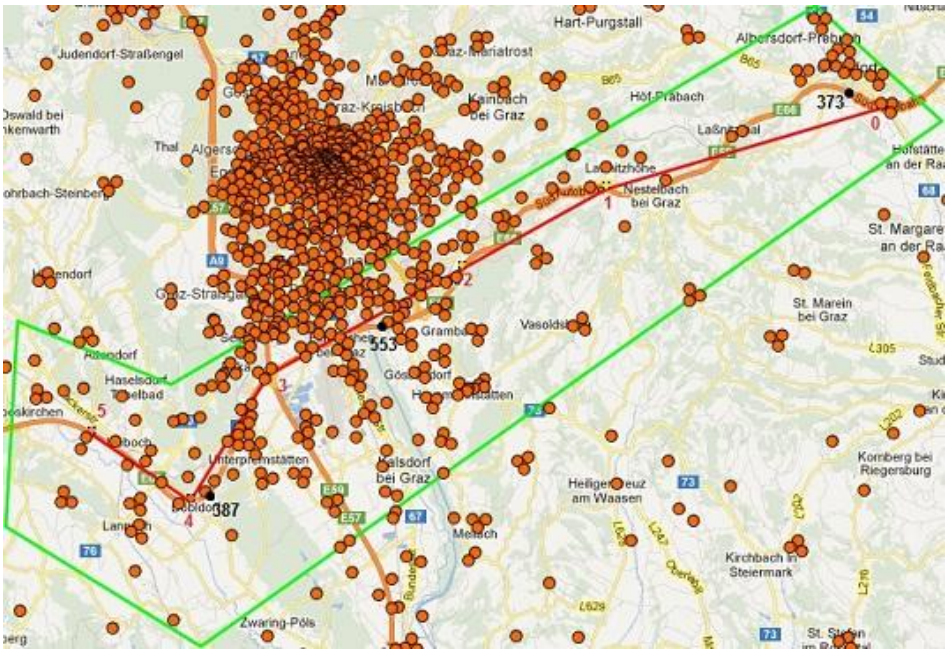


Figure 30: A setup for the algorithm to observe the route A2 from Gleisdorf to Mooskirchen showing possible locations, the area of interest as well as the segments of the route

Beside the polygon and the route a maximum distance for circles and for nervosities is needed. A good value for a maximum distance can be obtained by applying the method described in chapter 4.1.2. In this application it is 500 meters.

The minimum distance for a supersonic user is 100 km, the maximum speed is 300km/h.

A user's information expires if there is no event of a user in the area of interest within an hour. If a user's information would not expire, it would remain in the results as person moving on the route even if he or she already left the area of interest.

5.1.1.2 The Analyses Process

After the algorithm is set up it is fed with event blocks. While the algorithm is running it keeps track of users in the area of interest by putting them on a hash map which maps the anonymous Id of a user to his or her information. The algorithm has two sub processes: A dismissal process and a feature calculation. These two sub processes are explained after the overall algorithm is explained.

For each event block the following steps are done:

1. It is checked whether the event occurred in the area of interest. If it is not in the area check if the anonymous Id is in the hash map. If it is in the hash map mark it for the **dismissal process**. If not discard the event.

   If the event is in the area of interest and the anonymous Id isn't already in the hash map the user just entered the area of interest. Initialize new user information *UI* and store the location as well as the occurrence time in it. Allocate the closest way point to

the location of the event. Save the *UI* in the hash map and process the next event block.

If the event is in the area of interest and the anonymous Id is already in the hash map the user triggered a second or a further event. Load the *UI* for the user and save the current location to it.

2. Check the locations of the user for circles and nervosities. Remove them if there are some. If there is no more than one location known for the user process the next event.

   Circles are removed by simply removing all locations from the UI and only keep the current one. At least four locations are needed to detect a circle. The UI keeps the last three locations of a user. So if a circle is detected all known locations for a user can be removed and only the current location is kept.

   Nervosities are removed by simply removing all locations from the UI if there are only two locations known for a user. Three locations are needed to detect a nervosity. The UI keeps up to three locations. If two are kept and the current location turns out to be a nervosity all two locations can be removed and the current location is kept. If there are three locations already known for the user the last two locations are removed and the first as well as the current location is kept.

3. If there is more than one event known for the user allocate all locations of the user to the closest nodes of the graph defining the route.

4. Calculate the bearings of the edges connecting the allocated nodes. For *n* found nodes calculate *n* bearings for node 1-2, 1-3, …, 1-*n*.

5. Calculate the bearings of the locations of the user. For *n* locations calculate *n* bearings for location 1-2, 1-3, …, 1-*n*.

6. Calculate **feature λ** with the bearings and use the best feature value. Update the UI and process the next event.

The dismissal process:

If a user who's last event was in the area of interest triggers an event outside the area of interest is removed from the algorithm's hash map. It is assumed the user left the route.

Users who do not trigger an event within a user defined time are removed from the algorithm's hash map too.

Calculating λ:

An essential part of the algorithm is calculating λ which describes the situation on a route. The user's contribution to traffic is calculated by comparing the bearings of the user's movement and the route which has to be observed.

The user's contribution to traffic on the observed route is expressed by the match of his move's direction to the direction of the route. The more similar the bearings of his or her move and the direction are the more the user "contributes" to the traffic. Of course the direction of the route and the direction of the user's move have to be considered. A user may be on tour (he or she moves along the direction) or retour (he or she moves against the direction).

The user information data structure keeps a queue of the last three locations (A, B and C of a user). Three bearings (A-B = α_U, B-C = β_U, and A-C = γ_U) can be calculated with these locations.

In step 5 of the analyses process the locations of a user are allocated to the closest nodes (N_A, N_B and N_C) defining the graph which represents the route. With these nodes another three bearings (N_A - N_B = α_N, N_B – N_C = β_N, and N_A – N_C = γ_N) are calculated. These bearings are subtracted from each other and absolute value of the differences is used ($|\alpha_U - \alpha_N|$, $|\beta_U - \beta_N|$ and $|\gamma_U - \gamma_N|$). The smaller the difference, the better is the match. The closer the difference is at 90° the worse is the match. Figure 31 shows a route (red line) and several imaginary moves (blue, purple and green lines).

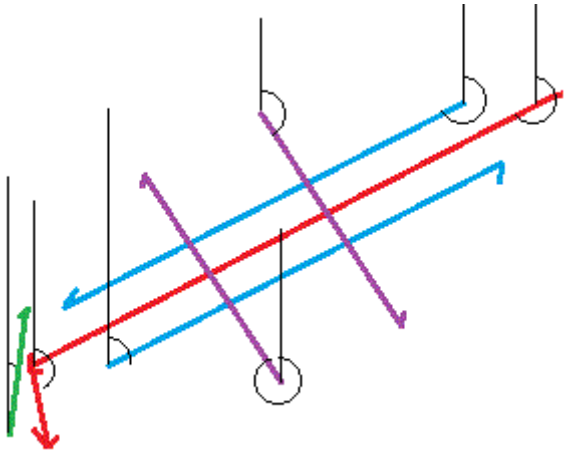


Figure 31: Possible directions and their bearings

By using Figure 31 regarding the direction is explained. When looking at the blue and the purple lines and their bearings (black thin lines and a circle), it can be seen that their difference is always 180 °. In other words, subtracting the bearings of opposite directions always yields ±180°. So the direction can be neglected by subtracting 180° from every bearing which is greater than 180°.

After neglecting the direction value between 0 and 180 degrees can be expected as difference of two bearings. An example for such a situation is the green line and the last segment of the red line.

Interesting λ are differences between 0 and 90 degree. 0° means the directions are perfectly parallel as seen in the blue lines. 90° means the move is absolutely normal to the direction and therefore follows it in no way as seen in the purple lines.

The remaining problem is the case of the green line. This situation is solved by subtracting again 180° from the difference of the bearings if the difference is greater than 90°.

The smallest difference is used to calculate the actual λ value. The smallest difference is now divided by 90° and subtracted from 1. *So the actual feature is a value between 0 and 1*.

The λ of all user in the area of interest are summed up. This sum is then representive for how many users are on the route. By considering the directions of the users two sums can be calculated to express how many users are on the route in which direction.

5.2 Road Surveillance Results

The algorithm described in chapter 5 was used to observe a route from Gleisdorf to Mooskirchen along the roadway A2 on 01.03.2012. The way points for the route are defined in Table 13; the points defining the polygon are defined in Table 14.

Latitude	Longitude
47.09229195595789	15.726662635803223
47.06359254442066	15.583454132080078
47.03540415526925	15.508695602416992
46.99801110330551	15.408016204833984
46.95326892003427	15.367504119873047
46.97716730919103	15.316005706787110

Table 13: Coordinates of the route's way points

Latitude	Longitude
47.12827587405629	15.69228744506836
47.08761693487146	15.757475852966309
46.90192160067718	15.372997283935547
46.94436286885973	15.271717071533203
47.01697423314641	15.277896881103516
46.99461575850479	15.357547760009766

Table 14: Coordinates of the polygon

The polygon and the route defined by the tables is shown in Figure 30. The ε for nervosity and circle detection is 500 meters. The minimum distance for a supersonic user is 100km and the minimum speed is 300 km/h. Information of users who are in the area of interest and do not trigger an event within an hour are removed. Using this setup an application which implements the algorithm evaluated user events and movements.

Figure 32 shows the output of the application. The data are read from log files which contained an image copy of the received stream data (first line in console). The application computes a mean value and a standard deviation of the timestamps of received events (third line in console). This is because of anachronisms and the possibility that the application may lag behind when processing data online instead of from log files. Bellow the third line are two tables which show the number of users nearby way points and the feature values of detected

movements. The evaluation is updated for every event and the updates are displayed every ten seconds.

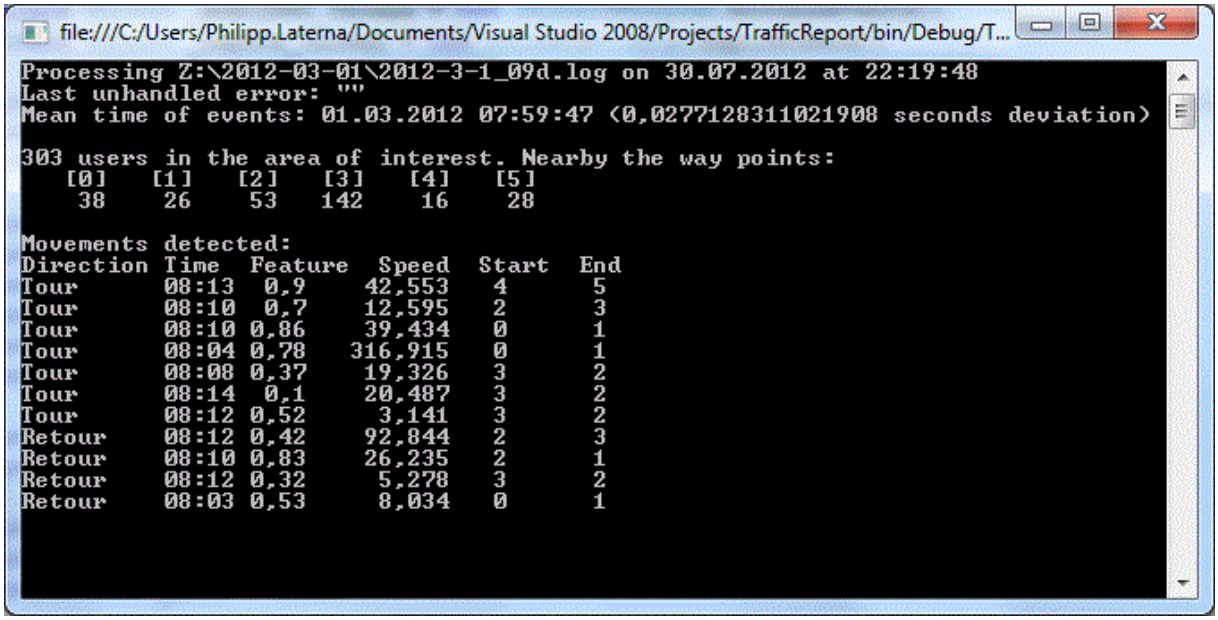


Figure 32: Online evaluation output of the road suveillance algorithm

A user is allocated to a way point and added to the area of interest as soon as he triggers his or her first event in the area. The user is removed when he or she triggers an event outside the area or expires. Updates do not change the allocation of a user. So once he or she is allocated to a way point the user remains there until he or she triggers an event outside the area or is cleaned out because of inactivity.

By doing so the concentration of users at way points is illustrated. A users appears at a way point because he or she was in the area and active his or her device (which caused an event) or because the user moved into the area and the device caused an event. If the first case is true for many users, the way point may be nearby an area where many people “rest” (i.e. they are at home, like in a town or a village). If the second case is true the way point may be on a supply route to the area of interest which leads many users into it. In either case the numbers of movements to this point or away from this point are interesting.

In the table of detected movements, each line represents the movement of one user. So if there are eleven lines, the moves of eleven users were detected. The direction can be tour (along the route definition) or retour (contrary to the route definition). In the case of this application tour means from Gleisdorf to Mooskirchen and retour means from Mooskirchen to Gleisdorf. The time column indicates when the movement was detected. The feature value indicates how much the bearing of the movement matches the segment it was allocated to. The speed value is the distance between the two locations of the movement divided by the difference of the occurrence time of the events. The start and the end columns are the way points which are the start and the end of the segment which is allocated to the movement.

If one and the same user triggers more events during the time span of surveillance only his best matching move is used and displayed.

The speed is only an estimate of the user's actual speed. Because of the inaccuracy of locations and the possibility of fast handovers impossible values (like 317 km/h) may be computed.

Figure 33 illustrates the movements detected by the algorithm as blue lines. The more opaque the lines are the lesser they match their allocated segment. Lines which lead outside the area of interest (green shape) are movements from users which left the area. The yellow and orange dots are way points. The yellow circle is the area which is zoomed-in in Figure 34.

Figure 34 illustrates the position of a way point nearby two hot spots for user's locations (marked with two yellow circles). These spots seem to be the origin and the destination of many detected moves. To have these spots in the area of interest is important to detect movements of users along the route.

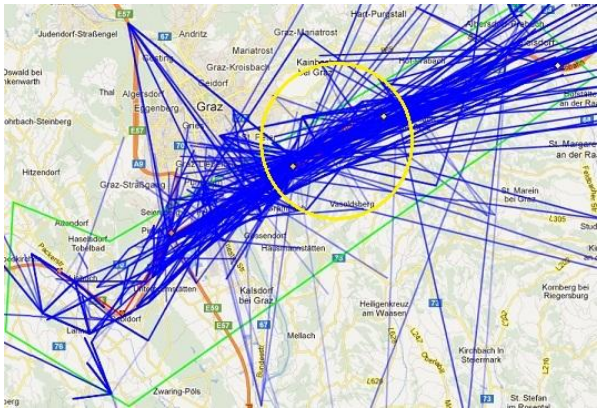


Figure 33: Movements detected by the algorithm

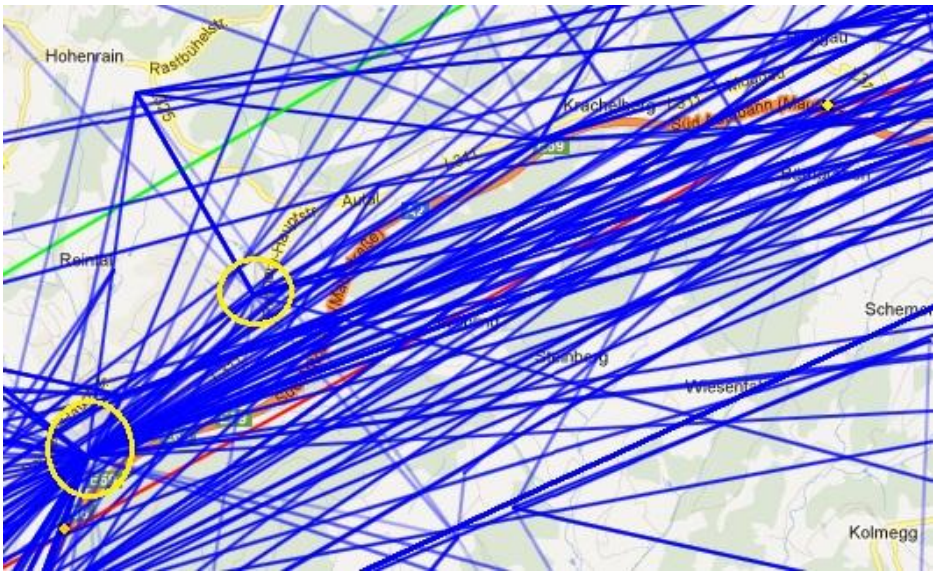


Figure 34: Movements detected by the algorithm and two highlighted hot spots for locations

As it can be seen the detected location changes were evaluated correctly according to their direction and therefore to their likelihood to contribute to road traffic.

5.2.1 Comparison with Counting Stations from ASFiNAG

The ASFiNAG provided data from three permanent counting stations on the roadway A2 from Gleisdorf to Mooskirchen for the 01.03.2012. These data include the number of vehicles detected at a station for each direction (tour or retour) every hour. The stations are station 373 at Pirching an der Raab (nearby Gleisdorf), 553 at Thondorf (in Graz) and 387 at Dobl. A graph of these numbers can be seen in Figure 35. The orange line shows detected vehicles at the station in Pirching an der Raab, the blue line shows numbers of the station in Thondorf and the green line shows detection of the station at Dobl. Filled lines show "Richtung 1" (users from Gleisdorf to Mooskirchen), opaque lines show "Richtung 2" (from Mooskirchen to Gleisdorf).

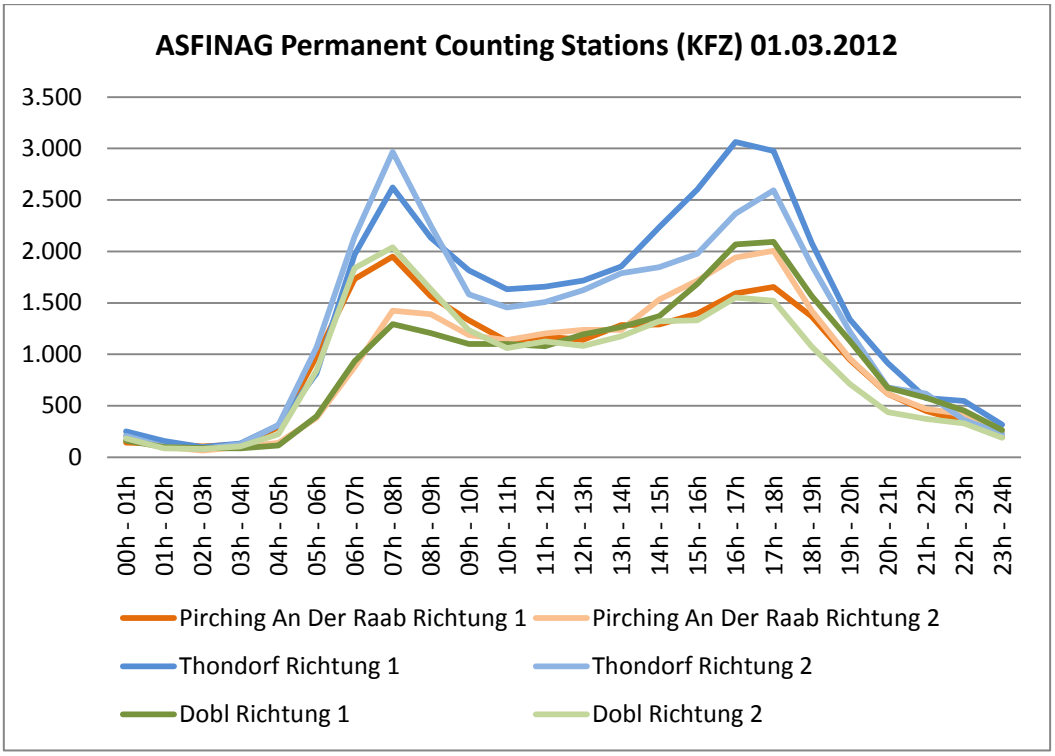


Figure 35: Vehicles counted on 01 of March 2012 at each station for every hour

A correlation of the graphs in Figure 35 with the detected movements in Figure 26, Figure 27 and Figure 28 can be seen. When adding the values of both directions from the counting stations, then computing the mean value of all counting stations and finally computing the correlation coefficients, then the coefficient for movements of Figure 26 is 0.988, for movements shown in Figure 27 is 0.981 and for movements shown in Figure 28 it is 0.953. Since the chance to get this correlation with random values is very small (smaller than 0.0001) the correlation seems to be significant.

The graph shows further that users seem to go to Graz in the morning time and leave Graz in the evening time. The green opaque graph has a higher peak in the morning but in the evening

the green filled graph has a higher peak. In opposite the orange filled graph has a peak in the morning and the orange opaque graph has a peak in the evening. These trends can be explained with commuters who visit Graz to work during the day and leave Graz for home in the evening.

To benchmark the movement featuring algorithm the outputs of the algorithm need a special refinement so the results can be compared to the data from ASFiNAG. The algorithm features a movement where it is detected. To use the results for counting, the detected features of a user are saved for each segment. If a user improved his feature on a segment, the value for the segment is updated. If a user shows features for several segments, the value for each segment is stored. For every hour the features of users are added together for each event. After an hour the value for each segment is reported; then the values are reset. After the reset only values from users who are still on the route are used (in addition to the values of the newly detected users).

Assuming there are two segments *A* and *B* and three users *x*, *y* and *z*. In the hour from 07:00 am until 08:00 am user *x* shows a feature of 0.8 on segment *A* and a feature of 0.9 on segment *B*; then user *x* leaves the area of interest. User *y* shows a feature of 0.4 on segment *A* and then leaves the area of interest. User *z* shows a feature of 0.6 on segment *B* but does not leave the area of interest. The report for the hour between 07:00 am and 08:00 would show 1.2 for segment *A* and 1.5 for segment *B*. In the hour from 08:00 am until 09:00 segment *A* would start with value 0 and segment *B* with 0.6.

This is of course no real user counting, but the feature values should express the number of users on the route. Figure 36 shows these values for the segments on the defined route and for both directions. Tour means from Gleisdorf to Mooskirchen and retour means from Mooskirchen to Gleisdorf. The graph is similar to the graph in Figure 35.

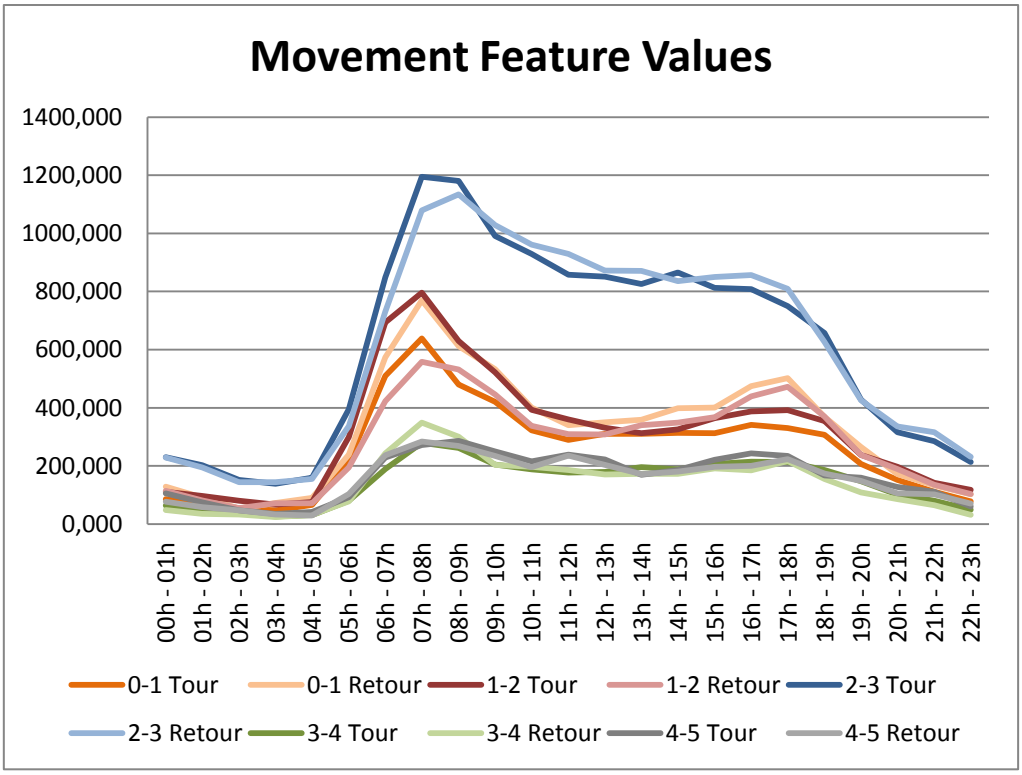


Figure 36: Movement feature values for route segments on roadway A2 on 01.03.2012 for each hour

The first common feature is the two peaks between 07:00 am and 08:00 am as well as between 17:00 pm and 18:00 pm. The peaks in the graph of the counting stations have the same height in the morning and in the evening. In the graph of the features the peaks in the morning are remarkable higher.

In the graphs of the counting stations all peaks are flipped. When a filled peak is higher than an opaque one in the morning, the opaque peak is higher than the filled peak in the evening and vice versa. This is also true for the graphs of feature of the segments 2-3 and 1-2.

The permanent counting station in Thondorf in Graz shows significant higher values than the other counting stations. In a similar fashion does segment 1-2 which goes along the A2 and passes Graz.

The values of segment 1-2 seem to match best the numbers of the ASFiNAG counting stations, followed by values of segment 0-1. The other segments seem to share only the rise in the morning and the decline in the evening. The peaks and the flipping are not visible. When computing the correlation coefficients of segments to their next counting stations, the coefficients are all above 0.8 as shown in Table 15.

Segment	Counting Station	Correlation Coefficient ρ
0-1 Tour	373 Richtung 1	0.9442
0-1 Retour	373 Richtung 2	0.8019
1-2 Tour	373 Richtung 1	0.9078
1-2 Retour	373 Richtung 2	0.8847
2-3 Tour	553 Richtung 1	0.8573

2-3 Retour	553 Richtung 2	0.8870
3-4 Tour	387 Richtung 1	0.8666
3-4 Retour	387 Richtung 2	0.9571
4-5 Tour	387 Richtung 1	0.8188
4-5 Retour	387 Richtung 2	0.9449

Table 15: Correlation coefficients of feature values for segments and numbers from counting stations

According to these coefficients the summed feature values of the algorithm express the number of vehicles on the street very well. The chance to get this correlation with random values is very small ($< 10^{-4}$) the correlation seems to be significant. But it has also to be seen that these features lack in counting vehicles like the permanent counting stations do. When looking at the peaks in the graphs of the counting stations, the numbers show that almost the same number of vehicles in the morning is on the other direction in the evening. The feature values for peaks are in all graphs higher in the morning than in the evening.

Differences between the graphs seem to have two reasons. The first reason is that the graphs of the counting stations were created by counting. They are a hard fact. The graph for movement feature values was created by adding feature values. Feature values are soft facts. They do not count movements, they express them. The second reason is the data source. A counting station counts a vehicle when it passes the station. A movement is featured when at least two locations of a user are known. So if there is little mobile phone activity for any reason fewer movements are likely to be featured. When seeing the feature values of movements in relation to detected movements they would express how many of the movements are along the route.

The comparison is therefore not optimal, but it was done since no other data for validation were available.

The algorithm also returned the number of new users detected at the way points for each hour. This result is shown in Figure 37.

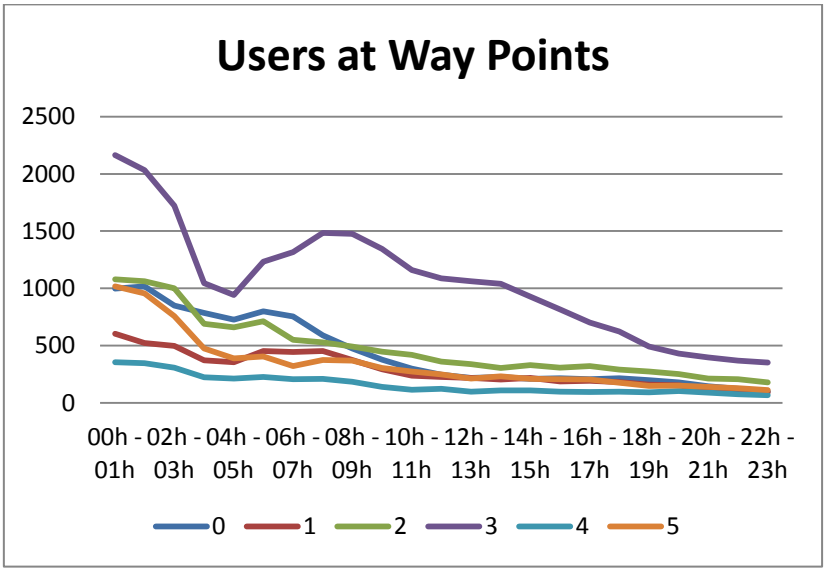


Figure 37: The number of new users detected at each way point for every hour

The graphs show a declining trend after the morning hours. The highest number of new users is after midnight. But these numbers must not be mistaken. The algorithm counts a new user when he or she is seen for the first time in the area of interest, or again in the area of interest after having left it or after expiring. After midnight no user is known to the algorithm and everyone who triggers an event in the area of interest is new. In the first three or four hours the algorithm seems to "adapt" itself to the area of interest by getting known to the users in the area. Significant is the increase in the morning hours between 05:00 am and 08:00 am. During these hours actually new users seem to enter the area of interest. This fact seems to support the higher movement feature values in the morning. The question is why there is no increase in the evening time when users move in the other direction. This behavior might be explained with the "greediness" of how users are counted. Once a user appears at one way point it cannot be counted on another way point until he or she leaves the area or expires.

It is hard to interpret these numbers further because there are no other data to correlate with than the numbers from the counting stations provided by the ASFiNAG.

6 Discussion

This thesis is an explorative thesis which tries to find new approaches to use mobile network data to support road traffic analysis. It does not try to prove or verify existing methods. This is why only twelve days were taken to work with. For methods which were CPU intensive only one day was taken to apply the methods. Data from other sources like permanent counting stations or similar would be necessary to feature assumptions or to show wrong ways. Still the application of the methods in this thesis showed some interesting results which seem promising to support traffic planning and traffic surveillance. Improvements of the methods and other applications of the data provided by the mobile phone network operator are discussed in this chapter.

6.1 Movement Detection Results

In chapter 4.1.2 a method was introduced to detect location changes which are no movements but have been caused by handovers or similar reasons. In chapter 4.1.3.1 this method was applied on the roadway A2 with different minima for distances. The result was the smaller the minimum the smaller the deviation for distances of detected circles and nervosities. Considering the architecture of a mobile phone network no deviation was interpreted as a positive result. In chapter 5.2.1 the results from the road surveillance algorithm was compared with results from permanent counting stations. Also the results from movement classifications were compared. The figures told a different story. The classification of movements with a greater minimum distance correlated better with the numbers from the counting stations.

The numbers from the counting station are a hard fact. It states how many vehicles passed the station and to do so the vehicle had to move. In other words the numbers from the counting stations express movement. But counting stations can only show vehicles which move by. Event reports from mobile phone networks also include not-moving persons. What seems to be a movement from the event reports does not need to be an actual movement. Looking at the tables of the movement detections the detection worked better using small minima because no deviations occurred. The deviations are no hard fact; they were interpreted to be false positives without being able to verify this assumption.

The conclusion from these contrary results is that movement classification needs to be explored better. The results in the previous chapters base only on data from one day and from one route. Therefore these data should be interpreted as idea for further research.

6.2 Mobile Network Data Usage

Extensions for methods of this thesis and ideas for other applications are discussed in this chapter. Of particular interest are traffic jam detections and detecting movement trends (i.e. on certain days more users move to certain locations or when users move to an event).

According to the results of this thesis mobile phone networks are a promising source of data for road traffic. Applications relying on these data will most likely not be able to accurately identify jams or trends but they will be able to give a good hint where a traffic jam is most likely to find and when it is over. The results of these applications then need to be verified

with data from a second source like cameras. The situation is similar for trend detection. If mobile phone network data indicate a movement trend they give a hint where to search for it.

Methods to improve these applications are discussed in this chapter.

6.2.1 Using Location Area Updates to Compute Trajectories

Johannes Schlaich presented an algorithm to compute trajectories of movements and to feature them using string comparison techniques in his doctor thesis "Nutzung von Mobilfunkdaten für die Analyse der Routenwahl"[9]. The application of his method is impossible with the data provided in this thesis because an event report about a Location Area Update does not include the Location Area Code but the Location Area Code for each location could be probably reverse engineered using the method described in chapter 2.1.3. Are more convenient way would be that the mobile network operator provides the code in his event reports.

After knowing the codes the layout of the Location Areas has to be respected. Location Areas are not like nations on a map which perfectly fit next to each others. There is overlapping as shown in Figure 38.

Figure 38: Four overlapping Location Areas

When more data than only Location Area Updates are available it is probably not necessary to use them explicitly in the methods. An advantage of Location Area Updates is that they are triggered by moving (from one area into another). The disadvantages are the unreliability (possible nervosities, no network coverage, etc.) and their wide range.

6.2.2 Creating Traffic Matrices Using Markov Models

The Markov Models presented in chapter 3.3.1 show the likelihood that a user triggers an event on a certain location after he or she triggered an event on another location before. This location change is most likely because the user moved. Other reasons like nervosities and circles can be filtered using methods from chapter 4.1.2. A model cleaned from nervosities and circles therefore shows the likelihood of moves from one location to another.

Traffic matrices (also transport matrix or origin-destination matrix/OD-matrix) show how many people move from one region to another. Such matrices are used for road planning and created by polling people. Markov Models from chapter 3.3.1 could be used to support OD matrix planning or to verify matrices created by polling.

6.2.3 Extending the Road Surveillance Algorithm

The road surveillance algorithm of chapter 5 only exploits little information of received event data. Considering the results of the Markov Model for event sequences in the appendix and connecting the feature values better with other data like new users in the area of interest or the ratio of movements to feature values traffic jam detection should be possible. The algorithm does not determine which event sequence created a movement feature. Movements which have a call set up as first event and then are followed by a handover indicate a person who is talking on the phone while moving. Sequences like that or an increase of short message events could indicate a traffic jam.

When doing an emergency call from an emergency call station at the roadway, the road administration is automatically informed about it. When doing an emergency call from a cell phone the administration does not know about it until informed by the authorities. An emergency call event in the area of interest could hint a person from the administration to check the part nearby the roadway for an accident. This way the administration becomes aware of the accident even though there was no call from the calling station.

Another consideration is the perfect set up for a route and the area of interest. When inspecting the traffic on a roadway like in chapter 5.2 the right area of interest and a good route set up are of essence. The results are featured by events at pre-defined locations (when considering only GSM and UMTS data). It should be possible to apply an artificial neural network or a support vector machine to learn the best parameters for an output that matches the trends or the numbers from permanent counting stations. Training and validation data are supplied by permanent counting stations. Parameters to adjust are the positions of way points as well as positions for points defining the area of interest. The input source is a combination of two locations. At least one of these locations has to be in the area of interest. There are at most $(k\text{-}1)*n$ input sources for n total possible locations and k locations in the area of interest.

The bottom line is that the algorithm of chapter 5 leaves pretty much room for extensions and improvements. Applying automatic pattern recognition on the data is discussed further in the next chapter.

6.2.4 Semantic Data Analysis

When discarding events from GPS sources then there are 18677 possible locations. At each location a user may trigger an event of several types. The events reported must be related to the environment somehow. Location Area updates for instance is the result of a movement; so are handovers. When looking at the iTraffic stream as a whole there must be a lot of pattern hidden in it which can describe the situation of the real world. These patterns and their meanings need to be discovered.

One approach could be to pin the 18677 possible locations on a map. Then classify the known events. Assign a color to each class. The more an event represents a class the more saturated is the color. Classes which only slightly represent a class are more opaque. If events represent two or more classes they become a mix of the colors. As soon as an event occurs at a certain location the location is painted in the mean value of all colors of events occurred at the location so far. This method should work very well if three classes are used and red, green and blue are the colors for the classes. The result at each location is a RGB color mix. When snap shots are taken at a given interval (i.e. every minute) trend changes should become visible. Which trend changes are shown depends on what is classified. A possible classification could be “communication events” – red, “movement events” – green “other events” – blue. Another classification could be “A interface related” – red, “Iu interface related” – green, “GPRS interface related” – blue. Clustering methods like EM-clustering (if clusters are Gaussians), DBSCAN (if there are no large differences in densities) could be used to process the results further.

Cluster analysis methods in general are expected to detect patterns. For instance an increase in events related to calls or short messages could be a hint on certain real life situations. The maximum-likelihood method could prove to be suitable if there are known situations which should be explained by the data. It is an efficient method to find parameters for a given model of normal-distributed data. This way parameters for algorithms could be detected or sequences of events could be detected which are a result of a real life situation somewhere.

Which method should be used depends on what is analysed. Events can have many reasons and causes. But also increases or decreases in detected nervosities might indicate certain real life events.

Processing the large amounts of data can be resource intensive. A better handling of the data is therefore a precondition. Better data processing strategies are discussed in chapter 6.3.

6.3 Better Data Processing

Processing the data could be optimized using databases or by transforming them into a suitable data format for later analyzes. In either ways parallelization is the key to more performance. Arguments and strategies for parallelization will be presented.

6.3.1 Using Databases

Most operations on the data, especially ordinary filtering operations like taking only events of a certain time span, are set operations. It is therefore self evident to use a database system which provides these operations as well as it provides the storage for and the access to the data. Operations on the data can then be performed via SQL queries. More complex operations like distance checks, bearing calculations and similar can be implemented as SQL functions and/or stored procedures.

In the very early phase of this thesis it was experimented with importing the data for a whole day into a MySQL database. The reference system was a server (AMD Sempron 2600+ and 1 GB RAM) with a 32bit Linux installed. The database engine was MySQL 5.0.51a-24+lenny2.

Table 16 shows the structure of a table used to store event blocks in the MySQL Database.

Field Name	Field Type
anonymous_id	VARBINARY(32)
timestamp	DATETIME
milliseconds	INT
longitude	FLOAT
latitude	FLOAT
radius	FLOAT
input_source	TINYINT
cell_id	INT
location_area_code	INT
event_type	TINYINT
cell_type	TINYINT
h_beam_width	SMALLINT
y_angle	INT
x_angle	INT
speed	INT

Table 16: The structure of a SQL table to store received events of any Option

1. Inserting the 217.038.823 entries into an empty MySQL table without indexes took about 12 hours. The insertions were done by a tool application. This application inserts the data via prepared statements.

2. Sorting of 217.038.823 entries without an index took 14 minutes and 7 seconds.

3. Filtering (searching and deleting) 66.551.727 entries out of 217.038.823 took 22 minutes and 18 seconds. These entries were events of a certain type. The event types were not indexed.

When indexes are used the more entries are already in the database the more time is consumed. There are about half a billion events each day. So the figures above and the fact that indexing has a significant bad influence on the performance have the usage of databases seeming a bad idea. This seem is corrected when the opportunities and remedies for the bad influence on the performance is considered.

The performance can be certainly improved by distributing the data to several databases. Each event is independent. This means there are no other data required to understand and interpret the information in an event. Therefore it is possible to spread the events over arbitrary databases (or to any storage in general). Since an anonymous Id maps an IMSI only for 24 hours each day should be recorded in a separate table. Of course it is possible to save all data in one table, but it makes no sense. Continuous operations like tracking a person, is only possible for 24 hours. Therefore saving more than 24 hours in one table would just cause losing performance but yield no benefit.

Such a table can be then spread over different databases. Since events do not contain unique ids when received from the mobile network provider it has to be assured that either each event is only saved once at one table or to create an Id that is unique among all tables as primary key.

The benefits of no primary keys and distinct events among all tables are that when receiving query results from each database, these results can be simply merged to one result set and since there is no redundancy the queries should run as fast as possible. The drawback are that if one table is lost all data of it are lost (unless there are no backups) and tagging event records which are phenomena isn't possible because the events of subject can't be found in the tables anymore.

The benefit of a primary key which is unique among all tables is the safety and integrity of the data. The data can then be stored redundant. Parts of one table can be saved into other tables. If one table is lost, its contents can be restored from the distributed parts of other tables. Since event records now have primary keys they can be tagged if they are part of a phenomenon. The drawback is that results from all tables have to be checked for double occurrences of a primary key before they can be processed.

Since data safety and backups can be solved with a redundant array of independent disks (RAID systems) the usage of primary keys is probably depreciated. All phenomena of chapter 2.2 involve two events so the first event may remain untagged, but the second event can be tagged as part of a phenomenon which should be sufficient to ignore it when processing the data. How the problem of delayed events (reported in chapter 2.2.2) is dealt with is an open question in this case. The events could be either saved afterwards into the table of the day they belong to or they could be dismissed.

Figure 39 illustrated the data flow described in this chapter. Computers are drawn as nodes – rectangles with round corners. On the right side there is the mobile network provider who sends the events via UDP/IP over the Internet (probably secured by a VPN tunnel). On the left side there are the nodes of the network which processes the data. One node for instance is responsible to receive the data from the mobile network provider. It does a preprocessing by for instance validating them, do small and fast statistics and probably buffering them. Then it distributes the data to all databases in its network. In other words the node with the receiving and validating application is responsible for the data management. The nodes with the

databases simply accept the data. Other possible tasks for these nodes are maintenance jobs. The node with the processing application is either the coordinator for distributed transactions or it is host for a simple application which send just on and the same SQL query to all databases, waits until all results are received and – in case there is no primary key – just unifies the results to one result set.

All procedures which can be parallelized could and should be packed in such a distributed SQL query. A complex parallel procedure could be implemented once as (extended) stored procedure and installed on all databases. A developer does not need to take care about the parallelization itself. Once the procedure works correctly in one database it is supposed to work in all other database and it is further supposed to yield a part of the complete result set on each database which is distinct from other partial result sets from other databases.

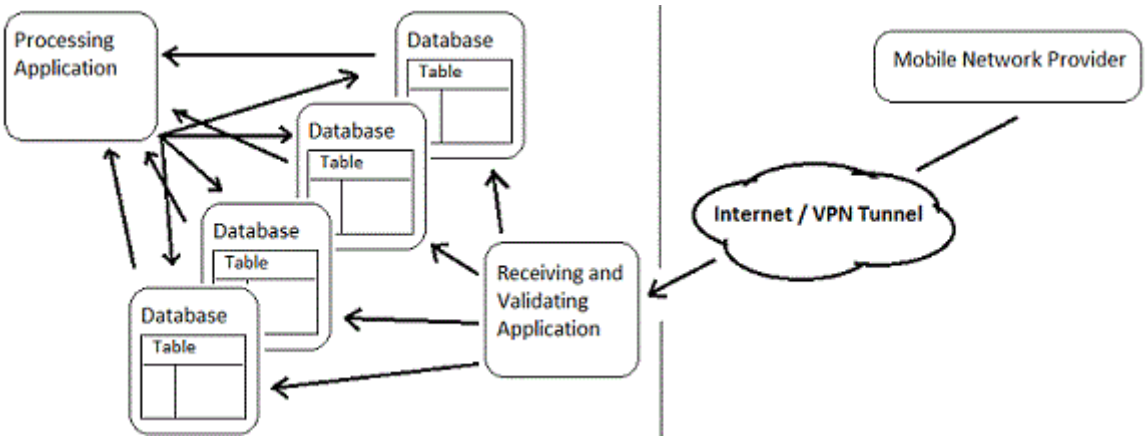


Figure 39: Example data flow for working with distributed tables and databases

6.3.2 A File Format for Received Events

A lot of information in the event reports received is redundant or it has an unnecessary overhead.

The anonymous Id for instance is encoded in 32 bytes and maps to one of about five million subscribers. Numbers from 0 to 4.294.967.295 can be encoded in a four byte long integer. If the anonymous Id only identifies a person, but the pseudonymized data concealed behind it are either not needed or not accessible it is unnecessary to store the whole 32 bytes. It would be more efficient to map every anonymous Id to a unique integer and then save the integer instead of the anonymous Id. The mapping has to be done from the scratch for every day since the anonymous Ids change at 00:00 AM. Every days is therefore saved into a new file or a different storage.

Further there are the three different input sources. The input sources for the 2G and 3G network always have about 18677 distinct locations (as stated in chapter 2.1.3). Only the third

input source has distinct positions. The sources could therefore be treated like different channels and stored therefore in different channels (i.e. a different file for each input source).

When the input sources are saved into different files it is further possible to improve the storage of the reported coordinates. For those channels where the positions are not unique, but member of a set (like those for 2G and 3G) a mapping could be created. In this mapping each longitude and latitude tuple is mapped to a two byte short integer key (which can identify up to 65.535 tuples/entries). Instead of the coordinates only the key is saved and the mapping is saved and maintained in a different file or storage. For channels where the positions are unique the longitudes and latitudes are stored as they are received.

Applying the strategies described before almost 59 bytes could be saved for each event report: *32 bytes for the anonymous Id plus 2*16 bytes for longitude/latitude replaced by 4+2 bytes for keys -1 byte for the input source Id*. The required storage would reduce to a half.

Another feature of the strategy is that the size for event reports remains static. Reports from 2G and 3G network would have 29 bytes (including unused information), reports from GPS notification would have 59 bytes (including unused information). Most of the events are received in an order according to the time they occurred. Since all events are blocks with a fixed size, events which are reported delayed could by inserted afterwards with relative small effort (compared to the effort needed if the blocks had different sized). It would involve replacing blocks only.

Last but not least also this file format could be stored distributed as long as there is one application which keeps track of the anonymous Id to integer mapping for one day. A copy of the mapping to coordinates should be saved at each place where the events are saved. This makes sure events can be processed distributed and parallel when their coordinates are involved.

6.3.3 Parallelizing the Road Surveillance Algorithm

The road surveillance algorithm presented in chapter 5 is designed to be an online algorithm. In other words the algorithm receives events, processes them and updates the situation in the area of interest. The computations of the algorithm and especially the memory usage are everything else but small. Because the algorithm has to validate the information of the event data it has to keep the last event of every user who reports an event (even from those outside the area of interest because they may move in the area). For users in the area of interest at least the last three locations for every user have to be saved and sorted according to their occurrence time. The information has to be kept for every user who is in the area of interest until he or she moves out or the information expires. Computing the feature involves several sub routines like checking for nervosities, allocating to segment, etc. The bottom line is that a computer with weaker hardware (especially small memory) may lag behind as soon as more data are received. This can be avoided by distributing the computations and splitting the steps up.

One strategy is to simply distribute the anonymous ids over several computers. So every computer receives only certain anonymous ids. This strategy is fast to implement and it does not require changes in the surveillance algorithm itself. The distributing computer just has to keep track of who receives events for which anonymous id and it should use a smart algorithm to compute the load balancing. Not every user causes the same amount of events.

Another strategy is to split up the procedures. The very first procedure is to validate the information and to detect phenomena. The computer who performs this task has to have enough memory (RAM) to keep the last event of every user and it has to have enough CPU power to validate each event as fast as possible. Sourcing the validation out on a separate computer has the advantage that if other applications rely on validated data, they can receive them from the validating computer. So the validating computer can be also a distributing computer.

The second process is to find out whether an event was triggered inside the area of interest. This process is at least $O(n)$ complex if the polygon is defined by n points. If a polygon has many points an improvement of the point-in-polygon test should be considered at first hand. On the second hand testing if an event is in an area of interest could be also out-sourced on a separate computer. This option should be especially considered if there are more areas of interest observed. A list of users who have been in an area of interest has to be kept as well, but this list is supposed to be always smaller than the list of previous events from all users. A computer which keeps track of users in an area of interest is also a distributing computer if it has more than one area of interest.

The actual computing of the features can be done on one or more other computers which are responsible for distinct anonymous ids. These computers could send their results to a presentation computer (like a web server) which simply shows the result it receives.

The advantage of this strategy is the reusability of results like validated information or events which are definitely in an area of interest. This way the algorithm can be extended in most ways so that it does not affect working applications.

6.4 Conclusion

Data from mobile phone networks seem to be a suitable source to assist traffic monitoring as well as they seem suitable to assist traffic planning (considering the Markov Model for location changes). There is no standard for delivering such data so any method which is implemented has to adapt to the characteristics (like phenomena) of the data source. Further it can be said that the reported locations are fuzzy data. But still it was possible to create an algorithm which reports the situation on the road. Of course the algorithm as well as other methods described in this thesis can be improved further, but the results and the methods are a good base for further applications.

7 Literature

[1] H. Fiby, „Workshop Verkehrsleitzentrale," 2009.

[2] FUTUREZONE GmbH, „TomTom: Echtzeitdaten gegen Gratis-Navis," 25 02 2011. [Online]. Available: http://futurezone.at/produkte/1986-tomtom-echtzeitdaten-gegen-gratis-navis.php. [Zugriff am 12 02 2012].

[3] Google Inc., „Time in current traffic," [Online]. Available: http://support.google.com/maps/bin/answer.py?hl=en&answer=2549020&topic=1687356&ctx=topic. [Zugriff am 6 9 2012].

[4] Autobahnen und Schnellstraßen Finanzierungs-Aktiengesellschaft, „Dauerzählstellen," 12 02 2012. [Online]. Available: http://www.asfinag.at/weitere-services/dauerzaehlstellen. [Zugriff am 12 02 2012].

[5] A1 Telekom Austria AG, „Das Netz von A1," 12 02 2012. [Online]. Available: http://www.a1.net/hilfe-support/netzabdeckung/. [Zugriff am 12 02 2012].

[6] H. G. Feiertag, *Mobilitätsrohdatengewinnung mittels Mobilfunkendgeräten und Verarbeitung in Informationssystemen,* Technical University Graz, 2005.

[7] F. D. R. Bolla, „Road Traffic Estimation from Location Tracking Data in the Mobile Cellular Network," Genoa, 2000.

[8] L. Wood, „GSM overview," 24 April 2007. [Online]. Available: http://personal.ee.surrey.ac.uk/Personal/L.Wood/constellations/tables/gsm.html. [Zugriff am 1 Januar 2012].

[9] J. Schlaich, „Nutzung von Mobilfunkdaten für die Analyse der Routenwahl," 24 10 2009. [Online]. Available: http://elib.uni-stuttgart.de/opus/volltexte/2010/5102/pdf/Dissertation_Schlaich.pdf. [Zugriff am 03 01 2011].

[10] D. Valerio, „Road Traffic Information from Cellular Network Signaling," Forschungszentrum Telekommunikation Wien, Vienna, 2009.

[11] European Telecommunications Standards Institute, „ETSI," 2011. [Online]. Available: http://www.etsi.org/WebSite/homepage.aspx. [Zugriff am 24 9 2012].

[12] 3rd Generation Partnership Project, „3GPP," 2012. [Online]. Available: http://www.3gpp.org/. [Zugriff am 24 9 2012].

[13] Cisco Systems, Inc., „Overview of GSM, GPRS, and UMTS,“ 29 12 2002. [Online]. Available: http://docstore.mik.ua/univercd/cc/td/doc/product/wireless/moblwrls/cmx/mmg_sg/cmxgsm.htm. [Zugriff am 5 Februar 2012].

[14] P. S. P.-J. P. Emmanuel Seurre, GPRS for mobile internet, Artech House, 2003.

[15] D. Valerio, „Exploiting cellular networks for road traffic: a survey and a research roadmap,“ IEEE VTC2009-Spring, Barcelona, 2009.

[16] TIS.Kis mobile messaging company, "iTraffic by A1 Output Stream Specification," 2009.

[17] D. E. Knuth, The Art of Computer Programming: Volume 2: Seminumerical Algorithms, Amsterdam: Addison-Wesley Longman, 1997.

[18] G. H. G. R. J. L. Tony F. Chan, „Updating Formulae and a Pairwise Algorithm for Computing Sample Variances,“ Standford University, 1979.

[19] R. Setchi, „Knowledge-based and Intelligent Information and Engineering Systems: 14th International Conference, Kes 2010, Cardiff, Uk, September 8-10, 2010, Proceedings,“ in s *Knowledge-based and Intelligent Information and Engineering Systems*, Springer, 2010, pp. 413-416.

[20] T. C. Urdan, „When to Use Correlation and What It Tells Us,“ in s *Statistics In Plain English*, Routledge, 2005, pp. 75-83.

[21] S. V. Vaseghi, Advanced Digital Signal Processing and Noise Reduction, Brunel: John Wiley & Sons, 2008.

[22] Robert.G.Chamberlain und William.H.Duquette, „Some Algorithms for Polygons on a Sphere,“ Jet Propulsion Laboratory California Institute of Technology, Pasadena, California, 2007.

[23] Microsoft Corporation, „Bounding Box, Radius, and Polygon Search,“ [Online]. Available: http://msdn.microsoft.com/en-us/library/cc451895. [Zugriff am 04 03 2012].

[24] D. R. Finley, „Point-in-Polygon Algorithm—Determining Whether a Point Is Inside a Complex Polygon,“ 2007. [Online]. Available: http://alienryderflex.com/polygon/. [Zugriff am 02 04 2012].

[25] AT&T Research, „Graphviz - Graph Visualization Software,“ [Online]. Available: http://www.graphviz.org/Home.php. [Zugriff am 12 05 2012].

[26] R. Riemer, „UMTSLink.at Mobilfunktechnik trasnparent & interessant,“ 2011. [Online]. Available: http://www.umtslink.at/. [Zugriff am 29 12 2011].

[27] G. Sanders, GPRS networks, West Sussex PO19 8SQ, England: John Wiley and Sons, 2003.

[28] pressetext Nachrichtenagentur GmbH, „A1 Traffic Data Stream,“ 17 12 2009. [Online]. Available: http://www.pressetext.com/news/20091217019. [Zugriff am 01 01 2012].

[29] A1 Telekom Austria AG, „Home,“ [Online]. Available: http://www.a1.net. [Zugriff am 01 01 2012].

[30] C. Johnson, Radio Access Networks for Umts: Principles and Practice, United Kingdom: John Wiley and Sons, 2008.

[31] 1 Feb 2012. [Online]. Available: http://www.osmocom.org/. [Zugriff am 26 Januar 2012].

[32] M. Buchheit, „Die Sprache C# im Detail,“ 14 Juni 2004. [Online]. Available: http://msdn.microsoft.com/de-de/library/bb979023.aspx. [Zugriff am 03 Mai 2012].

[33] Microsoft Corporation, „ServiceBase Class,“ 2012. [Online]. Available: http://msdn.microsoft.com/en-us/library/system.serviceprocess.servicebase.aspx. [Zugriff am 21 05 2012].

[34] Microsoft Corporation, „Installer Class,“ 2012. [Online]. Available: http://msdn.microsoft.com/en-us/library/system.configuration.install.installer%28v=vs.71%29.aspx. [Zugriff am 21 05 2012].

[35] High Beam Research LLC, „http://www.infoplease.com/ipa/A0933605.html,“ 2009. [Online]. Available: http://www.infoplease.com/ipa/A0933605.html. [Zugriff am 4 10 2011].

[36] WiGeoGIS GmbH, „A1 Traffic Data Stream: Movement Data in Mobile Telephone Network as Data Source for Marketing, Research and Planning,“ 17 12 2009. [Online]. Available: http://www.wigeogis.com/en/pdf/news/NEWS17122009.pdf. [Zugriff am 2 Dec 2011].

[37] S. St.Pölten, „Landeshauptstadt St. Pölten, Niederösterreich,“ 19 7 2012. [Online]. Available: http://www.st-poelten.gv.at/Content.Node/buergerservice/politik/Gemeinderat-Juli-2012.at.php. [Zugriff am 20 8 2012].

[38] H. M. R. K. F. M. K. H. D. W. E. e. a. Sammer Gerd, „Handbuch für Mobilitätserhebungen,“ Bundesministerium für Verkehr, Innovation und Technologie, Wien, 2011.

8 Appendix

Figure 40 shows the actual numbers of received MB per hour, detected devices per hour and caused KB per device. Each row has the number for a day from 01.03.2012 until 06.03.2012. The last two rows in each table contain the mean value and the standard deviation. In the columns there are the hours of each day. The received MB per hour and the detected devices per hour have a correlation coefficient of 0.9779.

Date	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00	08:00	09:00
01.03.2012	637	526	480	459	486	654	1.256	2.290	2.750	2.938
02.03.2012	578	468	425	423	431	545	1.050	1.874	2.267	2.463
03.03.2012	707	562	479	424	400	413	528	845	1.351	1.779
04.03.2012	767	629	545	478	421	382	410	556	862	1.218
05.03.2012	470	388	354	343	377	568	1.132	1.970	2.350	2.801
06.03.2012	624	524	482	454	477	654	1.276	2.254	2.726	2.892
Mean (μ)	631	516	461	430	432	536	942	1.632	2.051	2.349
StdDev (σ)	103	82	65	48	43	116	377	744	773	702

Date	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00	08:00	09:00
01.03.2012	1.062.063	997.252	963.530	940.808	978.179	1.103.854	1.511.434	1.944.410	2.049.684	2.119.097
02.03.2012	957.909	892.292	858.473	893.616	865.388	899.036	1.255.011	1.700.490	1.853.725	1.967.864
03.03.2012	929.587	816.110	750.066	694.257	684.938	694.945	823.379	1.136.915	1.522.014	1.772.346
04.03.2012	938.878	830.186	763.105	698.135	670.839	649.749	697.366	870.696	1.153.439	1.409.362
05.03.2012	761.098	701.399	668.615	631.966	666.672	817.465	1.254.846	1.718.244	1.841.187	2.041.156
06.03.2012	1.063.900	1.034.757	965.404	925.208	962.708	1.103.810	1.537.666	1.948.480	2.023.082	2.072.992
Mean (μ)	952.239	878.666	828.199	797.332	804.787	878.143	1.179.950	1.553.206	1.740.522	1.897.136
StdDev (σ)	111.165	123.531	121.544	137.127	148.363	195.814	349.032	446.616	343.739	268.116

Date	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00	08:00	09:00
01.03.2012	0,6142	0,5401	0,5101	0,4996	0,5088	0,6067	0,8509	1,2060	1,3739	1,4197
02.03.2012	0,6179	0,5371	0,5069	0,4847	0,5100	0,6208	0,8567	1,1285	1,2523	1,2816
03.03.2012	0,7788	0,7052	0,6539	0,6254	0,5980	0,6086	0,6567	0,7611	0,9089	1,0278
04.03.2012	0,8365	0,7758	0,7313	0,7011	0,6426	0,6020	0,6020	0,6539	0,7653	0,8850
05.03.2012	0,6323	0,5665	0,5422	0,5558	0,5791	0,7115	0,9238	1,1740	1,3070	1,4052
06.03.2012	0,6006	0,5186	0,5113	0,5025	0,5074	0,6067	0,8497	1,1846	1,3798	1,4286
Mean (μ)	0,6801	0,6072	0,5760	0,5615	0,5576	0,6260	0,7900	1,0180	1,1645	1,2413
StdDev (σ)	0,1010	0,1067	0,0945	0,0858	0,0574	0,0423	0,1286	0,2442	0,2619	0,2315

MB recieved per hour

10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
2.993	3.017	3.011	3.143	3.122	3.152	3.107	2.971	2.703	2.250	1.856	1.510	1.147	771
2.538	2.622	2.672	2.789	2.735	2.711	2.705	2.649	2.519	2.135	1.767	1.443	1.209	907
2.002	2.079	2.065	2.032	2.031	1.996	2.011	2.063	2.052	1.827	1.570	1.311	1.148	930
1.497	1.649	1.637	1.661	1.739	1.767	1.809	1.881	1.918	1.772	1.542	1.184	933	627
2.980	3.053	2.990	3.136	3.003	3.095	3.388	3.383	3.023	2.527	2.089	1.633	1.206	793
2.937	2.970	2.949	3.058	3.063	3.161	3.339	3.367	3.056	2.544	2.097	1.660	1.223	804
2.491	2.565	2.554	2.637	2.616	2.647	2.727	2.719	2.545	2.176	1.820	1.457	1.144	805
620	581	574	636	588	620	680	642	480	332	242	185	108	109

Unique Devices Received per Hour

10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
2.159.156	2.145.209	2.179.079	2.214.030	2.173.382	2.193.471	2.247.934	2.167.443	2.051.564	1.847.029	1.622.038	1.413.347	1.228.622	1.032.529
2.007.171	2.032.013	2.086.480	2.110.557	2.079.175	2.052.456	2.054.306	2.036.729	1.973.294	1.783.023	1.570.427	1.362.489	1.238.394	1.039.544
1.865.210	1.868.750	1.843.813	1.827.150	1.813.143	1.795.929	1.799.003	1.798.772	1.748.636	1.600.171	1.440.060	1.265.537	1.175.072	1.175.072
1.553.003	1.613.377	1.571.388	1.562.493	1.604.850	1.630.352	1.635.215	1.661.944	1.659.700	1.563.108	1.417.801	1.185.057	1.044.157	851.056
2.111.000	2.128.125	2.150.320	2.182.420	2.116.876	2.161.933	2.237.449	2.229.157	2.093.094	1.897.555	1.712.179	1.484.225	1.310.189	1.118.194
2.080.420	2.081.454	1.795.367	2.142.852	2.112.968	1.831.781	2.218.457	2.222.726	2.103.696	1.903.317	1.691.645	1.488.556	1.310.666	1.124.125
1.962.660	1.978.155	1.937.741	2.006.584	1.983.399	1.944.320	2.032.061	2.019.462	1.938.331	1.765.701	1.575.692	1.366.535	1.217.850	1.056.753
225.291	204.451	240.360	257.971	224.569	225.560	259.163	238.331	189.188	149.454	124.559	121.723	99.678	114.462

KB Received per Device Ratio

10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
1,4195	1,4401	1,4149	1,4537	1,4709	1,4715	1,4153	1,4036	1,3492	1,2474	1,1717	1,0940	0,9560	0,7646
1,2948	1,3213	1,3114	1,3532	1,3470	1,3526	1,3483	1,3318	1,3072	1,2261	1,1522	1,0845	0,9997	0,8934
1,0991	1,1392	1,1468	1,1388	1,1470	1,1381	1,1447	1,1744	1,2016	1,1692	1,1164	1,0608	1,0004	0,8104
0,9871	1,0466	1,0668	1,0886	1,1096	1,1098	1,1328	1,1590	1,1834	1,1608	1,1137	1,0231	0,9150	0,7544
1,4455	1,4690	1,4239	1,4714	1,4526	1,4659	1,5506	1,5540	1,4789	1,3637	1,2494	1,1266	0,9426	0,7262
1,4456	1,4611	1,6820	1,4613	1,4844	1,7671	1,5412	1,5512	1,4875	1,3687	1,2694	1,1419	0,9555	0,7324
1,2819	1,3129	1,3410	1,3278	1,3353	1,3842	1,3555	1,3623	1,3346	1,2560	1,1788	1,0885	0,9615	0,7803
0,1965	0,1809	0,2202	0,1720	0,1679	0,2440	0,1845	0,1742	0,1309	0,0915	0,0664	0,0434	0,0333	0,0630

Figure 40: Actual numbers of received MB per hour, detected devices per hour and KB caused by device ratio

Table 17 shows the absolute numbers of unique devices detected per hour from the 21.02.2012 until the 26.02.2012. The mean values of the number of detected devices from 01.03.2012 until 06.03.2012 (Figure 40) correlate with the mean values of Table 17 with a coefficient of 0.9875.

Unique Devices Detected per Hour (Feb)

	21.02.2012	22.02.2012	23.02.2012	24.02.2012	25.02.2012	26.02.2012	Mean (μ)	StdDev (σ)
00:00	1.046.580	1.065.821	1.034.333	1.047.554	1.116.609	1.141.363	1.075.376,7	43.436,3
01:00	1.003.109	1.013.867	977.668	970.719	1.031.924	1.081.306	1.013.098,8	40.394,0
02:00	953.875	959.629	941.393	942.756	978.359	1.028.649	967.443,5	32.858,9
03:00	919.187	914.394	901.639	928.092	926.869	952.917	923.849,7	17.184,2
04:00	947.318	955.478	961.199	954.462	934.377	936.798	948.272,0	10.798,0
05:00	1.069.456	1.072.852	1.076.124	1.072.616	977.564	932.160	1.033.462,0	62.589,0
06:00	1.402.653	1.411.098	1.411.171	1.393.628	1.073.791	978.428	1.278.461,5	197.889,5
07:00	1.807.931	1.841.070	1.823.581	1.804.133	1.339.826	1.134.107	1.625.108,0	307.887,4
08:00	1.997.491	2.000.451	1.994.231	1.990.270	1.669.585	1.381.756	1.838.964,0	259.205,9
09:00	1.056.911	2.073.582	1.049.507	2.077.866	1.896.269	1.608.174	1.627.051,5	476.251,1
10:00	2.157.874	2.154.750	2.121.569	2.063.858	2.003.848	1.738.025	2.039.987,3	159.235,9
11:00	2.182.170	2.167.326	2.111.875	1.989.047	2.015.562	1.807.581	2.045.593,5	140.438,3
12:00	2.209.033	2.167.181	2.124.378	2.026.422	1.962.647	1.745.023	2.039.114,0	170.318,2
13:00	2.198.165	2.182.920	2.147.425	2.054.514	1.942.995	1.712.212	2.039.705,2	186.573,0
14:00	2.125.740	2.145.606	2.119.136	2.053.055	1.926.717	1.745.826	2.019.346,7	156.044,2
15:00	2.113.529	2.161.786	2.142.100	2.159.159	1.915.143	1.737.579	2.038.216,0	174.293,0
16:00	2.134.595	2.230.472	2.202.299	2.153.656	1.915.945	1.793.347	2.071.719,0	175.884,7
17:00	2.123.632	2.206.923	2.193.713	2.136.728	1.932.156	1.828.108	2.070.210,0	154.203,9
18:00	2.023.083	2.056.313	2.058.404	2.030.969	1.877.516	1.816.610	1.977.149,2	103.513,6
19:00	1.854.877	1.849.497	1.864.568	1.865.029	1.732.213	1.734.634	1.816.803,0	64.858,3
20:00	1.663.211	1.665.032	1.670.565	1.661.962	1.582.588	1.583.024	1.637.730,3	42.646,0
21:00	1.461.659	1.453.938	1.491.653	1.478.580	1.452.957	1.373.839	1.452.104,3	41.196,8
22:00	1.342.282	1.335.265	1.319.357	1.374.279	1.353.071	1.270.856	1.332.518,3	35.340,4
23:00	1.178.906	1.134.358	1.151.522	1.253.753	1.265.918	1.096.990	1.180.241,2	67.241,5

Table 17: Unique devices detected per hour from 21.02.2012 to 26.02.2012

Table 18, Table 19, Table 20, Table 21 and Table 22 have actual figures to the diagrams in chapter 4.1.3. Blue rows show rows with no standard deviation, red rows show rows with standard deviations for circles and for nervosities.

Hour	Nervosities	Circles	Movements	Total	*Nervosity μ*	*Circle μ*	*Nervosity σ*	*Circle σ*
00:00	17.609	408	16.118	34.135	*0,4658*	*1,0812*	*0,0000*	*0,0000*
01:00	12.742	401	8.279	21.422	*0,4672*	*5,6094*	*0,0000*	*0,2333*
02:00	9.779	280	6.171	16.230	*1,4835*	*0,9433*	*0,0133*	*0,1626*
03:00	8.387	294	7.613	16.294	*0,3619*	*1,5649*	*0,0000*	*0,0698*
04:00	10.193	358	11.017	21.568	*9,7017*	*1,7649*	*0,0781*	*0,0000*
05:00	22.974	771	36.720	60.465	*0,4668*	*3,8616*	*0,0143*	*0,0877*
06:00	72.153	2.380	112.161	186.694	*0,5474*	*6,1513*	*0,0000*	*0,1157*
07:00	158.247	5.968	215.084	379.299	*0,4663*	*1,0888*	*0,0000*	*0,0000*
08:00	188.797	9.373	190.147	388.317	*2,2846*	*6,5501*	*0,0040*	*0,0537*
09:00	206.722	11.182	166.024	383.928	*0,0975*	*0,6266*	*0,0000*	*0,0131*
10:00	213.830	12.520	159.410	385.760	*0,1240*	*1,3717*	*0,0000*	*0,0000*
11:00	220.492	13.305	159.152	392.949	*0,9593*	*0,7459*	*0,0000*	*0,0000*
12:00	224.255	13.902	165.996	404.153	*0,2481*	*0,3026*	*0,0000*	*0,0222*
13:00	229.737	14.313	178.201	422.251	*1,1013*	*0,7225*	*0,0000*	*0,0000*
14:00	224.867	13.634	184.572	423.073	*0,2242*	*0,6865*	*0,0000*	*0,0000*
15:00	226.070	13.451	199.208	438.729	*0,5968*	*0,7013*	*0,0000*	*0,0000*
16:00	229.808	14.332	222.945	467.085	*0,6461*	*5,8196*	*0,0000*	*0,0428*
17:00	212.940	12.978	219.577	445.495	*0,4663*	*1,6607*	*0,0000*	*0,0097*
18:00	189.603	12.162	183.886	385.651	*0,6927*	*29,1423*	*0,0000*	*0,2614*
19:00	155.175	9.746	130.507	295.428	*0,4768*	*0,9402*	*0,0000*	*0,0238*
20:00	124.966	7.695	87.729	220.390	*0,3617*	*1,2932*	*0,0042*	*0,0000*
21:00	92.358	5.856	61.888	160.102	*4,5179*	*1,3787*	*0,0141*	*0,0000*
22:00	65.480	4.283	45.514	115.277	*4,0903*	*0,9941*	*0,0092*	*0,1081*
23:00	38.701	2.546	27.999	69.246	*0,9602*	*1,7437*	*0,0000*	*0,0000*
∑=	3.155.885	182.138	2.795.918	6.133.941	*0,4658*	*1,0807*	*0,0000*	*0,0000*

Table 18: Classification results

Hour	Nervosities	Circles	Movements	Total	*Nervosity μ*	*Circle μ*	*Nervosity σ*	*Circle σ*
00:00	10.549	1.374	29.177	41.100	*0,4658*	*0,5674*	*0,0000*	*0,0000*
01:00	8.009	975	17.302	26.286	*0,4672*	*0,8253*	*0,0000*	*0,0000*
02:00	6.235	629	13.086	19.950	*0,4671*	*0,6480*	*0,0000*	*0,0000*
03:00	5.070	561	14.393	20.024	*0,3619*	*0,5610*	*0,0000*	*0,0000*
04:00	5.420	720	20.950	27.090	*0,4129*	*0,9103*	*0,0000*	*0,0000*
05:00	9.495	1.692	65.769	76.956	*0,4667*	*0,5585*	*0,0000*	*0,0000*
06:00	28.294	6.020	206.352	240.666	*0,1725*	*0,5585*	*0,0000*	*0,0000*
07:00	72.558	15.173	392.342	480.073	*0,4663*	*0,5541*	*0,0000*	*0,0000*
08:00	94.631	20.851	374.621	490.103	*0,3059*	*0,7632*	*0,0000*	*0,0000*
09:00	107.753	24.524	352.781	485.058	*0,0975*	*0,7120*	*0,0000*	*0,0000*
10:00	111.257	25.883	353.135	490.275	*0,1240*	*0,7408*	*0,0000*	*0,0000*

11:00	116.741	26.526	357.541	500.808	*0,4406*	*0,7459*	*0,0000*	*0,0000*
12:00	119.321	26.604	368.861	514.786	*0,2481*	*0,3024*	*0,0000*	*0,0000*
13:00	119.608	27.086	392.986	539.680	*0,2754*	*0,7225*	*0,0000*	*0,0000*
14:00	114.198	26.021	402.169	542.388	*0,2242*	*0,5585*	*0,0000*	*0,0000*
15:00	112.313	25.388	425.950	563.651	*0,3107*	*0,5968*	*0,0000*	*0,0000*
16:00	112.426	25.004	462.852	600.282	*0,3124*	*0,6461*	*0,0000*	*0,0000*
17:00	101.953	21.680	452.924	576.557	*0,4663*	*0,6697*	*0,0000*	*0,0000*
18:00	91.835	19.437	388.581	499.853	*0,4670*	*0,6927*	*0,0000*	*0,0000*
19:00	76.445	16.573	290.639	383.657	*0,4451*	*0,9399*	*0,0000*	*0,0000*
20:00	62.687	13.539	212.410	288.636	*0,3616*	*0,5733*	*0,0000*	*0,0000*
21:00	47.077	9.660	154.193	210.930	*0,4664*	*0,5543*	*0,0000*	*0,0000*
22:00	34.374	6.705	109.870	150.949	*0,4664*	*0,5195*	*0,0000*	*0,0000*
23:00	21.532	3.398	64.465	89.395	*0,3189*	*0,6087*	*0,0000*	*0,0000*
∑=	1.589.781	346.023	5.923.349	7.859.153	*0,4658*	*0,5673*	*0,0000*	*0,0000*

Table 19: Classification results for the diagram in Figure 23

Hour	Nervosities	Circles	Movements	Total	*Nervosity μ*	*Circle μ*	*Nervosity σ*	*Circle σ*
00:00	2.304	53	3.743	6.100	*2,5209*	*2,2008*	*0,0324*	*0,5933*
01:00	1.511	44	1.712	3.267	*0,9273*	*5,5145*	*0,1223*	*0,7066*
02:00	1.259	43	1.406	2.708	*0,4408*	*1,3371*	*0,0000*	*2,7880*
03:00	1.286	36	2.035	3.357	*0,8990*	*2,4016*	*0,0000*	*1,9882*
04:00	1.841	96	3.384	5.321	*0,8479*	*1,7579*	*0,0000*	*0,0000*
05:00	5.828	214	13.366	19.408	*1,7837*	*3,9390*	*0,0000*	*1,1779*
06:00	18.093	576	38.802	57.471	*0,5474*	*4,2655*	*0,0000*	*0,1111*
07:00	31.780	1.278	57.971	91.029	*0,4663*	*2,1952*	*0,0000*	*0,0492*
08:00	39.650	2.066	52.704	94.420	*2,2846*	*6,5489*	*0,0094*	*0,0797*
09:00	41.079	2.306	43.573	86.958	*1,1263*	*0,4521*	*0,0000*	*0,1383*
10:00	40.875	2.391	40.231	83.497	*0,4675*	*3,1112*	*0,0050*	*0,0359*
11:00	42.915	2.698	40.107	85.720	*0,9593*	*1,5896*	*0,0000*	*0,0000*
12:00	42.724	3.043	41.222	86.989	*0,9042*	*13,0658*	*0,0000*	*0,2635*
13:00	45.951	3.242	45.373	94.566	*1,0637*	*0,8537*	*0,0096*	*0,1347*
14:00	48.438	3.253	50.421	102.112	*0,4786*	*5,1618*	*0,0070*	*0,0000*
15:00	49.762	3.449	54.653	107.864	*0,5968*	*0,7024*	*0,0141*	*0,0642*
16:00	52.774	3.994	65.839	122.607	*0,6461*	*40,3217*	*0,0121*	*0,6010*
17:00	49.973	3.752	67.804	121.529	*0,6106*	*1,5440*	*0,0000*	*0,0363*
18:00	42.174	3.306	52.556	98.036	*0,4090*	*25,5120*	*0,0000*	*0,3631*
19:00	32.709	2.551	35.377	70.637	*0,4768*	*6,6780*	*0,0000*	*0,1038*
20:00	25.267	1.847	22.488	49.602	*2,6874*	*18,4752*	*0,0000*	*0,2816*
21:00	18.033	1.441	15.432	34.906	*0,3898*	*6,1583*	*0,0000*	*0,1199*
22:00	11.943	969	9.991	22.903	*1,3115*	*5,2231*	*0,0000*	*0,1152*
23:00	6.261	615	5.458	12.334	*0,6388*	*5,2729*	*0,0146*	*0,0937*
∑=	654.430	43.263	765.648	1.463.341	*2,5216*	*2,1194*	*0,0000*	*0,0000*

Table 20: Classification results for the diagram in Figure 26

Hour	Nervosities	Circles	Movements	Total	*Nervosity μ*	*Circle μ*	*Nervosity σ*	*Circle σ*
00:00	1.922	113	4.431	6.466	*2,5208*	*4,1326*	*0,0354*	*0,0000*
01:00	1.306	82	2.044	3.432	*0,9238*	*3,7221*	*0,0000*	*0,1187*
02:00	1.101	71	1.675	2.847	*0,4408*	*0,9635*	*0,0000*	*0,4346*
03:00	1.081	60	2.407	3.548	*0,8990*	*2,6971*	*0,0000*	*0,0000*
04:00	1.531	112	4.047	5.690	*0,8479*	*1,7593*	*0,0000*	*0,0000*
05:00	4.669	306	15.758	20.733	*1,7837*	*3,8606*	*0,0000*	*0,0000*
06:00	14.547	905	45.844	61.296	*0,5474*	*3,7329*	*0,0000*	*0,0708*
07:00	26.387	2.023	67.833	96.243	*0,4663*	*2,1947*	*0,0000*	*0,0391*
08:00	33.682	2.981	63.123	99.786	*2,2846*	*0,9937*	*0,0102*	*0,0355*
09:00	35.487	3.246	52.893	91.626	*1,1263*	*2,7396*	*0,0000*	*0,0000*
10:00	35.274	3.385	49.475	88.134	*0,4675*	*3,1114*	*0,0053*	*0,0302*
11:00	37.179	3.619	49.895	90.693	*0,9593*	*3,2176*	*0,0000*	*0,0377*
12:00	37.052	3.791	51.310	92.153	*0,9042*	*2,4261*	*0,0000*	*0,0000*
13:00	39.543	4.124	56.674	100.341	*1,0636*	*5,1187*	*0,0000*	*0,0580*
14:00	41.518	4.139	63.041	108.698	*0,4786*	*1,7527*	*0,0075*	*0,0512*
15:00	42.427	4.194	68.918	115.539	*0,5968*	*0,7019*	*0,0058*	*0,0422*
16:00	44.150	4.833	82.511	131.494	*0,6461*	*2,5740*	*0,0000*	*0,0000*
17:00	41.242	4.532	85.016	130.790	*0,6106*	*3,9164*	*0,0000*	*0,0000*
18:00	34.917	4.028	66.521	105.466	*0,4090*	*2,7547*	*0,0000*	*0,0000*
19:00	27.257	3.131	45.614	76.002	*0,4768*	*3,3200*	*0,0000*	*0,0337*
20:00	21.375	2.317	29.707	53.399	*0,5734*	*2,6878*	*0,0086*	*0,0000*
21:00	15.003	1.692	21.369	38.064	*0,3898*	*2,6612*	*0,0000*	*0,0247*
22:00	10.030	1.136	13.704	24.870	*1,3115*	*2,8226*	*0,0000*	*0,0351*
23:00	5.241	626	7.651	13.518	*0,4889*	*0,9635*	*0,0180*	*0,0796*
∑=	553.921	55.446	951.461	1.560.828	*2,5216*	*4,1287*	*0,0000*	*0,0050*

Table 21: Classification results for the diagram in Figure 27

Hour	Nervosities	Circles	Movements	Total	*Nervosity μ*	*Circle μ*	*Nervosity σ*	*Circle σ*
00:00	583	122	7.418	8.123	*0,2555*	*0,9733*	*0,0000*	*0,0000*
01:00	500	98	3.781	4.379	*0,4664*	*0,8263*	*0,0000*	*0,0000*
02:00	372	83	3.190	3.645	*0,4405*	*0,5249*	*0,0000*	*0,0000*
03:00	360	95	3.862	4.317	*0,4663*	*0,7333*	*0,0000*	*0,0000*
04:00	515	106	6.312	6.933	*0,4661*	*0,5104*	*0,0000*	*0,0000*
05:00	1.205	341	23.521	25.067	*0,4672*	*0,9273*	*0,0000*	*0,0000*
06:00	3.567	1.069	70.398	75.034	*0,1725*	*0,7769*	*0,0000*	*0,0000*
07:00	7.415	2.205	110.026	119.646	*0,4663*	*0,7707*	*0,0000*	*0,0000*
08:00	10.444	3.294	113.495	127.233	*0,2181*	*0,7632*	*0,0000*	*0,0000*
09:00	11.690	3.625	103.689	119.004	*0,3814*	*0,4493*	*0,0000*	*0,0000*
10:00	11.603	3.680	100.160	115.443	*0,4674*	*0,7992*	*0,0000*	*0,0000*
11:00	12.134	3.883	103.892	119.909	*0,3175*	*0,8309*	*0,0000*	*0,0000*
12:00	12.416	3.943	104.754	121.113	*0,4785*	*0,8745*	*0,0000*	*0,0000*
13:00	12.919	4.113	115.571	132.603	*0,4673*	*0,7147*	*0,0000*	*0,0000*
14:00	13.551	4.322	124.118	141.991	*0,4785*	*0,5195*	*0,0000*	*0,0000*
15:00	13.785	4.264	132.067	150.116	*0,4672*	*0,5968*	*0,0000*	*0,0000*
16:00	14.221	4.152	151.615	169.988	*0,4664*	*0,6460*	*0,0000*	*0,0000*
17:00	12.853	3.599	152.364	168.816	*0,4669*	*0,9335*	*0,0000*	*0,0000*
18:00	10.864	2.939	124.232	138.035	*0,4090*	*0,5741*	*0,0000*	*0,0000*
19:00	8.500	2.398	89.994	100.892	*0,4657*	*0,8150*	*0,0000*	*0,0000*
20:00	7.027	1.975	63.025	72.027	*0,4422*	*0,5733*	*0,0000*	*0,0000*
21:00	4.806	1.327	45.286	51.419	*0,3898*	*0,7286*	*0,0000*	*0,0000*
22:00	3.232	916	29.516	33.664	*0,4664*	*0,6504*	*0,0000*	*0,0000*
23:00	1.736	437	15.994	18.167	*0,4886*	*0,6217*	*0,0000*	*0,0000*
∑=	176.298	52.986	1.798.280	2.027.564	*0,2552*	*0,9754*	*0,0000*	*0,0000*

Table 22: Classification results for the diagram in Figure 28

8.1 Interpretation of the Markov Models for Sequences

The list and the matrix are illustrated in graphs created by the tool Graphviz using the road map "neato", a multi scale version of the "spring layout"[25]. These graphs have labeled nodes in different colors and labeled edges in different colors. The more intense a color is the greater is the likelihood. The paler a color is the smaller is the likelihood. Another feature of this layout is that nodes are clustered according to the intense of their relations: "Neato attempts to minimize a global energy function, which is equivalent to statistical multi-dimensional scaling." [25]. Absolute numbers can also be found in end of the chapter.

8.1.1.1 Night Time Sequences

When starting to analyze the data at 00:00 AM there is a more than 40 % chance that the first event a user triggers is a Location Area update event. Location Area updates are to 54.18 % followed by another Location Area update. To 33 % there is no following event. The remaining 12.82 % are distributed over a few other events. When looking at the absolute numbers, the numbers of Location Area updates is constant. During the night time the ratio of

Location Area updates to devices is about 1.1 – so every device causes at least one area update.

Since Location Area updates occur when a mobile device enters a new location area, they indicate a moving user. Therefore since more than 40 % of the initial events are Location Area updates, during night time there is a good chance to find a moving subscriber!

Since the Location Area updates are hardly followed by another event than the Location Area update these subscribers seem to be moving subscribers who do not use their mobile device (in other words: The device is turned on and idle and the subscriber is moving).

Because of the device to Location Area update ratio it also seems that those devices which cause events are moving devices.

The initial probability for a handover is less than one percent. Handovers follow another handover with a probability of about 56 %. The absolute number of handovers declines during the night time and starts to rise again at the end of the night time.

The initial probability for other events from the A interface is about 5 %. 22 % of these events are followed by another event from the A interface. To 33 % there is no following event. 30 % of the handovers lead to an event from the A interface. As supposed by the nature of the GSM technology mobile originated or terminated calls, call set ups and connection acknowledges as well as short message service events also have a good chance (about 10 % to 20 %) to be followed by an event from the A interface. The mobile network provider did not explain what exactly the event from the A interface is.

In absolute numbers there are fewer (about three to five times fewer) reports from the A interface than Location Area updates. The number of reports from the A interface declines during the night time similar to the number of handovers. Their number also rises again at the end of night time.

The few numbers of event reports from the A interface and from events leading to reports from the A interface indicates a low usage of mobile devices.

The second highest initial probabilities have events from the Iu interface after Location Area updates. Other TMSI events have an initial probability of about 15 %, other TLLI event have an initial probability of about 8 %. What exactly other TLLI or TMSI events are is not explained by the mobile network provider. Other TMSI and TLLI events usually follow each other (with a chance of more than 50 % for each). With a chance of about 9 % they follow themselves and with a chance of less than 5 % the events aren't followed by any event. They are hardly lead or followed by events from the 2G network. Most previous events are either sent or received short messages from the Iu CS domain (about 20 %) followed by paging events in the Iu CS domain. From the packet switched domain the predecessors are detaching or attaching events (about 12 %). Routing area updates usually lead to TMSI/TLLI events. The periodic Routing Area updates follow about 2 % of the TLLI events and about 6 % of the updates are followed by TMSI events and about 9 % are followed by TLLI events again.

A Routing Area update indicates an active data connection. Therefore subscribers which trigger TLLI and TMSI event sequences seem to have an active data connection (their device is in dedicated mode).

About 12 % of the initial events found are from the Gb interface. About 5 % are cell updates another about 5 % are other events. 2 % are Routing Area updates. Less than 1 % is attaching to GPRS. Routing Area updates are followed to about 19 % by further Routing Area updates. Similar numbers for Routing Area updates are reported from the packet switched domain of the Iu interface. Cell updates are followed to 46 % by another cell update.

The low number of attaching to the GPRS and the high number of subsequent Routing Area updates and cell updates suggest a high movement rate of a few GPRS users.

8.1.1.2 Morning Time Sequences

There is a significant change in the likelihood of Location Area updates to be the first event of a user in this time span. The probability dropped from about 40 % down to about 20 %. The probability that one Location Area update is followed by another is almost unchanged. The absolute number of Location Area updates is almost the same than the number during the night time. There are no significant changes of the likelihoods of handovers. But looking at absolute numbers the amount of handovers sharply increased.

Other events like originating and terminating calls or short messages are now more likely to be the first events seen. The initial probability for events from the A interface is now at about 16 %.

The decreased likelihood of Location Area updates to be the first event of a user and the increased initial probability of other events indicate a higher usage of devices by the subscribers!

The increase of handovers with the increase events from the A interface indicates an increase of users who actively use their mobile phone while they are moving.

Remark: For unresolved matters of fact, other events from the A interface as well as call disconnects reported from the A interface seem to originate from the 3G network. Though the A interface is supposed to be an interface of the 2G network.

There is no significant change in the likelihoods reported from the Iu interface. Although the number of devices found in the morning time increases from hour to hour TMSI and TLLI events have almost the same initial probability and their likelihood to follow each other is almost the same.

The chance that the first event reported for a user is reported from the Gb interface only slightly increased. The chance that such an event is a cell update is now lower and initial probabilities for other events from the Gb interface is now higher. Significant is the change of the transition probability for Gb cell update. The chance to follow each other dropped about

20 %. In similar fashion the likelihood for Routing Area updates dropped about 7 %. The chance to be followed by another event from Gb increased for both events.

8.1.1.3 Day Time Sequences

During day time the initial probability for Location Area updates drops down to about 8 %. There is still a chance with more than 51% that a Location Area update is followed by another. The chance that one handover of an active call is followed by another handover about 57 %.

In absolute numbers Location Area updates are almost constant from 09:00 AM until 14:00 PM and then they start to decline. In the same time span the ratio of handovers to mobile originated and mobile terminated calls is less than 0.1. After 14:00 PM this ratio is slightly but constantly increasing up to 0.12 and the number of Location Area updates decreases sharply. The number of distinct devices remains almost constant at about 2.1 million ± 100.000.

The initial probability for other events from the A interface increases to about 18 %. Further there is a 4 % initial chance for mobile originated and mobile terminated calls. The chances for call set ups and call acknowledgements is also about 4 %.

A high initial probability for other events from the A interface indicates a high general use of mobile devices to do calls or texts. Comparing it to the initial probability for Location Area updates leads to the assumption that more users use their phones without moving (far enough to change a Location Area).

The initial probability of other TMSI events from the Iu interface is about 14 %, other TLLI events happen the first time for a user with a chance of about 7 %. The probability that an TMSI event is followed by another TLLI is still a bit more than 60 % while the probability that a TLLI even is followed by another TMSI event is still a bit less than 60 %. The statistics in absolute numbers show that after 15:00 PM the number of other TMSI and TLLI events increases. In the same time periodic Routing Area update reports start to decline and regular Routing Area update reports start to increase. The reporting interface is the Iu interface.

In the afternoon the mobility of users seem to increase compared to the mobility about noon. Further there seem to be many persons who are moving with or without using their phones.

Other events from the A interface have the highest initial probability and there is about a 32 % chance that they follow each other. There is no further significant number or pattern for this interface during day time.

There is no significant change of probabilities for events from the Gb interface.

8.1.1.4 Evening Time Sequences

In the evening time the initial chance for Location Area updates decline to about 5 %. The chance that a Location Area update is followed by another is only a bit more than 13 %. The

chance that a handover is followed by another handover drops down to a bit more than 23 %. In the statistics there are more handover events than location area updates at 18:00 PM and 19:00 PM from 20:00 PM until 24:00 PM handovers decline sharply and Location Area updates increase slightly. The ratio of handovers per mobile originated or terminated calls is 0.45 from 18:00 PM to 20:00 PM. From 21:00 PM to 24:00 PM the ratio goes from 0.44 to 0.41. The number of mobile phones starts to decline at 17:00 PM. The trend of fewer devices continues over the evening time into the night time until 04:00 AM. At 04:00 AM the number starts to increase again.

It seems that there are more subscribers using their mobile device while moving in the evening time from 18:00 PM to 24:00 PM. While there are fewer and fewer devices the ratio of handovers per mobile call is almost constant.

There are no significant changes in probabilities of events reported from the Iu interface and there is no significant change in probabilities of events reported from the Gb interface.

During the evening time the initial probability for events from the A interface is almost the same as during the day time. But the probability that another A events follows an A event is about 7 % higher in the evening time. The probability that another A event follows a mobile originated or terminated call, a short message service, a call set up or a connection acknowledgement is significant higher in the evening time than during day time.

Although the increase of the probabilities is significant there is no explanation why it happens. One reason why this fact cannot be explained here is that it is not known what "other events from the A interface" are in detail. No explanation was provided by the mobile network provider.

Figure 41 shows absolute numbers of Location Area updates in whole Austria from 21.02.2012 until 26.02.2012 and 01.03.2012 until 06.02.2012.

	Location Update Events (UMTS) per Hour								Location Update Events (GSM) per Hour							
	21.02.2012	22.02.2012	23.02.2012	24.02.2012	25.02.2012	26.02.2012	Mean (µ)	StdDev (σ)	21.02.2012	22.02.2012	23.02.2012	24.02.2012	25.02.2012	26.02.2012	Mean (µ)	StdDev (σ)
00:00	58.741	58.751	60.419	50.942	59.135	57.190	57.529,7	3.388,1	988.028	1.004.017	1.020.600	1.003.824	943.792	970.566	988.471,2	27.664,6
01:00	60.384	60.918	62.746	46.946	60.037	57.276	58.051,2	5.719,6	1.037.487	1.039.118	1.023.286	1.016.578	964.132	962.491	1.007.182,0	35.032,9
02:00	59.957	59.641	61.685	45.734	58.097	59.459	57.428,8	5.843,6	1.041.636	1.033.008	1.050.122	1.046.197	972.343	961.541	1.017.474,5	39.702,2
03:00	59.128	59.629	59.306	56.538	59.005	59.023	58.771,5	1.118,4	1.002.506	989.767	1.004.528	1.032.565	940.271	934.744	984.063,5	38.714,4
04:00	58.723	57.926	57.608	60.591	58.196	58.164	58.534,7	1.071,9	1.030.633	1.036.218	1.069.470	1.048.086	973.196	964.481	1.020.347,3	42.153,1
05:00	56.392	56.603	55.815	57.974	55.202	57.260	56.541,0	991,6	1.064.755	1.068.791	1.080.478	1.080.183	1.013.316	1.009.973	1.052.916,0	32.581,0
06:00	50.805	48.949	49.455	53.681	52.871	54.355	51.686,0	2.269,5	1.136.545	1.146.993	1.142.332	1.141.539	1.034.690	1.027.736	1.104.972,5	57.272,1
07:00	45.255	44.745	46.852	48.518	53.149	54.853	48.895,3	4.204,3	1.127.065	1.134.569	1.126.241	1.126.942	1.067.010	1.059.429	1.106.876,0	34.037,2
08:00	44.429	46.951	46.025	45.512	52.605	54.597	48.353,2	4.193,4	1.015.334	1.010.312	1.011.581	1.018.747	1.061.491	1.061.587	1.029.842,0	24.730,0
09:00	12.673	47.132	13.242	54.898	52.159	53.776	38.980,0	20.332,0	09:00	247.369	846.436	245.202	597.076	957.959	996.316	648.393,0	341.209,4
10:00	55.901	55.509	59.478	49.793	54.085	54.983	54.958,2	3.132,5	10:00	827.462	835.604	835.887	530.130	902.816	938.538	811.739,5	144.966,6
11:00	52.283	55.023	54.410	39.665	54.709	57.040	52.188,3	6.319,9	11:00	823.427	827.995	823.513	514.228	873.759	911.383	795.717,5	142.328,9
12:00	48.965	51.950	53.403	39.648	53.772	55.755	50.582,2	5.812,6	12:00	845.560	849.613	843.733	524.406	826.894	866.247	792.740,5	132.058,3
13:00	50.928	53.544	54.354	40.556	57.068	58.314	52.460,7	6.392,5	13:00	859.030	877.642	873.688	533.659	829.506	844.371	802.982,7	133.166,3
14:00	53.212	52.998	53.265	36.789	56.856	59.692	52.135,3	7.975,9	14:00	843.630	850.999	858.599	584.722	838.469	860.755	806.195,7	108.833,5
15:00	53.440	54.127	55.353	53.334	56.657	58.168	55.179,8	1.933,1	15:00	840.937	846.949	892.183	771.055	866.244	882.750	850.019,7	43.450,2
16:00	53.447	55.837	54.105	56.075	57.726	57.377	55.761,2	1.713,1	16:00	850.424	855.444	821.729	794.970	872.111	910.794	850.912,0	40.092,9
17:00	51.585	53.352	52.675	53.299	56.699	58.512	54.353,7	2.660,5	17:00	865.528	854.566	825.386	808.602	891.101	931.442	862.770,8	44.553,6
18:00	50.111	51.564	53.422	54.534	56.079	56.949	53.776,5	2.622,9	18:00	840.773	811.158	807.755	777.827	840.953	864.045	823.751,8	30.766,0
19:00	53.545	57.709	57.651	56.735	57.348	56.960,7	1.800,0	19:00	802.808	773.296	772.969	760.905	801.687	800.725	785.398,3	18.461,9
20:00	57.321	61.406	59.165	56.999	60.298	58.798	58.997,8	1.695,0	20:00	789.897	764.322	762.514	743.747	798.485	768.641	774.597,7	20.996,7
21:00	60.123	62.683	59.270	61.058	60.125	58.536	60.299,2	1.448,3	21:00	807.782	790.439	791.694	778.035	826.043	801.992	799.330,8	16.630,9
22:00	59.899	62.029	60.412	61.402	61.350	61.698	61.131,7	810,4	22:00	860.258	863.197	852.262	823.030	861.993	836.230	849.495,0	16.393,9
23:00	60.348	61.585	60.426	58.352	59.419	62.343	60.412,2	1.438,0	23:00	932.373	949.540	938.121	892.724	929.498	896.169	923.070,8	23.236,2

	Location Update Events (UMTS) per Hour								Location Update Events (GSM) per Hour							
	01.03.2012	02.03.2012	03.03.2012	04.03.2012	05.03.2012	06.03.2012	Mean (µ)	StdDev (σ)	01.03.2012	02.03.2012	03.03.2012	04.03.2012	05.03.2012	06.03.2012	Mean (µ)	StdDev (σ)
00:00	75.566	79.877	77.768	74.907	78.526	82.110	78.125,7	2.688,6	00:00	995.419	235.860	0	0	1.009.911	373.531,7	455.833,4
01:00	79.571	81.455	80.022	77.946	79.997	84.003	80.499,0	2.051,9	01:00	1.032.158	266.357	0	0	1.045.040	390.592,5	512.450,9
02:00	76.981	79.257	78.216	77.655	77.282	78.331	77.953,7	824,2	02:00	1.058.846	282.530	0	0	1.068.623	401.666,5	524.388,7
03:00	74.262	76.216	79.582	77.830	77.198	78.877	77.660,8	1.858,7	03:00	1.019.434	254.021	0	0	1.013.557	381.168,7	501.863,1
04:00	74.266	77.274	79.012	76.888	77.543	77.240	77.037,2	1.546,3	04:00	1.056.101	203.450	0	0	1.039.132	383.119,8	520.745,5
05:00	72.143	74.657	76.015	76.249	74.899	75.631	74.932,3	1.500,1	05:00	1.067.991	170.255	0	0	1.061.310	383.259,3	531.910,5
06:00	62.770	67.348	71.819	71.073	65.642	65.933	67.430,8	3.455,3	06:00	1.151.015	83.432	0	0	1.163.948	399.732,5	587.852,9
07:00	59.254	63.286	71.471	72.233	59.300	59.215	64.126,5	6.189,1	07:00	1.116.318	28.415	0	0	1.118.299	377.172,0	573.413,2
08:00	57.882	61.721	66.015	70.254	58.496	58.870	62.206,3	4.961,9	08:00	976.643	10.792	0	0	969.379	326.135,7	501.090,2
09:00	66.218	69.264	64.360	67.026	67.368	70.711	67.491,2	2.243,5	09:00	848.105	0	0	0	511.336	365.333,8	417.874,0
10:00	74.921	77.190	64.910	67.310	75.636	81.069	73.506,0	6.157,8	10:00	800.169	0	0	0	772.833	394.034,5	431.733,1
11:00	71.600	69.600	65.035	66.416	70.468	73.038	69.359,5	3.072,6	11:00	796.665	0	0	0	795.845	398.053,8	436.046,2
12:00	68.543	68.166	62.947	64.786	69.888	68.375	63.784,2	7.978,1	12:00	836.299	0	0	0	828.708	586.154	375.193,5
13:00	70.944	70.268	69.284	66.653	74.631	64.249	69.338,2	3.593,6	13:00	874.244	0	0	0	857.567	875.124	434.489,2
14:00	68.383	64.938	71.898	67.653	67.180	56.325	66.062,8	5.278,3	14:00	831.311	0	0	0	824.877	837.572	415.626,7
15:00	70.003	67.063	69.360	68.445	71.047	43.456	64.895,7	10.590,6	15:00	741.791	0	0	0	809.281	591.163	357.039,2
16:00	74.166	69.210	70.575	72.136	78.861	61.588	71.089,3	5.745,1	16:00	353.344	0	0	0	830.710	840.602	337.442,7
17:00	70.582	63.841	71.162	73.881	73.158	65.733	69.726,2	4.059,9	17:00	105.842	0	0	0	846.057	832.559	297.409,7
18:00	68.802	68.092	67.515	70.732	71.879	69.446	69.411,0	1.645,8	18:00	86.526	0	0	0	799.979	798.376	280.813,5
19:00	75.249	69.424	72.031	73.855	80.896	76.200	74.609,2	3.913,2	19:00	87.089	0	0	0	756.822	757.080	266.831,8
20:00	79.182	74.949	74.555	75.293	79.895	78.983	77.142,8	2.451,6	20:00	63.595	0	0	0	742.612	744.643	258.475,0
21:00	80.888	78.157	77.654	77.443	83.796	80.631	79.761,5	2.477,6	21:00	74.967	0	0	0	765.965	771.275	268.701,2
22:00	81.265	78.669	76.929	79.233	84.482	81.787	80.394,2	2.673,6	22:00	100.189	0	0	0	831.749	830.285	293.703,8
23:00	81.650	77.501	76.925	80.416	84.084	81.362	80.323,0	2.702,0	23:00	166.554	0	0	0	926.570	924.324	336.241,3

Figure 41: Location Area updates in whole Austria from 21.02.2012 until 26.02.2012 and 01.03.2012 until 06.03.2012

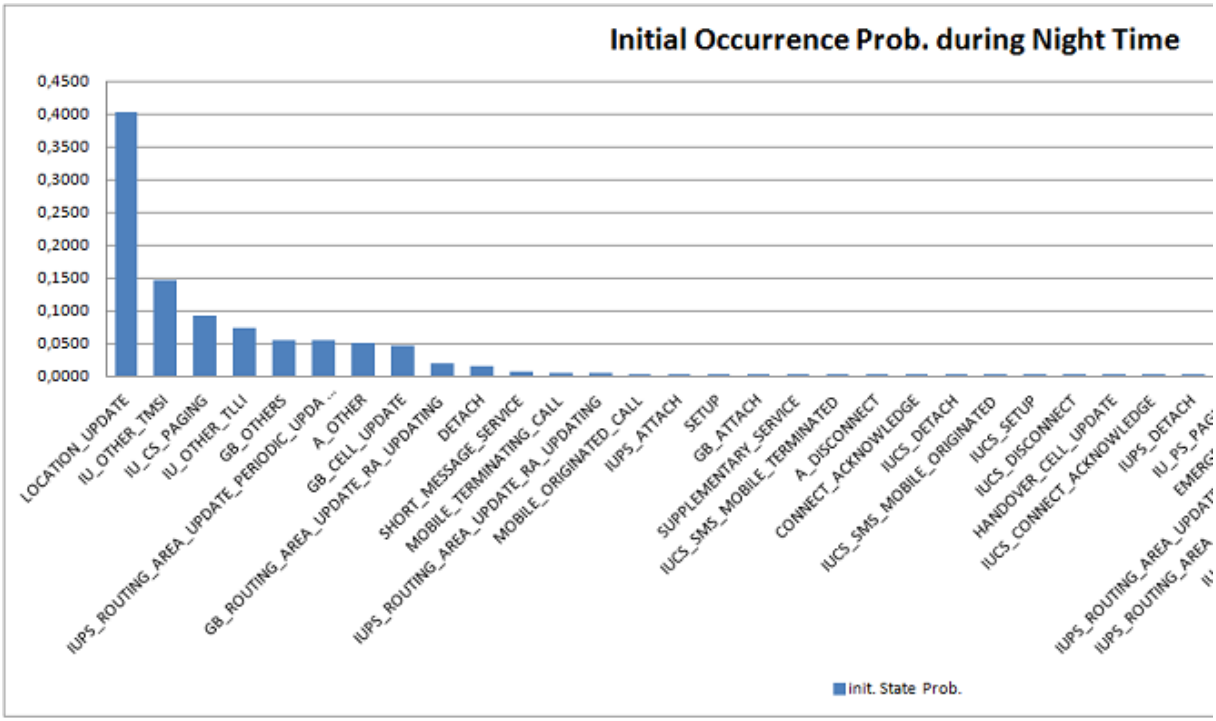
	21.02.2012	22.02.2012	23.02.2012	24.02.2012	25.02.2012	26.02.2012	Mean (μ)	StdDev (σ)
00:00	32.640	45.665	23.149	29.295	56.053	65.561	42.060,5	16.555,4
01:00	18.415	25.988	9.744	12.789	34.001	48.953	24.981,7	14.703,4
02:00	13.287	17.130	5.403	7.933	23.855	37.859	17.577,8	11.921,3
03:00	10.578	11.884	4.469	5.949	16.846	29.013	13.123,2	8.952,4
04:00	10.545	10.463	5.832	7.651	13.924	21.561	11.662,7	5.583,4
05:00	21.295	21.157	18.328	18.825	13.464	13.959	17.838,0	3.416,4
06:00	82.644	82.693	75.793	79.371	26.544	15.684	60.454,8	30.771,2
07:00	283.196	281.108	282.359	289.352	99.105	45.641	213.460,2	110.622,2
08:00	554.981	548.915	556.474	571.188	268.203	132.080	438.640,2	189.830,2
09:00	180.302	656.305	176.295	447.939	445.091	265.352	361.880,7	188.545,1
10:00	749.556	722.320	717.266	447.523	536.556	369.429	590.441,7	161.845,4
11:00	769.813	734.387	722.246	448.919	566.917	420.419	610.450,2	153.202,1
12:00	769.251	694.970	687.711	447.403	542.964	387.929	588.371,3	152.354,9
13:00	749.883	719.732	726.392	454.948	530.929	386.785	594.778,2	157.406,2
14:00	687.186	715.885	733.979	519.869	502.950	389.677	591.591,0	140.450,9
15:00	685.675	742.673	761.363	741.914	496.849	389.824	636.383,0	155.399,6
16:00	744.974	827.254	825.007	774.572	515.055	434.244	686.851,0	169.232,1
17:00	801.059	885.772	890.573	809.137	570.557	484.884	740.330,3	171.011,3
18:00	752.164	785.280	822.239	775.163	593.681	514.561	707.181,3	123.261,9
19:00	613.062	605.963	635.941	616.829	486.479	472.678	571.825,3	72.272,0
20:00	447.768	443.989	458.611	447.016	362.810	350.061	418.375,8	48.401,5
21:00	280.178	262.834	297.275	282.959	235.375	198.702	259.553,8	36.590,6
22:00	177.695	159.414	164.038	187.723	162.158	129.747	163.462,5	19.729,6
23:00	86.890	62.224	70.842	105.722	106.889	49.342	80.318,2	23.542,6

Handovers (GSM) per Hour

	01.03.2012	02.03.2012	03.03.2012	04.03.2012	05.03.2012	06.03.2012	Mean (μ)	StdDev (σ)
00:00	26.067	9.149	0	0	0	19.945	9.193,5	11.435,7
01:00	10.989	4.597	0	0	0	8.297	3.980,5	4.809,7
02:00	6.388	2.834	0	0	0	5.073	2.382,5	2.846,6
03:00	5.500	1.923	0	0	0	4.558	1.996,8	2.481,9
04:00	7.490	1.577	0	0	0	6.481	2.591,3	3.472,8
05:00	21.185	2.979	0	0	0	19.215	7.229,8	10.131,9
06:00	100.642	4.597	0	0	0	101.178	34.402,8	51.547,3
07:00	336.168	5.859	0	0	0	338.001	113.338,0	173.329,1
08:00	601.436	5.094	0	0	0	607.980	202.418,3	311.625,4
09:00	702.426	0	0	0	464.937	708.744	312.684,5	353.626,1
10:00	720.755	0	0	0	745.621	716.134	363.751,7	398.596,2
11:00	727.495	0	0	0	760.038	719.837	367.895,0	403.234,8
12:00	705.385	0	0	0	724.289	488.550	319.704,0	359.883,3
13:00	745.032	0	0	0	767.417	740.022	375.411,8	411.346,5
14:00	763.101	0	0	0	777.924	751.870	382.149,2	418.705,0
15:00	711.619	0	0	0	810.174	553.965	345.959,7	387.693,7
16:00	405.952	0	0	0	886.257	868.850	360.176,5	430.531,3
17:00	137.364	0	0	0	933.362	925.941	332.777,8	465.393,1
18:00	111.339	0	0	0	845.486	857.416	302.373,5	427.510,6
19:00	93.732	0	0	0	666.652	673.732	239.019,3	335.959,5
20:00	53.825	0	0	0	485.625	490.290	171.623,3	245.921,0
21:00	37.444	0	0	0	290.791	301.605	104.973,3	148.869,5
22:00	27.585	0	0	0	152.902	159.447	56.655,7	77.851,3
23:00	18.251	0	0	0	56.390	58.558	22.199,8	28.231,1

Figure 42: Absolute numbers of Handovers from 21.02.2012 until 26.02.2012 and 01.03.2012 until 06.03.2012

Figure 43 shows initial occurrence probabilities of events on the 1st March 2012 for the Night as well as for the Morning time. Figure 44 shows initial occurrence probabilities of events on the 1st March 2012 for the Day and the Evening time. The occurrences are Pareto-distributed.



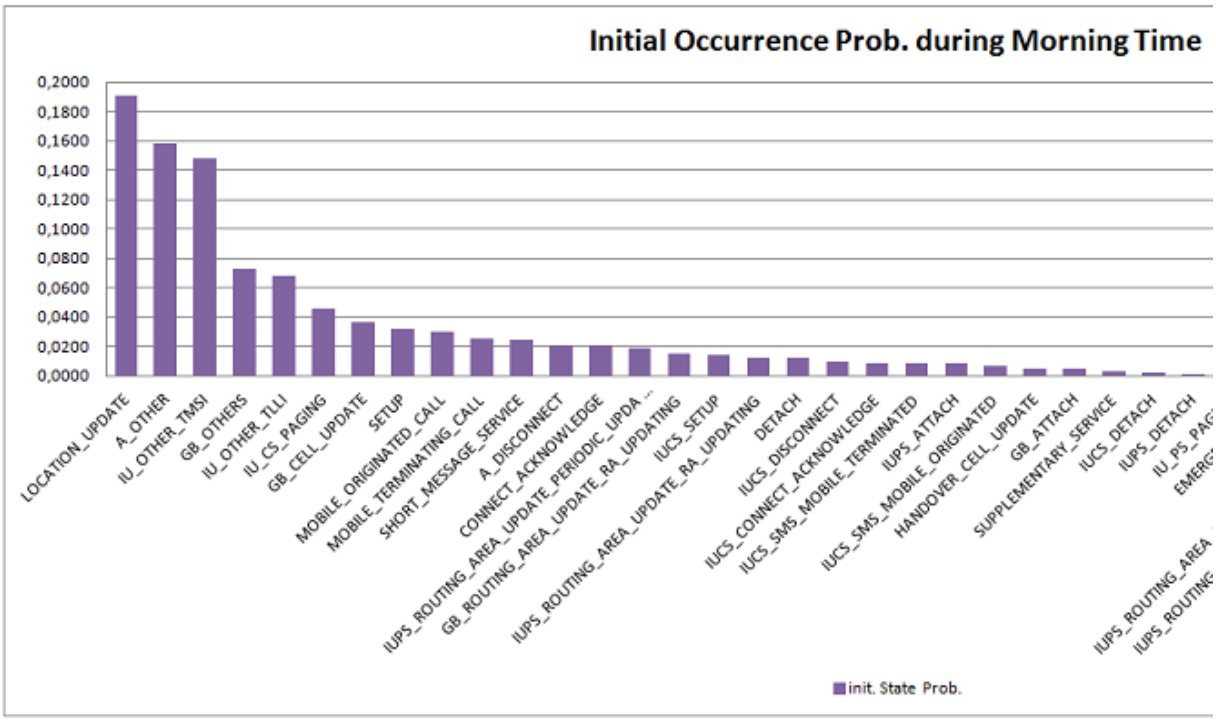


Figure 43: Initial occurrence probabilities of events on 01.03.2012 for the Night and the Morning time

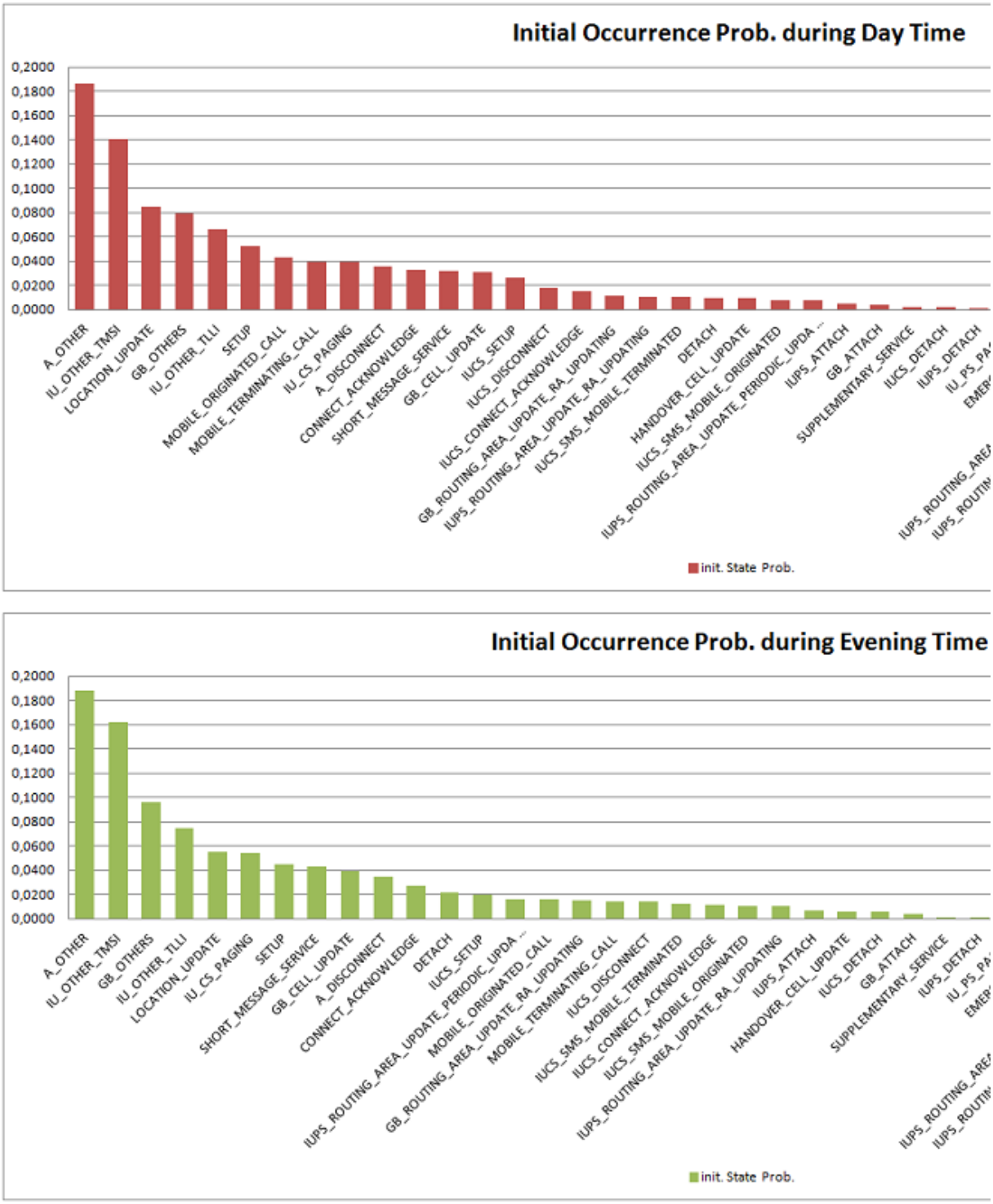


Figure 44: Initial occurrence probabilities of events on 01.03.2012 for the Day and the Evening time